PatNet: A lexical database for the patent domain

Tannebaum and Andreas Rauber

Institute of Software Technology and Interactive Systems,
Vienna University of Technology, Austria
http://www.ifs.tuwien.ac.at
{tannebaum, rauber}@ifs.tuwien.ac.at

Abstract. In the patent domain Boolean retrieval is particularly common. 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*, to provide synonymous expansion terms. In this paper we present a new approach to support patent searchers in the query generation process. We extract a lexical database, which we call *PatNet*, from real query sessions of patent examiners of the United Patent and Trademark Office (USPTO). *PatNet* provides several types of synonym relations. Further, we apply several query term expansion strategies to improve the precision measures of *PatNet* in suggesting expansion terms. Experiments based on real query sessions of patent examiners show a drastic increase in precision, when considering support of the synonym relations, US patent classes, and word senses.

Keywords: Patent searching, Query term expansion, Query log analysis;

1 Introduction

In the patent domain Boolean retrieval is particularly common. 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 [3]. But despite the importance of Boolean retrieval in patent searching, as shown in [8], there is not much work in current research assisting patent experts in formulating such queries, preferable via automatic query term expansion.

In this paper we present a new approach to support patent searchers in the query generation process. We extract a lexical database, which we call *PatNet*, from real query sessions of patent examiners of the USPTO. First, we review related work on automatic query term expansion in patent searching. We then describe the approaches to detect several types of synonym relations in the query logs. Following we present the lexical database *PatNet*. Finally, we provide the experiments to improve the precision measures of *PatNet* followed by conclusions and an outlook on future work.

2 Related Work

Related approaches to enhance query term expansion in patent searching are mostly limited to computing co-occurring terms in a patent corpus for query expansion, while patent searchers predominately use synonyms and equivalents for query term expansion [1,5]. An analysis of real query sessions of patent examiners has shown that about 60% of the used expansion terms (*ETs)* are synonyms and equivalents [8]. Further, [9] shows that the highly specific vocabulary used in the patent domain is not included in standard dictionaries, such as *WordNet.* Patent examiners use the terms created by the patent applicants, such as "*pocketpc*" for "*notebook*", "*watergas*" for "*steam*", or "*passcode*" for "*password*" for synonym expansion. Hence, the challenge is to learn the synonyms directly from the patent domain to assist patent searchers in formulating Boolean queries. An approach to extract synonyms directly from patent documents is presented in [5]. Claim sections of granted patent documents from the European Patent Office 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. Contrary to the extraction of the synonyms from patents, as indicated in [5], we propose to extract them from query logs as presented in [7] and in particular from query logs of patent examiners as suggested in [9]. This allows us to extract specific terms, in particular the query and expansion terms to the patent applications.

3 Extracting synonyms from query logs of patent examiners

For our experiments we downloaded and preprocessed 103,896 query log files of USPTO patent examiners from Google as mentioned in [9].[1] We kept 7,500 log files as a hold-out set for evaluation and used 96,396 files for the following experiments.

In [9] the Boolean Operator "OR", which indicates that two query terms are synonyms or can at least be considered as equivalents, was used for detecting synonyms (single term relations) in the text queries. Expanding the approach, we now use the proximity operator "ADJ" to detect keyword phrases and the Boolean operator "OR" to learn synonyms thereto. Table 1 shows several types of synonym relations provided by the search operators "OR" and "ADJ" and for each type of relation an example.

Table 1. 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)

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

The process to detect single term relations works as follows: We filter all 3-grams generated from the text queries in the form “X *b* Y”, where *b* is the Boolean operator “OR” and X and Y are query terms. To exclude mismatches and misspellings, we consider those 3-grams that were encountered at least three times. 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 operator “OR”. To exclude mismatches, we consider the correctly set parentheses. Table 2 shows the detected synonym relation frequencies.

Table 2. Detected Synonyms based on the Search Operators.

Type of Relation	Code	#Relations	#Terms
single term	STR	27,798	17,105
single term to phrase	STPR	628	928
phrase to phrase	PPR	409	701
Σ	-	**28,835**	**17,643**

In addition, we learned that patent examiners may also rely on a default operator, which can be set to “OR” or “AND”. This is indicated by the default operator element in the query logs. To detect these synonyms, we use all text queries where the default operator is set on “OR” and the approach to detect synonyms as mentioned above, but we excluded the “OR” operator in the 3-, 5- and 7-grams. We obtained 1,871 single term relations, 394 single term to phrase, and 165 phrase to phrase relations.

4 *PatNet*: A lexical database

Based on the detected synonym relations, we learn in this section a lexical database for the patent domain, which we call *PatNet*. The lexical database resembles a thesaurus of English concepts that can be used for semi-automatic query term expansion. To query the lexical database we use the open source thesaurus management software *TheW32* [2].

Table 3. 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**

As shown in Table 3, *PatNet* 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.

Table 4. Suggested *STR*, *STPR* and *PPR* for the query term "*voice*".

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	-	-	-

Table 4 shows the provided *ETs* for the term "*voice*". *PatNet* suggests 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*).

5 Experiments

In this section we apply several query term expansion strategies to suggest *ETs* in a useful order to avoid time-consuming term selection. For the single terms *PatNet* provides, on average, 11 *ETs*. But the maximum number rise up to 92 terms, for common terms, such as "*sensor*". For the experiments we use the test set from Subsection 3.1. and measure the performance of *PatNet* based on real query sessions of patent examiners (gold standard), because (1) benchmark data sets with synonym relations are not available for the patent domain and (2) the performance of thesauri in *IR* depends on contextual factors, as shown [4].

At first, we rank the synonym relations of *PatNet* according to their support in the training set and carry out five expansion steps ($Step_1$ *to* $Step_5$) which is a realistic value in real query 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* $Step_5$. For each expansion step we calculate recall (we compare the suggested *ETs* from *PatNet* with the synonyms used by the examiners in the test set) and precision (we compare the synonyms used by the examiners with all *ETs* suggested by *PatNet*). For recall we consider the obtained scores of the previous expansion steps.

Table 5. Recall and Precision achieved when successively suggesting the highest ranked *ETs*.

Expansion Step	***Ranking***	***Positions***	**Recall**	**Precision**
$Step_1$	r_1	*1 – 5*	**38.46**	23.10
$Step_2$	r_2	*6 – 10*	48.72	**24.81**
$Step_3$	r_3	*11 – 15*	55.38	22.31
$Step_4$	r_4	*16 – 20*	58.38	20.45
$Step_5$	r_5	*21 - 25*	**62.54**	**20.00**

As shown in Table 5, in $Step_1$ *to* $Step_5$, on average, 1 out of 5 terms that are suggested by *PatNet* as synonyms were used by the examiners for query expansion (on average

22% precision). Further, after *Step$_2$ PatNet* already provides almost half of the *ETs* used (49% recall). Compared to 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%).

Next, we consider specific and related US patent classes, as presented in [9], to suggest *ETs* in a certain context (patent class). In addition, we use the idea behind Relevance Feedback *RF* to take the *ETs* that are initially suggested for a *QT* and to use information about whether or not those are relevant to perform a new expansion step. At first, we consider the US patent classes of the *QTs* and expand the terms with class-specific *ETs* (*Step$_1$*). Then, we expand the relevant *ETs* 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 6. Recall and Precision achieved when using intersections between US patent classes.

Expansion Step	*Expansion Terms*	**Recall**	**Precision**
Step$_1$	*class-specific*	**49.38**	**18.50**
Step$_2$	*class-related*	50.86	17.37
Step$_3$	*class-independent*	**54.99**	**12.21**

Table 6 shows that after *Step$_1$* almost half of the used *ETs* are provided by the class-specific *ETs* with best precision (19%). In *Step$_2$*, the recall measure could be further improved, while we notice only a minor decrease in precision (17%). In *Step$_3$* precision fall to 12% and recall rises to 55%. In light of suggesting all possible *ETs* in one step, there is a significant increase in precision, but also a major decrease in recall.

Finally, we perform word sense disambiguation (*WSD)* to suggest the most suitable *ETs*. We determine the sense of an *ET* based on the overlap of the sense definitions of the target word, as mentioned in [6]. We consider the *QTs*, which appear before the *STR* in the training and test set (reflecting real query expansion scenarios, where information from past queries can be used). We use a context size of n = 20 words. We rank the *ETs* according the number of common words (highest overlap) and initially suggest the highest ranked *ETs* followed by additional ones.

Table 7. Recall and Precision achieved when using *WSD*.

Expansion Step	*Ranking*	*Overlap*	**Recall**	**Precision**
Step$_1$	r_1	≥ 5	**6.06**	**44.44**
Step$_2$	r_2	*4*	9.09	37.50
Step$_3$	r_3	*3*	12.12	36.36
Step$_4$	r_4	*2*	18.18	19.35
Step$_5$	r_5	*1*	**30.30**	**11.24**

As shown in Table 7, compared to the expansion strategies applied before, there is a further increase in precision (up to 44% in *Step$_1$*). But now also a decrease in recall has to be noticed. Recall measures already decrease from 70% to 30%, when considering only one common term in the context words. Further experiments show that also a considerable decrease in recall has to be noticed (from 70% to 56%), when using a context size of 50 terms, while now the precision scores, on average, rise up to 20%.

6 Conclusions and Future Work

In this paper we presented a new approach to support patent experts in formulating Boolean queries. We used real query expansion sessions of patent examiners to learn the lexical database *PatNet*. We have shown that *PatNet* can be used to support patent searchers in the time-consuming query generation process. Experiments showed that the achieved precision scores significantly exceed the scores achieved in related work for patent searching and are comparable to numbers reported for professional academic search [3,9,10]. Specifically, we notice only a minor decrease in recall, when considering support of the extracted relations and successively suggesting the highest ranked *ETs* (while precision increases). In future work we want to evaluate *PatNet* based on the relevant documents cited by the patent examiners in their search reports to measure the performance of our query expansion approach in document retrieval.

References

1. 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 2013), Moscow, Russia, pp. 676-679.
2. 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.
3. 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 (SIGIR 2011), Beijing, China, pp. 825-834.
4. 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.
5. 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.
6. Navigli, R. 2009. Word sense disambiguation: A survey. In ACM Computing Surveys (CSUR), Volume 41, Issue 2, Article 10.
7. 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.
8. 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, pp. 136-142.
9. 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.
10. 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.