Mining Query Logs of USPTO Patent Examiners

Wolfgang 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 this paper we analyze a highly professional search setting of patent examiners of the United Patent and Trademark Office (USPTO). We gain insight into the search behavior of USPTO patent examiners to explore ways for enhancing query generation in patent searching. We show that query generation is highly patent domain specific and patent examiners follow a strict scheme for generating text queries. Means to enhance query generation in patent search are to suggest synonyms and equivalents, co-occurring terms and keyword phrases to the searchable features of the invention. Further, we show that term networks including synonyms and equivalents can be learned from the query logs for automatic query expansion in patent searching.

Keywords: Patent Searching, Query Log Analysis.

1 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. Patent searching is usually performed by examiners in patent offices and patent searchers in private companies.

There is an increasing need to assist patent searchers in formulating queries, because query formulation is very time-intensive [1,5,6]. Yet, in the patent domain no sources, such as patent domain specific lexica or thesauri, are available. Actual queries being posed by patent experts could be valuable resources to explore the requirements for supporting patent searchers in query generation. The United Patent and Trademark Office (USPTO) has stored and published the query logs of the patent examiners. The goal of this paper is to analyze the query logs of the USPTO patent examiners to gain insights into the search behavior and characteristic of patent examiners queries. We first review state-of-the-art techniques for mining query logs. We then describe the nature of the query logs of USPTO patent examiners and analyze them. Following we present lexical term networks learned from the query logs. Finally, we provide conclusions and an outlook on future work.

P. Forner et al. (Eds.): CLEF 2013, LNCS 8138, pp. 136–142, 2013.

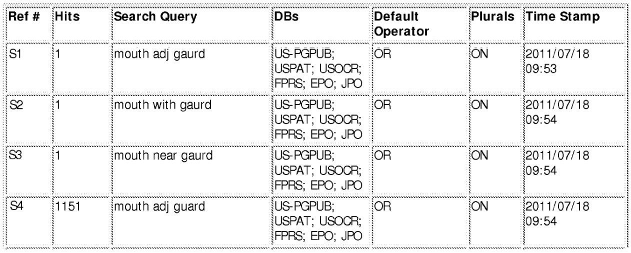
2 Related Work

In several information retrieval applications query logs are being intensively studied. The purpose of all studies is to enhance either effectiveness or efficiency of searching by discovering patterns from query logs of search engines [2]. The main focus is on the analysis of web queries to enhance web searches [7]. Large-scale data sets of web queries, which have been made publicly available, such as AltaVista log or AOL log, have been studied [8]. Predominantly, basic statistics, such as query and term popularity, average query length, or co-occurring terms are used for characterizing the queries. Further specific analysis of the logs, such as distribution of the queries over time, variations of topics over time or distance between repetitions of queries over time, has been carried out. The classification of the queries, particularly through topic popularity, is a further task in mining query logs. The distribution of large-scale data sets across general topics enables to retrieve domain specific characteristics [7,8].

Finding query logs in the patent domain has been a difficult task [4]. Private companies and searchers are not interested in making their logs available as these may include terms revealing their current R&D activities. In earlier work we provided initial analyses of query logs of US Patent and Trademark Office (USPTO) patent examiners. We manually downloaded a limited set (346 log files) for one specific patent domain from the USPTO portal PAIR [10]. Initial results indicated that specialized term networks can be extracted directly from the query logs to complement resources for standard English [9]. In this paper we present a more in-depth analysis of this high professional search setting. We collect and analyzed the by now largest corpus of patent query logs to gain insight into query generation behavior as basis for automatic query expansion.

3 Query Logs of the USPTO

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) and can be downloaded from (http://www.uspto.gov/). The download is limited by the USPTO. For each patent application a verification code has to be entered. Google has begun crawling the USPTO's public PAIR sites and provides free download of all patent applications published until now (http://www.google.com/googlebooks/uspto-patents.html). Google created single zip file for each patent application. Each file contains several folders including information on: 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 is of concern to us here. This folder can contain one or several query log files. 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. Each query has several elements. We focus on the search query element showing the query formulated by the patent examiner. Further elements are reference, hits, database(s), default operator, plurals, and time stamp.



Ref #	Hits	Search Query	DBs	Default Operator	Plurals	Time Stamp
S1	1	mouth adj gaurd	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO	OR	ON	2011/07/18 09:53
S2	1	mouth with gaurd	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO	OR	ON	2011/07/18 09:54
S3	1	mouth near gaurd	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO	OR	ON	2011/07/18 09:54
S4	1151	mouth adj guard	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO	OR	ON	2011/07/18 09:54

Fig. 1. Example of a USPTO query log

There are several kinds of queries in the search query element. Text queries are used for querying whole documents (fulltext search) or only sections of patent documents, such as the title section (title search). Non-text queries are used for number search or classification search, for example "148/674.ccls." for searching the class 148/674 for "Metal treatment", specifically for "Cobalt or cobalt base alloy". For query formulation text queries include search operators between the query terms. The types of search operators are (1) Boolean operators, such as "AND or OR" and (2) Proximity operators, like "SAME, ADJ(acent), NEAR, or WITH". Furthermore, Truncation Limiters, such as "$", are used for query formulation. If the search operators are added manually, they are shown between the query terms in the text query element, else they are indicated by the default operator element. We are interested in the queries including the search operators.

4 Query Log Analysis

The USPTO published about 2.7 million patent applications, since 2003. The applications are classified into 473 US classes each including several subclasses. Hence, on average, about 6000 application documents are available for each US class. Because patent searchers use the classification system to narrow the search, we selected three collections of query logs each for a specific US class. We selected the US class 433 called "Dentistry", the US class 128 called "Surgery" (a similar domain to the US class 433) and the US class 126 for "Stoves and Furnaces" (a domain very different from the US classes 433 and 128). For our query log analysis experiments we downloaded 2,721 files for the US class 126, 4,025 files for the US class 433 and 8,758 files for the US class 128. Through OCR conversion and segmentation of the 15,504 query log files we extracted the Boolean and Proximity Queries and the search operators between the query terms. We filtered all 3-grams in the form "X *b* Y", where *b* is an Boolean or Proximity operator and X and Y are query terms.

4.1 Vocabulary Analysis

In this section we show for each US class some basic statistical properties of the vocabulary. At first we learn from the USPTO query logs how terms co-occur in

Table 1. 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 query logs based on the Boolean and proximity operators. In Table 1 we present the five most frequently co-occurring terms for the three US classes based on the Boolean operator "OR".

The majority out of the top-200 co-occurring terms are synonyms or equivalents at least for each specific domain. This show, that patent examiners use the Boolean operator "OR" to generate synonyms or equivalents. In Table 2 we show the top-five co-occurring terms based on the proximity operators "SAME, "ADJ(cent)", "NEAR" and WITH". In all classes studied the majority of term pairs are keyword phrases. Hence, to narrow a search, particularly to limit a general query term, for example "mouth", a keyword phrase is generated by the patent examiners, such as "mouth piece".

Table 2. Co-Occurring Terms based on Proximity Operators

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

Further, we analyze the query terms of each class w.r.t. the part of speech using the CLAWS part of speech tagger [3], and if the query terms used by the patent examiners are domain specific (the terms appear only in one specific US class). 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 terms are nouns followed by verbs (about 13%) and adjectives (about 10%). This can be useful for suggesting additional query terms from patent documents. The class 128 for "Surgery" and class 433 for "Dentistry" have the most common terms (3,673 terms) followed by the class 126 "Stoves" and US Class 433 "Dentistry" (having 3,483 common terms). Fewest common terms (1,751 terms) 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 domain specific. Less than 5% of the query terms of the specific classes appear in the other classes, even across similar domains.

4.2 Search Operator Analysis

In this section we present for each class some basic statistical properties on the used search operators. First we analyze operator popularity for each domain based on the usage of the Boolean and proximity operators. Tab. 3 shows the relative spread of the used operators for formulating Boolean and proximity queries for each class.

Table 3. Search Operator Popularity

Search Operator	Stoves and Furnaces	Dentistry	Surgery
Boolean "OR"	57.65 %	46.92 %	48.24 %
Boolean "AND"	22.37 %	29.99 %	29.37 %
Proximity	19.98 %	23.09 %	22.39 %

In each domain about half of the queries are built using "OR", nearly one third of the queries are generated using "AND" and the remaining queries are built by the proximity operators. The analysis shows, that the examiners' behavior in formulating queries in the three domains is similar. For all domains they generate in the same proportions synonyms and equivalents, co-occurring terms and keyword phrases. Comparisons of the kinds of 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. 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". The Boolean query is used to search for the synonyms or equivalents "drill" or "bit".

5 Detecting Synonyms and Equivalents

In the patent domain significant efforts are invested to assist researchers in formulating better queries, preferably via automated query expansion. Currently, automatic query expansion in patent search is mostly limited on computing co-occurring terms. Learning synonyms and equivalents in the patent domain has been a difficult task. As we learned in Section 4 in patent searching the Boolean operator "OR" is used to expand a query term with an expansion term, which has the same meaning. We use that for automatically learning term networks from the query logs of USPTO patent examiners. Our approach resulted in 26,653 unique synonyms and 29,702 unique synonym relations for the three patent US classes as presented in Table 4 in detail.

Table 4. Learned Term Networks

US Class	unique relations	unique query terms
126	4,155	3,058
433	7,441	7,547
128	18,106	16,048
Σ	**29,702**	**26,653**

The learned lexical databases, particularly term networks, resemble thesauri of English terms for each specific patent domain. In each term network terms that have the same meaning are linked to each other. Finally, the learned term networks can be used in each specific US class for (semi-) automated query suggestion, particularly query expansion.

6 Conclusions and Future Work

In this paper we introduced and analyzed query logs of USPTO patent examiners. We show that query generation in patent searching is highly domain specific. Patent examiners follow a strict scheme for generating text queries. In each domain they use the Boolean 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. Finally, means to enhance query generation in patent search are to suggest synonyms and equivalents, co-occurring terms and keyword phrases. Further we show, that specialized term networks including synonyms and equivalents can be extracted to complement resources for standard English. As shown in [9] this has positive effects on automated query expansion in patent searching. Currently, we are collecting and preprocessing a larger corpus of patent query logs to obtain a broader basis of USPTO classes. In future work we will focus on evaluating the performance of the learned term networks based on real query sessions done by the patent examiners. Further we want to use the proximity operators to learn term networks of keyword phrases, which we use for query limitation in patent searching.

References

1. Azzopardi, L., Vanderbauwhede, W., Joho, H.: 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 (2010)
2. Clough, P., Berendt, B.: Report on the Treble CLEF query log analysis workshop 2009. SIGIR Forum 43, 71–77 (2009)
3. Garside, R., Smith, N.: A hybrid grammatical tagger: CLAWS4. In: Garside, R., Leech, G., McEnery, A. (eds.) Corpus Annotation: Linguistic Information from Computer Text Corpora, pp. 102–121. Longman, London (1997)
4. Jürgens, J.J., Hansen, P., Womser-Hacker, C.: Going beyond CLEF-IP: The ‘Reality’ for Patent Searchers? In: Catarci, T., Forner, P., Hiemstra, D., Peñas, A., Santucci, G. (eds.) CLEF 2012. LNCS, vol. 7488, pp. 30–35. Springer, Heidelberg (2012)
5. Magdy, W., Jones, G.J.F.: A Study of Query Expansion Methods for Patent Retrieval. In: Proceedings of PaIR 2011, Glasgow, Scotland, pp. 19–24 (2011)
6. Piroi, F., Lupu, M., Hanbury, A.: Effects of Language and Topic Size in Patent IR: An Empirical Study. In: Catarci, T., Forner, P., Hiemstra, D., Peñas, A., Santucci, G. (eds.) CLEF 2012. LNCS, vol. 7488, pp. 54–66. Springer, Heidelberg (2012)

7. Silverstein, C., Marais, H., Henzinger, M., Moricz, M.: Analysis of a very large web search engine query log. SIGIR Forum 33, 6–12 (1999)
8. Silvestri, F.: Mining Query Logs: Turning Search Usage Data into Knowledge. Foundations and Trends in Information Retrieval 4(1-2), 1–174 (2010)
9. Tannebaum, W., Rauber, A.: Acquiring lexical knowledge from Query Logs for Query Expansion in Patent Searching. In: The Proceedings of the 6th IEEE International Conference on Semantic Computing (IEEE ICSC 2012), Palermo, Italy (2012)
10. Tannebaum, W., Rauber, A.: Analyzing Query Logs of USPTO Examiners to Identify Useful Query Terms in Patent Documents for Query Expansion in Patent Searching: A Preliminary Study. In: Salampasis, M., Larsen, B. (eds.) IRFC 2012. LNCS, vol. 7356, pp. 127–136. Springer, Heidelberg (2012)