Inf Retrieval (2014) 17:412–429
DOI 10.1007/s10791-013-9232-5

INFORMATION RETRIEVAL IN THE INTELLECTUAL PROPERTY DOMAIN

The effect of citation analysis on query expansion for patent retrieval

Parvaz Mahdabi · Fabio Crestani

Received: 3 May 2013 / Accepted: 13 September 2013 / Published online: 21 November 2013

Abstract Patent prior art search is a type of search in the patent domain where documents are searched for that describe the work previously carried out related to a patent application. The goal of this search is to check whether the idea in the patent application is novel. Vocabulary mismatch is one of the main problems of patent retrieval which results in low retrievability of similar documents for a given patent application. In this paper we show how the term distribution of the cited documents in an initially retrieved ranked list can be used to address the vocabulary mismatch. We propose a method for query modeling estimation which utilizes the citation links in a pseudo relevance feedback set. We first build a topic dependent citation graph, starting from the initially retrieved set of feedback documents and utilizing citation links of feedback documents to expand the set. We identify the important documents in the topic dependent citation graph using a citation analysis measure. We then use the term distribution of the documents in the citation graph to estimate a query model by identifying the distinguishing terms and their respective weights. We then use these terms to expand our original query. We use CLEF-IP 2011 collection to evaluate the effectiveness of our query modeling approach for prior art search. We also study the influence of different parameters on the performance of the proposed method. The experimental results demonstrate that the proposed approach significantly improves the recall over a state-of-the-art baseline which uses the link-based structure of the citation graph but not the term distribution of the cited documents.

Keywords Citation analysis · Patent retrieval · Query expansion

P. Mahdabi (✉) · F. Crestani
Faculty of Informatics, University of Lugano, Lugano, Switzerland
e-mail: parvaz.mahdabi@usi.ch

F. Crestani
e-mail: fabio.crestani@usi.ch

1 Introduction

A patent is a legal document that, when granted by a country's patent office, gives a set of rights of exclusivity and protection to the owner of an invention. Patent search is a general term that covers different types of search processes such as *technology survey, prior art search, freedom to operate, validity* and *patent portfolio search*. These search processes differ in terms of the information need of the searcher, the corpora and the output of the search. Notice, however, that the precise names and definitions of these search processes vary between those who deal with patents, like for example, information specialists, private patent searchers, patent examiners, and patent lawyers (Lupu and Hanbury 2013). In this paper we focus our attention on the prior art search which is a critical step in the examination and evaluation process of a patent application.

Patent prior art search (also referred to as *Patentability* and *Novelty*) is mainly composed of a search over previously published patent and non-patent documents with the aim of verifying whether the idea of a patent application is novel, i.e. has not been previously patented by someone else, has not been described in a scientific paper or disclosed to the public through any other medium. The objective of this type of search is to retrieve all relevant documents that may invalidate or at least describe prior art work in a patent application (Lupu et al. 2011). This type of search addresses the challenging task of finding relevant, highly technical and domain-specific content of patents (Atkinson 2008).

The challenges of patent prior art search are different from those of standard ad hoc text and web search (Baeza-Yates and Ribeiro-Neto 2011). The first difference is associated with the query length: patent prior art queries are full patent applications comprising hundreds of words as opposed to ad hoc and web searches where queries are usually rather short (Magdy et al. 2009).

The second issue is related to the fact that patent prior art search is a recall oriented task where the goal is to retrieve all relevant documents at early rank positions. Ad hoc and web search, on the other hand have the goal of retrieving only a few relevant documents at the top of a ranking and thus achieving high precision (Bashir and Rauber 2009). Patent examiners are therefore required to have a vast knowledge of all relevant and related patents. In such a search scenario even missing one relevant patent can lead to a multimillion Euro law suit due to patent infringement, and so a high recall is demanded in this type of search.

The third property is attributed to the vocabulary usage in the patent domain. The language of patents is unique and contains highly specialized or technical words not found in everyday language (Joho et al. 2010). The abstract and description section of a patent use a technical terminology while the claims section uses a legal jargon. Patent retrieval is often cumbersome and distinct from other information retrieval tasks. This is because of the inherent properties of patents, namely, exceptional vocabulary, curious grammatical constructions, and legal requirements (Atkinson 2008). Patent authors purposely use many vague terms and a non-standard terminology in order to avoid narrowing down the scope of their invention. This exacerbates the retrieval problem and can confuse standard IR approaches and systems.

Because of these challenges, the work performed by patent examiners involves manual query formulation from the query patent in order to find invalidating claims. They consider high term frequency in the document to be a strong indicator of a good query term. The keyword-based searches built from the query patent are then completed using other metadata associated with the patent applications such as International Patent Classification

(IPC classes)[1] and date tags. IPC classes are language independent keywords assigned as metadata to the patent documents with the purpose of categorizing their content. Such classes describe the field of technology that a patent document belongs to. These IPC classes resemble tags assigned to documents in standard information retrieval tasks.

Behavioral studies of patent examiners in patent offices show that, besides the keyword based query and the classification based query, the other sources that are influencing the most the searching practice of patent examiners are the bibliographic information (Lupu et al. 2011). This includes both backward and forward citations. Forward citations denote the citations to a given patent document from patents which are *forward in time* from the patent of interest. In contrast, the backward citations indicate the citations made by the patent to patents which are *backward in time* with respect to the given patent.

The question is how can building queries from different information sources such as classifications and citations lend additional power to the original query itself. As patent authors try to obfuscate their invention by using non standard terminology, there is often a gap between the terms in the query document and the documents relevant to that query (Bashir and Rauber 2010; Magdy et al. 2009). We must cope with the fact that documents relevant to a given query may not contain the exact terms used by the author, which are given to our system as specific query terms.

We are interested in overcoming this gap by tapping the power of the community of inventors related to the subject of the invention of the query. To this end, we want to boost the original query with the terms used in the cited documents. In other words, through citation link analysis we identify a set of terms which are relevant to a given query document and appear in the cited documents. These terms can be exploited for improving the original ranking. Thus, the main research questions we aim to answer in this paper are:

- Does using the content of the cited documents in addition to their link-based structure lead to improvements?
- Does the language model of the cited documents complement the language model of the original query?

We try to capture the influence of the citation links in the graph structure of patent documents in two scenarios and compare them. We first use a link-based measure to compute the importance of each document in the graph in a topic-sensitive manner. We then use the term distribution of cited documents to estimate a query model from the cited documents by identifying distinguishing terms and their corresponding weights. We perform a query expansion using the estimated query model from the cited documents to improve the language model of the original query.

We evaluate our work on CLEF-IP 2011 patent retrieval corpus. The experimental results show that our query expansion model using the distribution of the cited documents, achieves significant improvement in terms of recall over a baselines which uses solely the citation links.

The rest of this paper is organised as follows: Sect. 2 reviews the related work. Section 3 defines the original query model. Section 4 explains the construction of a citation graph for a given patent application and describes the citation analysis over the graph. Section 5 describes a method to estimate a query model from the cited documents exploiting the citation-based measures. Sections 6 and 7 report the setup and results of the experiments aimed at proving the validity of the approach. Section 8 describes the analysis carried out

[1] http://www.wipo.int/classifications/ipc/en/.

to study the influence of different parameters on the performance of the proposed method. Finally, Sect. 9 reports the conclusions of the work and some directions for future work.

2 Related work

Recently, patent processing has attracted considerable attention in the academic research community, in particular from information retrieval and natural language processing researchers (Fujii et al. 2007).

The main research in patent retrieval started after the third NTCIR workshop in 2003(Iwayama et al. 2003), where a few patent test collections were released. Starting from the fourth NTCIR workshop in 2004 (Fujii et al. 2004), a search task related to the prior-art search was presented which was referred to as an invalidity search run[2]. The goal was to find prior-art before the filing date of the application in question that conflicts with the claimed invention. The citation parts of the applications are removed and counted as relevant documents used for the evaluation of results. Participants performed different term weighting methods for query generation from the claims. They applied query expansion techniques by extracting effective and concrete terms from the description section of the patent document to enhance the initial query.

2.1 Generating query from a patent application

In (Takaki et al. 2004) the authors study the rhetorical structure of a claim. They segmented a claim into multiple components, each of which is used to produce an initial query. They then searched for candidate documents on a component by component basis. The final result was produced from the candidate documents. Similar work was introduced in (Mase et al. 2005) where the authors analyzed the structure of claims to enhance retrieval effectiveness. A claim structure usually consists of the *premise* and *invention* parts, which describe existing and new technologies, respectively. The authors proposed a two stage process where they first extract query terms from the premise part to increase the recall. They then aim to increase the precision by extracting another query from the invention part. The final relevance score of each document is calculated by merging the scores of the two stages.

IPC classification has been used as an extra feature besides the content of the patent. Different methods for combining text content and classification were proposed. In (Takaki et al. 2004) the authors use IPC codes for document filtering and show how this feature can help in patent retrieval. In (Fujita 2004) the authors integrate IPC codes into a probabilistic retrieval model, employing the IPC codes for estimating the document prior.

In (Fujii 2007), author applied link analysis techniques to the citation structure of patents. He calculated two different scores based on textual information and citation information. He showed that by combining these two scores he can achieve better performance.

A recent line of work advocated the use of the full patent application as the query to reduce the burden on patent examiners. This direction was initiated by Xue and Croft (2009), who conducted a series of experiments in order to examine the effect of different

[2] Invalidity search (also called validity search) is performed over all public documents prior to the priority date of a granted patent. The difference between invalidity search and prior art search is that the input of the former is a granted patent, while the input of the latter is a patent application.

patent fields on the query formulation and concluded with the observation that the best Mean Average Precision (MAP) is achieved using the text from the description section of the query patent with raw term frequencies.

The current developments in the patent search are driven by the Intellectual Property task within the CLEF[3] initiative. Several teams participated in the prior art search task of the CLEF-IP 2010 and proposed approaches to reduce the query patent by extracting a set of key terms from it. Different participating teams experimented with term distribution analysis in a language modeling setting employing the document structure of the patent documents in various ways (Piroi and Tait 2010). We now discuss with details the two best performing approaches in CLEF-IP 2010. Lopez et al. (Lopez and Romary 2010) construct a small corpus by exploiting the citation structure and IPC metadata. They then perform the retrieval on this initial corpus. In (Magdy and Jones 2010a) generate the query out of the most frequent unigrams and bigrams. In this work the effect of using bigrams in query generation was studied but the retrieval improvement was not significant. This is perhaps because of the unusual vocabulary usage in the patent domain.

So far, one of the most comprehensive descriptions of the problem and possible solutions for the prior art search is presented by Magdy et al. (2010). The authors showed that the best performing run of CLEF-IP 2010 (Lopez and Romary 2010) uses citations extracted by training a Conditional Random Field (CRF). The second best run (Magdy and Jones 2010a) used a list of citations extracted from the patent numbers within the description field of patent queries. They also showed that the best run employed sophisticated methods of retrieval using two complementary indices, one constructed by extracting terms from the patent collection and the other built from external resources such as Wikipedia. They compared this two approaches and concluded with an interesting observation that the second best run achieves a statistically indistinguishable performance compared to the best run.

A recent study (Ganguly et al. 2011) studies the effect of using Pseudo Relevance Feedback (PRF) for reducing patent queries. The authors decompose a patent application into constituent text segments and compute language modeling similarities by calculating the probability of generating each segment from the top ranked documents. Another work (Mahdabi and Crestani 2012) employs a learning to rank based framework for estimating the effectiveness of a document in terms of its performance in PRF. They use the knowledge of effective feedback documents on past queries to estimate effective feedback documents for new queries. They introduced features correlated with feedback document effectiveness. A recent work on query expansion used a learning-based approach to predict the quality of a query (Mahdabi et al. 2012). This work uses noun phrases for query expansion from the query patent document. Because cited documents can be a good source for extracting noun phrases, our citation based measure for estimating the importance of a citation document can be integrated as a feature in their learning-based model to improve their performance.

2.2 Citation analysis

We now explain in more detail the previous approaches that is more in line with our experiments in this paper. In (Fujii 2007; Lopez and Romary 2010) the authors focused on studying the link-based algorithms by using citations as links between documents with the goal of improving the ranking of the documents.

[3] http://www.ir-facility.org/clef-ip.

Fujii applies link analysis techniques to the graph structure of patent documents (Fujii 2007) and uses the citation links for re-ranking an initially retrieved list. He computes a composite score based on textual information and citation links. The cited paper, shows that ranking based on this composite score improves upon the ranking based on the textual-derived score alone. Fujii also uses the citation link structure of patent documents to measure the influence of each patent document, developing two distinct methods. In the first method he calculates the PageRank score (Brin and Page 1998) for each document by considering the graph structure of all documents in the collection. This method is not specific to the query. In the second method, he computes the PageRank score for a query-specific citation graph, which is composed of the top-k documents initially retrieved for a given query topic and their cited documents. His experimental results on the NTCIR-6 test collection demonstrate that the citation analysis is helpful for invalidity patent search and the query-specific PageRank score is more effective than the traditional PageRank score. As a baseline for this paper, we implemented the work of Fujii on the CLEF-IP 2011 collection. Similar to his work, we use the PageRank measure on a query-specific citation graph to calculate a score for quantifying the authoritativeness of each document.

Lopez and Romary use citation information in a different way (Lopez and Romary 2010). They extract all patent and non patent literature references in the collection using a Conditional Random Field model on an annotated corpus. They estimate the importance of each key term in a supervised manner. They use key terms selected by authors and readers and feed them into a bagged decision tree. Our work in this paper is different from this work, as our approach is completely unsupervised and we do not have any annotated tag terms.

2.3 An evaluation metric for patent retrieval

In addition to the well known MAP and Recall metrics, we report the effectiveness of our proposed method in terms of the Patent Retrieval Evaluation Score (PRES) (Magdy and Jones 2010b). This metric is a modification over the well known IR evaluation metric called Normalized Recall (R_{norm}) (Rocchio 1964; van Rijsbergen 1979). R_{norm} measures the effectiveness in ranking documents relative to the best and worst ranking case, where the best ranking case is the retrieval of all relevant documents at the top of the list, and the worst case is the retrieval of all relevant documents only after retrieving the full collection. R_{norm} is calculated as the area between the actual and worst cases divided by the area between the best an worst cases. Normalized recall is greater when relevant documents are retrieved earlier in the ranked list thus it can be seen as a good representative measure for recall-oriented applications. However, the disadvantage of the normalized recall is related to the fact that it requires ranking the full collection which might not be feasible for very large collections.

In order to address this problem, Magdy and Jones (2010b) propose a modification of the calculation of R_{norm}. They suggest an approximation of the worst case scenario by considering any relevant document not retrieved in the top N_{max} to be ranked at the end of the collection. The new assumption for the worst case scenario is to retrieve all the relevant documents just after the maximum number of documents to be checked by the user, N_{max}. PRES uses this new assumption for the worst case scenario and the following equation shows how PRES is calculated.

$$PRES = 1 - \frac{\frac{\sum r_i}{n} - \frac{n+1}{2}}{N_{max}} \qquad (1)$$

where N_{max} is the number of documents to be checked by the user (cut-off value), n is the number of relevant documents, and $\sum r_i$ is the summation of ranks of relevant documents, which is shown in the following:

$$\sum r_i = \sum_{i=1}^{nR} r_i + nR(N_{max} + n) - \frac{nR(nR-1)}{2} \tag{2}$$

where R denotes the Recall value defined as the number of relevant and retrieved documents in the first N_{max} documents.

3 Establishing a baseline query

We now describe our approach to estimate a unigram query model from the query patent document. We first create a language model Θ_Q for the query patent:

$$P(t|\Theta_Q) = P_{ML}(t|D) \tag{3}$$

where the maximum likelihood estimate P_{ML} is calculated as follows:

$$P_{ML}(t|D) = \frac{n(t,D)}{\sum_{t'} n(t',D)} \tag{4}$$

where $n(t, D)$ denotes the term frequency of term t in document D.

We then introduce a unigram query model by estimating the importance of each term according to a weighted log-likelihood-based approach as expressed below:

$$P(t|Q_{orig}) = Z_t\, P(t|\Theta_Q)\ log\left(\frac{P(t|\Theta_Q)}{P(t|\Theta_C)}\right) \tag{5}$$

where $Z_t = 1/\sum_{t\in V} P(t|Q_{orig})$ is a normalization factor that is defined as the Kullback-Leibler divergence between Θ_Q and Θ_C. This approach favors terms that have high similarity to the document language model Θ_Q and low similarity to the collection language model Θ_C (Mahdabi et al. 2011). All sections of the query document are considered in this estimation.

We build a query by selecting the top k terms from Q_{orig}. This query is used to retrieve an initial ranked list of documents to build the root set. In the remainder of this paper we refer to Q_{orig} as the unigram baseline. In the next sections we explain how the original query is expanded utilizing citation links and the term distribution of documents in the citation graph. Figure 1 illustrates the general scheme of our proposed method of query expansion.

4 Query-specific citation graph

In this section we discuss the basics of the patent collection as a graph and explain how to build it.

In the CLEF-IP 2011 dataset, the citations of query topics have been removed by the organizers and used for building the relevance judgments. However, we have access to the citations of all other documents apart from the query topics in the collection. A recent work (Mahdabi et al. 2011) tried to extract these citations with regular expressions but the

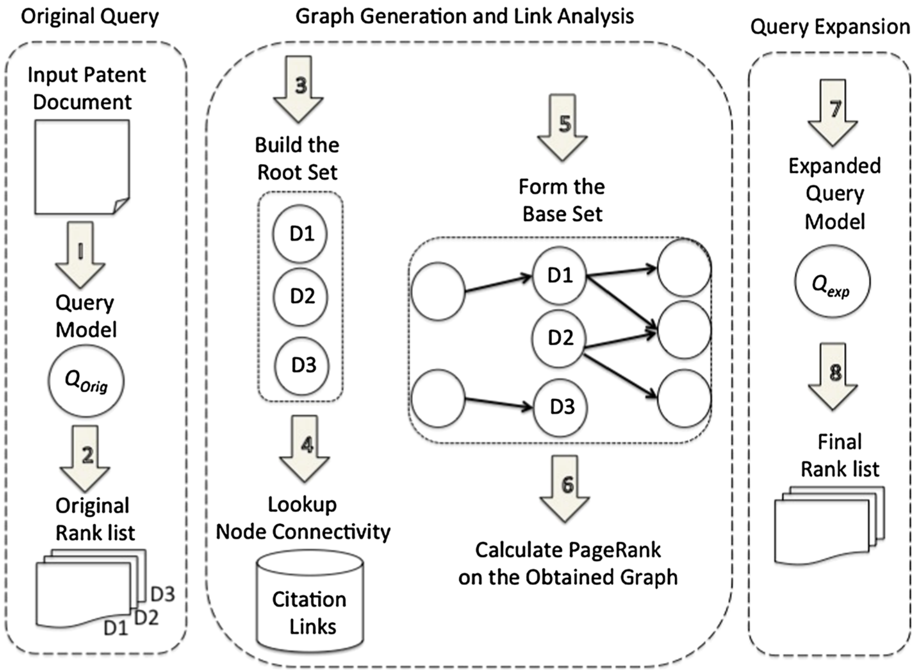


Fig. 1 The general scheme of our proposed method for query expansion using citation information. *Numbers* indicate the sequence flow of operations

reported accuracy was low. Another previous work (Lopez and Romary 2009) used a web service offered by the European Patent Office (EPO)[4] to retrieve the citations of documents in the collection. We used this web service to extract all the citations of the documents in the collection with the exception of the query documents. We extracted the citation links and stored them in a relational database.

4.1 Building the query-specific citation graph

As previous work (Fujii 2007) suggests, computing PageRank values as a measure of static quality of documents in the collection (calculated independently of any query a system might receive) has a clear disadvantage compared to conditioning the computation of PageRank values on the query being served. Thus we will focus on how to assemble a subset of patent documents around the topic of the query, from the graph induced by their citation links. By doing so we are able to derive PageRank values relative to particular queries.

Our approach is inspired by the technique used by the HITS algorithm (Kleinberg 1999), where a small subgraph of the entire web related to the query (as opposed to the whole web graph) is chosen for estimating the importance of a webpage.

We build such a graph by gathering a subset of linked documents in the patent collection related to a query following the two steps below:

[4] http://www.epo.org/searching/free/ops.html.

1. Given a patent query, we perform a search and we retrieve an initial ranked list of documents. We take the top-n documents from this list and call this the *root set* of documents.
2. We construct the *base set* of documents, by including the root set as well as any document that either cites a document in the root set, or is cited by a document in the root set. Lookups are performed to retrieve node and connectivity information from the citation links stored as a relational database.

4.2 Citation analysis of the graph structure

In our query-specific citation graph, each vertex is a patent document and there is an edge between two vertices if one of the vertices cites the other. The assumption is that if patent p is cited by patent q, then the author of patent q is implicitly saying that patent p is somehow important to q. The basic idea is that each citation link from patent document q to patent document p can be seen as an endorsement of document p. Therefore, the vocabulary usage of patent p might be useful to bridge the gap between the query q and its relevant documents. We now compute the query-specific PageRank values for all nodes in the citation graph to estimate the importance of each node.

The PageRank value of the document p is determined by the sum of the votes from other documents in the citation graph. The computation of the PageRank value for any document p is performed based on the following equation:

$$PR(p) = \sum_{q \in D_{* \to p}} \frac{PR(q)}{D_{q \to *}} \tag{6}$$

where $D_{* \to p}$ is the set of patent documents that cites p, and $D_{q \to *}$ is the set of patent documents cited by q. If p is cited by a large number of documents, a high score is given to p. However, if a document cites n documents, the vote for each cited documents is divided by n. This means that the influence from citing documents is shared among the documents it cites.

We calculate the PageRank values for all the documents in our query-specific citation graph iteratively. We start by assuming equal PageRank values for all the nodes in the citation graph. This value is set to $\frac{1}{N_{G_{cit}}}$, where $\frac{1}{N_{G_{cit}}}$ denotes the size of the graph. We then perform multiple iterations of this calculation until PageRank values converge to the final values.

The calculated PageRank values are used to guide the priority assignment to documents while estimating a query model from citations. This procedure is described in Sect. 5

5 Query expansion guided by the citation-based measure

Our approach for query modeling aims to improve the language model of the original query by using the term distribution of documents in the citation graph. The key assumption of this paper is that documents with more importance in the citation graph are more likely to be relevant and thus term selection from them is more effective.

We identify and weight the most distinguishing terms in the documents in the citation graph and we use the calculated PageRank value as a document prior in a language modeling framework. This term sampling is performed as follows:

Table 1 Comparison between the list of terms derived from the patent query and the terms sampled from documents belonging to the query-specific citation graph for the patent application "EP-1832953-A2" belonging to CLEF-IP 2011 topic set, with title "method and apparatus for managing a peer-to-peer collaboration system"

Query document	Documents in the citation graph
manage, server, collaborate	transact, handle, service
client, soap, peer	access, command

$$P(t|Q_{cit}) = Z_d \frac{1}{N_{G_{cit}}} \sum_{D \in G_{cit}} P(t|D)P(D|G_{cit}) \tag{7}$$

where G_{cit} is the citation graph and N_{G_cit} is the number of documents in this graph. $P(D|G_{cit})$ is the probability of a document given the citation graph. We use the PageRank value of a document D, as previously explained in Sect. 4.2, to denote this probability. $Z_d = 1/\sum_{D \in G_{cit}} P(D|G_{cit})$ is a normalization factor.

We interpolate the citation query with the original query (as estimated in Eq. 5):

$$P(t|Q) = \lambda P(t|Q_{orig}) + (1-\lambda)P(t|Q_{cit}) \tag{8}$$

The M highest terms from the updated query model is then used as a query to retrieve a final ranked list of documents.

As an example, Table 1 shows the terms selected from two different information resources, namely the query itself and the documents in the citation graph. The terms are selected based on Eqs. 5 and 7, respectively.

6 Experimental setup

In order to answer the research questions listed in the first Section of the paper, we run a set of experiments. We now discuss our experimental setup.

We described the procedure of building a query-specific citation graph in Sect. 4 Generation of this graph is sensitive to the choice of the following parameters. The first parameter is top-k query terms, selected from the estimated unigram query model Q_{Orig}, which is used for retrieving an initial ranked list of documents. The second parameter is top-n documents selected from the initial ranked list to form the root set. The third parameter is top-m feedback terms extracted from the expanded query model, which is used for retrieving the final ranked list of documents. The fourth parameter is related to the citation depth that is considered while assembling the base set. The number of query terms used in the original query model is experimentally set to 100. The influence of the citation depth on the performance of the proposed method for query expansion is analysed in Sect. 7 We study the influence of the rest of the parameters in Sect. 8.

The value of λ for interpolation in Eq. 8 is empirically set to 0.5. We used the Language Modeling approach with Dirichlet smoothing (Zhai and Lafferty 2001) to score documents and build the initial and final ranked lists. We only calculate score for documents that have one IPC class in common with the query topic and not for the entire collection. We empirically set the value for the smoothing parameter to 1500.

We note that patent topics are unexamined patent applications and they are temporally prior to all other documents in the corpus. The task of prior art search is defined as a search backward in time, to find possible relevant documents. Therefore our query-specific citation graph includes documents that are backward in time compared to our query document. This ensures that topics (patent applications) could have not been cited inside the collection and thus the setting of our experiments utilizing citation links is valid.

6.1 Test set and pre-processing of the data

CLEF-IP 2011 contains 2.6 million patent documents pertaining to 1.3 million patents from the European Patent Office (EPO). These documents are extracts of the MAREC dataset. This documents are extended by documents from World Intellectual Property Organization (WIPO) and their content is available in English, German and French.

We used the Terrier Information Retrieval System[5] to index the CLEF-IP 2011 collection. We used the default stemming and stop-word removal. According to our experiments we obtained better results by removing terms with length shorter than 3 characters from the query. We also removed terms including numbers from the query. In our experiments we used the English subsection of the collection. The English topic set of CLEF-IP 2011 consists of 1351 topics.

6.2 Evaluation

We report the retrieval effectiveness of our proposed method in terms of Mean Average Precision (MAP), Recall and Patent Retrieval Evaluation Score (PRES) (Magdy and Jones 2010b). MAP and Recall are popular metrics used for search engines and are applied to report results at the Text Retrieval Conference (TREC)[6] and Cross Lingual Evaluation Forum (CLEF).[7] PRES (Magdy and Jones 2010b) is a modification of the well known R_{norm} metric. This metric measures the system recall and the quality of ranking in one single score. We used a Perl script,[8] provided by authors of (Magdy and Jones 2010b), for calculating these evaluation metrics.

To be consistent with the reports of CLEF-IP participants, we also report the evaluation results in terms of normalized Discounted Cumulative Gain (nDCG) and geometric Mean Average Precision (gm-map). nDCG measures the usefulness of a document based on its grade of relevance and position in the ranked list. The gm-map measure is designed for situations where one is interested to highlight the improvements of the low-performing topics. We used trec-eval for calculating these two evaluation metrics.

7 Experimental results

In this section we describe the experiments that we conducted to evaluate the usefulness of our proposed method, present their results and formulate answers to the research questions.

We now describe the structure of our experiments. We compare three methods using the CLEF-IP 2011 corpus. The first method is related to the original query we refer to which as the baseline method. Our original query model, which was explained in Sect. 3, is built

[5] See: http://ir.dcs.gla.ac.uk/terrier/.

[6] http://trec.nist.gov/.

[7] http://www.clef-initiative.eu.

[8] http://www.computing.dcu.ie/~wmagdy/PRES.htm.

from the query document. The second method corresponds to our implementation of the work reported in (Fujii 2007). This method is focused on computing a composite score using the textual information of the query together with the link-based structure of the query-specific citation graph. This method is referred to as *Score-cit*. The last method is our proposed model which estimates a query model from the documents in the citation graph and expands the original query using the estimated model from the term distribution of the documents in the citation graph. This method is referred to as *QM-cit*.

We study the influence of the size of the citation graph on the effectiveness of query expansion by considering two alternative versions of Score-cit and QM-cit. The first version considers a citation graph exploiting one level depth of citation links, constructed by collecting documents in the root set and base set as explained in Sect. 4 We call these methods Score-cit1 and QM-cit1. The second variation takes into account a citation graph using two levels of citation links. We refer to the methods in this category as Score-cit2 and QM-cit2.

Table 2 shows the evaluation results of different methods of our experiments using the CLEF-IP 2011 dataset. Results marked with † achieved statistically significant improvement over the baseline, while ‡ represents a statistical significant difference compared to Score-cit1 and Score-cit2. The reported statistical difference is calculated using t test and has a *p*-value of 0.05. The reported results for QM-cit1 and QM-cit2 are obtained using the top 100 feedback terms selected from the expanded query model. Top 30 feedback documents are selected and used to generate the root set.

The results of Table 2 suggest that neither of the versions of Score-cit method achieves statistical significance over the baseline. However, we can see that QM-cit1 and QM-cit2 achieve statistical significant difference compared to the baseline in terms of recall without decreasing the precision. Comparing the performance of QM-cit1 and QM-cit2 with Score-cit1 and Score-cit2 allows us to answer our first research question by concluding that using the link-based structure of the citations together with their textual content is more useful than using the link-based structure alone.

By comparing the performance of the proposed methods (QM-cit1 and QM-cit2) with the baseline, we can also observe that the proposed methods achieved a better performance in comparison to the baseline. In other words, using the term distribution of the documents in the citation graph enabled us to estimate a query model which improves over the language model of the original query. This is due to the fact that the estimated query from the cited documents complements the original query (made from the patent application) and alleviates the term mismatch between the query document and documents relevant to it. This observation answers our second research question.

Table 2 Performance of different methods over CLEF-IP 2011 dataset with a cut-off value of 1,000

CLEF-IP 2011 test set				
Method	Run description	MAP	Recall	PRES
baseline	-	0.099	0.540	0.450
Score-cit1	citation depth level 1	0.091	0.543	0.453
Score-cit2	citation depth level 2	0.095	0.550	0.459
QM-cit1	citation depth level 1	0.105	0.560 †	0.465
QM-cit2	citation depth level 2	0.105	0.579 †‡	0.481 †‡

As shown in the explanation of the experiments, increasing the depth of the citation graph (from depth 1 to depth 2) has a positive effect on the performance of both Score-cit and QM-cit methods. We also carried out experiments with a citation graph of depth 3, where 3 consecutive iterations of the steps described in Sect. 4 are considered. The obtained performance is statistically indistinguishable from the results for Score-cit2 and QM-cit2 presented in Table 2. We therefore did not present these results here.

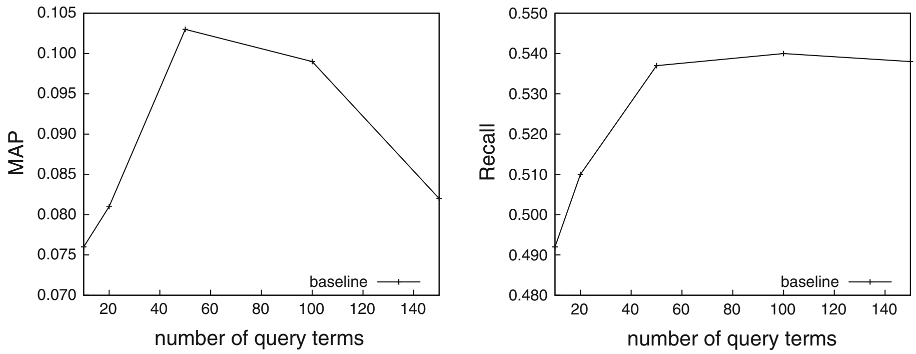
We hypothesize that the expansion of the root set into the base set as explained in Sect. 4, results in including documents in the base set that have different languages compared to the language of the query. This cross-language retrieval effect of the base root is very desirable for capturing patent documents relevant to the query but in other languages. In this paper, we ignore non English documents, as we only perform the search and indexing on the English subsection of the collection.

8 Sensitivity analysis of different parameter settings

As mentioned before, the reported results in Table 2 are obtained using the top 100 terms extracted from the expanded query model using the top 30 feedback documents. In this section, we conduct experiments to study the impact of these parameters on the retrieval effectiveness of our proposed method QM-cit2. We used test topics of CLEF-IP 2011 during these evaluations.

8.1 Effect of the number of query terms

We study the impact of the number of query terms selected from the original query model on the retrieval effectiveness of the baseline method. Figure 2 shows the results of this study. We can observe that by increasing the number of query terms we achieve improvement in terms of recall. In Fig. 2, the best performance in terms of recall is achieved when the number of query terms is around 100. On the other hand, when selecting more than 100 query terms MAP drops.



(a) MAP@1000 of baseline

(b) Recall@1000 of baseline

Fig. 2 Sensitivity analysis of the baseline method to the number of query terms selected from the original query model on CLEF-IP 2011

8.2 Effect of the number of feedback terms

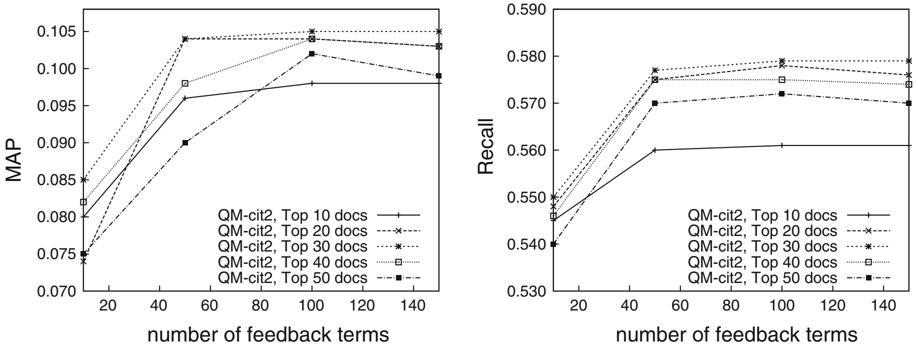
We run QM-cit2 by varying the number of feedback terms from 10 to 150. Table 3 shows the effect of these parameters on the performance of the system in terms of MAP, Recall, PRES, nDCG and gm-map at cut-off value of 1000. Results marked with † achieved statistically significant improvement over the baseline. We observe that QM-cit2 achieves the best performance, selecting 100 feedback terms, regardless of the number of feedback documents. In Table 3, selecting more than 100 query terms does not lead to an improvement. We notice the positive effect of increasing the number of expansion terms on all the evaluation metrics.

We report the performance of QM-cit2 method considering MAP and Recall metrics in Fig. 3.

Tables 4 and 5 report the results of QM-cit2 method in terms of MAP, Recall, PRES at cut-off value of 100 and 500, respectively. The results of Tables 4 and 5 are consistent with the observations made from Table 3.

Table 3 QM-cit2 results over CLEF-IP 2011 dataset with a cut-off value of 1,000

Feedback terms	Metric	Feedback documents				
		10	20	30	40	50
10	MAP	0.080	0.074	**0.085**	0.082	0.075
	Recall	0.545	0.548	**0.550**	0.546	0.540
	PRES	0.445	0.445	**0.450**	0.445	0.440
	nDCG	0.200	0.203	**0.236**	0.227	0.211
	gm-map	0.010	0.010	**0.015**	0.013	0.011
20	MAP	0.082	0.082	**0.097**	0.082	0.076
	Recall	0.549	0.551	**0.564**	0.551	0.539
	PRES	0.447	0.451	**0.467**	0.451	0.440
	nDCG	0.210	0.232	**0.254**	0.232	0.223
	gm-map	0.011	0.013	**0.019**	0.013	0.013
50	MAP	0.096	**0.104**	**0.104**	0.098	0.090
	Recall	0.560	0.575	**0.577**	0.575	0.570
	PRES	0.463	0.479	**0.480**	0.479	0.476
	nDCG	0.221	0.250	**0.264**	0.250	0.232
	gm-map	0.013	0.018	**0.022**	0.018	0.014
100	MAP	0.098	0.104	**0.105**	0.104	0.102
	Recall	0.561	0.578 †	**0.579** †	0.575	0.572
	PRES	0.465	**0.481**	**0.481**	**0.481**	0.480
	nDCG	0.225	**0.252**	0.251	0.250	0.235
	gm-map	0.014	0.018	**0.019**	0.018	0.015
150	MAP	0.098	0.103	**0.105**	0.103	0.099
	Recall	0.561	0.576	**0.579** †	0.574	0.570
	PRES	0.465	0.479	**0.481**	0.479	0.477
	nDCG	0.225	0.250	**0.251**	0.250	0.235
	gm-map	0.013	0.018	**0.019**	0.018	0.015



(a) MAP@1000 of QM-cit2

(b) Recall@1000 of QM-cit2

Fig. 3 Sensitivity analysis of QM-cit2 to the number of feedback terms on CLEF-IP 2011

Table 4 Recall, MAP and PRES results over CLEF-IP 2011 dataset with a cut-off value of 100

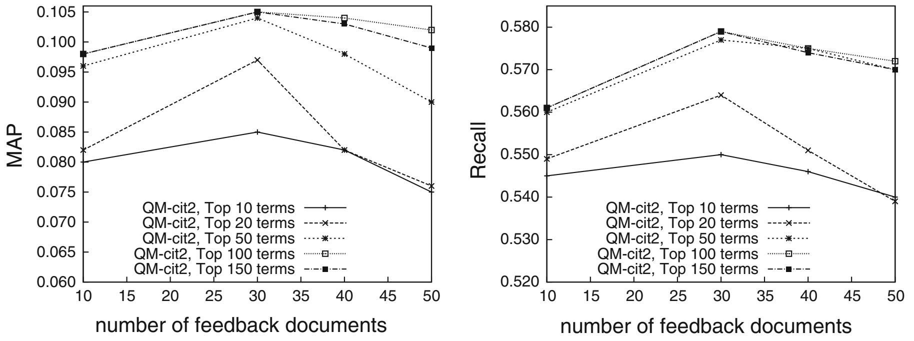
Feedback terms	Metric	Feedback documents				
		10	20	30	40	50
10	MAP@100	0.076	0.077	0.078	0.082	0.078
	Recall@100	0.300	0.310	0.310	0.313	0.314
	PRES@100	0.221	0.227	0.227	0.230	0.228
20	MAP@100	0.077	0.080	0.080	0.082	0.078
	Recall@100	0.310	0.324	0.325	0.315	0.328
	PRES@100	0.226	0.245	0.245	0.2327	0.247
50	MAP@100	0.078	0.080	0.080	0.084	0.080
	Recall@100	0.327	0.342	0.342	0.354	0.344
	PRES@100	0.247	0.257	0.257	0.262	0.257
100	MAP@100	0.079	0.081	0.081	0.090	0.082
	Recall@100	0.329	0.344	0.344	0.354	0.342
	PRES@100	0.250	0.258	0.258	0.266	0.258
150	MAP@100	0.079	0.081	0.081	0.090	0.082
	Recall@100	0.329	0.344	0.344	0.353	0.342
	PRES@100	0.250	0.258	0.258	0.266	0.258

8.3 Effect of the number of feedback documents

We investigate the effect of the number of feedback documents by varying this number from 10 to 50. We plot the sensitivity of QM-cit2 method for varying values of feedback documents in Fig. 4. We observe the best performance is achieved when the number of feedback documents is around 30. We can see that values higher that 30 hurt the performance.

Table 5 Recall, MAP and PRES results over CLEF-IP 2011 dataset with a cut-off value of 500

Feedback terms	Metric	Feedback documents				
		10	20	30	40	50
10	MAP@500	0.079	0.078	0.080	0.081	0.082
	Recall@500	0.476	0.482	0.483	0.489	0.489
	PRES@500	0.374	0.380	0.381	0.389	0.389
20	MAP@500	0.080	0.080	0.082	0.082	0.083
	Recall@500	0.495	0.497	0.497	0.490	0.492
	PRES@500	0.394	0.398	0.398	0.389	0.392
50	MAP@500	0.084	0.084	0.086	0.087	0.087
	Recall@500	0.495	0.505	0.505	0.492	0.494
	PRES@500	0.398	0.401	0.401	0.398	0.398
100	MAP@500	0.960	0.977	0.977	0.960	0.960
	Recall@500	0.494	0.508	0.508	0.494	0.493
	PRES@500	0.405	0.411	0.411	0.405	0.405
150	MAP@500	0.968	0.978	0.978	0.968	0.968
	Recall@500	0.494	0.508	0.508	0.494	0.494
	PRES@500	0.407	0.411	0.411	0.407	0.407



(a) MAP of QM-cit2 on CLEF-IP 2011 **(b)** Recall@1000 of QM-cit2 on CLEF-IP 2011

Fig. 4 Sensitivity analysis of QM-cit2 to the number of feedback documents on CLEF-IP 2011

9 Conclusions and future work

Previous work showed that using the link-based structure of the citations leads to improvements over a strictly textual-based method (using the term distribution of the query document). It remained to investigate whether the link-based structure of the citation graph together with the term distribution of cited documents can be effective to improve the ranking. To answer this question, we introduced a query model built from the citation graph. This query model provides a principled way to calculate the importance of terms selected from the linked documents.

We analyzed the effectiveness of this query model on the CLEF-IP 2011 test collection. The results demonstrated significant improvements in terms of recall, without decreasing precision. The results showed the advantage of using the term distribution of the cited documents for query expansion.

As future work we could quantify the language model of different information resources such as classifications and the citation graph to perform a comparison. This would enable us to use characteristics of these vocabularies to better estimate a unified query model composed from all of these resources.

An interesting extension to this work could be to use the publication date tags of the patent documents in the citation graph to detect the change in the vocabulary over time. The importance of each document (taken from the citation graph) in estimating the query model can be discounted based on its time difference to the query document. We plan to investigate these directions in the future.

Acknowledgments Authors would like to thank the reviewers for the very useful comments. We also thank Morgan Harvey for proofreading this manuscript.

References

Atkinson, K. H. (2008). Toward a more rational patent search paradigm. In J. Trait (Ed.), *Proceedings of the 1st ACM workshop on Patent Information Retrieval (PaIR 2008)* (pp. 37–40), Napa Valley, CA, 30 October 2008. ACM.

Baeza-Yates, R. A., & Ribeiro-Neto, B. A. (2011). *Modern information retrieval—The concepts and technology behind search, Second edition*. Harlow, England: Pearson Education Ltd.

Bashir, S., & Rauber, A. (2009). Improving retrievability of patents with cluster-based pseudo-relevance feedback documents selection. In D. W. -L. Cheung, I. -Y. Song, W. W. Chu, X. Hu, & J. J. Lin (Eds.), *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)* (pp. 1863–1866), Hong Kong, China, 2-6 November 2009. ACM.

Bashir, S., & Rauber, A. (2010). Improving retrievability of patents in prior-art search. In C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, S. Rüger, & K. van Rijsbergen, (Eds.), *32nd European Conference on IR Research (ECIR 2010)* (Vol. 5993, pp. 457–470), Milton Keynes, UK, 28-31 March 2010.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, *30*(1–7), 107–117.

Fujii, A. (2007). Enhancing patent retrieval by citation analysis. In W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, & N. Kando (Eds.), *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007)* (pp. 793–794), Amsterdam, The Netherlands, 23-27 July 2007. ACM.

Fujii, A., Iwayama, M., & Kando, N. (2004). Overview of patent retrieval task at NTCIR-4. In N. Kando & H. Ishikawa (Eds.), *NTCIR Workshop: Proceedings of the Fourth NTCIR Workshop Research in Information Access Technologies: Information Retrieval, Question Answering and Summarization*. Tokyo, Japan, April 2003–June 2004. NII.

Fujii, A., Iwayama, M., & Kando, N. (2007). Introduction to the special issue on patent processing. *Information Processing Management*, *43*(5), 1149–1153.

Fujita, S. (2004). Revisiting the document length hypotheses- NTCIR-4 CLIR and patent experiments at Patolis. In *Proceedings of NTCIR-4 Workshop*.

Ganguly, D., Leveling, J., Magdy, W., & Jones, G. J. F. (2011). Patent query reduction based on pseudo-relevant documents. In C. Macdonald, I. Ounis, & I. Ruthven (Eds.), *Proceedings of the 20th ACM Conference on Information and Knowledge Management (CIKM 2011)* (pp. 1953–1956), Glasgow, UK, 24-28 October 2011. ACM.

Iwayama, M., Fujii, A., Kando, N., & Takano, A. (2003). Overview of the third NTCIR workshop. In M. Iwayama & A. Fujii (Eds.), *Proceedings of the ACL-2003 workshop on patent corpus processing* (pp. 24–32).

Joho, H., Azzopardi, L. A., & Vanderbauwhede, W. (2010). A survey of patent users: an analysis of tasks, behavior, search functionality and system requirements. In D. Kelly & N. J. Belkin (Eds.), *Proceedings of the third symposium on information interaction in context (IIiX)* (pp. 13–24). ACM.

Kleinberg, J.M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, *46*(5), 604–632.

Lopez, P., & Romary, L. (2009). Patatras: Retrieval model combination and regression models for prior art search. In C. Peters, G. M. Di Nunzio, M. Kurimo, T. Mandl, D. Mostefa, A. Peñas, & G. Roda (Eds.), *Proceedings of CLEF (Notebook Papers/LABs/Workshops)* (pp. 430–437). Lecture Notes in Computer Science.

Lopez, P., & Romary, L. (2010). Experiments with citation mining and key-term extraction for prior art search. *CLEF (Notebook Papers/LABs/Workshops).*

Lupu, M., & Hanbury, A. (2013). Patent retrieval. *Foundations and Trends in Information Retrieval*, *7*(1), 1–97.

Lupu, M., Mayer, K., Tait, J., & Trippe, A. (2011). *Current challenges in patent information retrieval.* Berlin:Springer.

Magdy, W., & Jones, G. J. F. (2010a). Applying the KISS principle for the CLEF-IP 2010 prior art candidate patent search task. In M. Braschler, D. Harman, & E. Pianta (Eds.), *CLEF (Notebook Papers/LABs/Workshops).* Lecture Notes in Computer Science.

Magdy, W., & Jones, G. J. F. (2010b). PRES: A score metric for evaluating recall-oriented information retrieval applications. In *Proceedings of ACM SIGIR conference on research and developement in information retrieval* (pp. 611–618).

Magdy, W., Leveling, J., & Jones, G. J. F. (2009). Exploring structured documents and query formulation techniques for patent retrieval. In C. Peters, G. M. Di Nunzio, M. Kurimo, T. Mandl, D. Mostefa, A. Peñas, & G. Roda (Eds.), *CLEF* (pp. 410–417). Lecture Notes in Computer Science.

Magdy, W., Lopez, P., & Jones, G. J. F. (2010). Simple vs. sophisticated approaches for patent prior-art search. In C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, S. Rüger, & K. van Rijsbergen (Eds.), *32nd European Conference on IR Research,(ECIR 2010)* (Vol. 5993, pp. 725–728), Milton Keynes, UK, 28–31 March 2010.

Mahdabi, P., Andersson, L., Hanbury, A., & Crestani, F. (2011). Report on the CLEF-IP 2011 experiments: Exploring patent summarization. In A. Hanbury, A. Rauber, & A. P. de Vries (Eds.), *CLEF (Notebook Papers/Labs/Workshop).* Lecture Notes in Computer Science.

Mahdabi, P., Andersson, L., Keikha, M., & Crestani, F. (2012). Automatic refinement of patent queries using concept importance predictors. In W. R. Hersh, J. Callan, Y. Maarek, & M. Sanderson (Eds.), *Proceedings of ACM SIGIR conference on research and developement in information retrieval* (pp. 505–514). ACM.

Mahdabi, P., & Crestani, F. (2012). Learning-based pseudo-relevance feedback for patent retrieval. In M. Salampasis & B. Larsen (Eds.), *Proceedings of information retrieval facility conference (IRFC)* (pp. 1–11). Lecture Notes in Computer Science.

Mahdabi, P., Keikha, M., Gerani, S., Landoni, M., & Crestani, F. (2011). Building queries for prior-art search. In *Proceedings of information retrieval facility conference (IRFC)* (pp. 3–15).

Mase, H., Matsubayashi, T., Ogawa, Y., & Iwayama, M. (2005). Proposal of two stage patent retrieval method considering the claim structure. *ACM Transaction on Asian Language Information Processing*, *4*(2), 190–206.

Piroi, F., & Tait, J. (2010). CLEF-IP 2010: Retrieval experiments in the intellectual property domain. In *CLEF-2010 (Notebook Papers/LABs/Workshops).*

Rocchio, J. (1964). Performance indices for document retrieval systems. In *Report ISR-8, The Computation Laboratory of Harvard University.*

Takaki, T., Fujii, A., & Ishikawa, T. (2004). Associative document retrieval by query subtopic analysis and its application to invalidity patent search. In D. A. Grossman, L. Gravano, C. Zhai, O. Herzog, & D. A. Evans (Eds.), *ACM conference on information and knowledge management (CIKM)* (pp. 399– 405). ACM.

van Rijsbergen, C. J. (1979). *Information retrieval.* London: Butterworths.

Xue, X., & Croft, W. B. (2009). Transforming patents into prior-art queries. In J. Allan, J. A. Aslam, M. Sanderson, C. Zhai, & J. Zobel (Eds.), *Proceedings of ACM SIGIR conference on research and developement in information retrieval* (pp. 808–809). ACM.

Zhai, C., & Lafferty, J. D. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. In W. B. Croft, D. J. Harper, D. H. Kraft, & J. Zobel (Eds.), *Proceedings of ACM SIGIR conference on research and developement in information retrieval* (pp. 334–342). ACM.