World Patent Information 41 (2015) 15–22

Contents lists available at ScienceDirect

World Patent Information

journal homepage: www.elsevier.com/locate/worpatin

Learning keyword phrases from query logs of USPTO patent examiners for automatic query scope limitation in patent searching

Wolfgang Tannebaum, Andreas Rauber*

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

ARTICLE INFO

Article history:
Available online 14 March 2015

Keywords:
Patent searching
Query term expansion
Lexical resources

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 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 extract 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.

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 earlier 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 detected 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 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 extracting only synonyms from the query logs. We present an approach to detect 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-

* Corresponding author. Tel.: +43 1 58801 18826.
E-mail addresses: tannebaum@ifs.tuwien.ac.at (W. Tannebaum), rauber@ifs.tuwien.ac.at (A. Rauber).

http://dx.doi.org/10.1016/j.wpi.2015.02.005
0172-2190/

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 2. In Section 3 we present our approach to extract keyword phrases from the query logs and the lexical databases generated for specific US patent classes. Experiments on 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 [6–9]. 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 assist patent searchers in query generation [12–14].

To provide synonyms for conceptual search, standard dictionaries, such as *WordNet*, or lexica, like *Wikipedia*, are used for query refinement [8,15]. To learn them directly from the patent domain, as described in Refs. [8], 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. These are the so-called corpora. Based on the language pairs having the same translation terms, synonyms are learned in English, French and German [16].

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 [17–19]. In the same way as for learning synonyms, standard dictionaries and lexica are used to learn the keyword phrases. Further, in Ref. [16] the claims are aligned to detect keyword phrases based on the translations, in particular based on term to phrase translations, phrase to term translations and phrase to phrase translations.

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 Ref. [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 Ref. [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 6000 patent applications per class). Each query log of the USPTO is a PDF file consisting of a series of queries. Fig. 1 shows an example, particularly an extract of four text queries of such a 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

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