Patent Query Formulation by Synthesizing Multiple Sources of Relevance Evidence

PARVAZ MAHDABI and FABIO CRESTANI, University of Lugano

Patent prior art search is a task in patent retrieval with the goal of finding documents which describe prior art work related to a query patent. A query patent is a full patent application composed of hundreds of terms which does not represent a single focused information need. Fortunately, other relevance evidence sources (i.e., classification tags and bibliographical data) provide additional details about the underlying information need. In this article, we propose a unified framework that integrates multiple relevance evidence components for query formulation. We first build a query model from the textual fields of a query patent. To overcome the term mismatch, we expand this initial query model with the term distribution of documents in the citation graph, modeling old and recent domain terminology. We build an IPC lexicon and perform query expansion using this lexicon incorporating proximity information. We performed an empirical evaluation on two patent datasets. Our results show that employing the temporal features of documents has a precision enhancing effect, while query expansion using IPC lexicon improves the recall of the final rank list.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Query formulation*

General Terms: Experimentation, Performance, Measurement

Additional Key Words and Phrases: Patent search, query expansion, proximity, citation analysis

ACM Reference Format:
Parvaz Mahdabi and Fabio Crestani. 2014. Patent query formulation by synthesizing multiple sources of relevance evidence. ACM Trans. Inf. Syst. 32, 4, Article 16 (September 2014), 30 pages.
DOI: http://dx.doi.org/10.1145/2651363

1. INTRODUCTION

A patent is a legal document, granted by a country's patent office, that gives a set of rights of exclusivity and protection to the owner of an invention. In order to be granted a valid patent, an invention needs to meet certain criteria such as novelty, that is, it should not have been previously patented by someone else, described in a scientific paper, or disclosed to public through any other medium. A patent examiner has to perform a search over previously published patents and non-patent data with the aim of verifying whether the idea of a patent application is novel. This type of search is called *prior art search* and is also referred to as *patentability* or *novelty* search. The objective of this 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].

There are other types of search processes in the patent domain, such as *technology survey*, *freedom to operate*, *validity*, and *patent portfolio search*. These search processes differ in terms of the information need of the searcher, the corpora used, and the

This article is an extension of a paper that appeared in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* and a manuscript accepted for publication in the *Journal of IR*.
Authors' addresses: P. Mahdabi and F. Crestani, Faculty of Informatics, University of Lugano, Switzerland; corresponding author's email: parvaz.mahdabi@usi.ch.

DOI: http://dx.doi.org/10.1145/2651363

output of the search. Notice, however, that the precise names and definitions of these search processes vary between those who deal with patents, for example, information specialists, private patent searchers, patent examiners, and patent lawyers [Lupu and Hanbury 2013]. In the remainder of this article, we focus our attention on prior art search which is a critical step in the examination process of a patent application.

Prior art search is executed by patent examiners using keyword-based searches from the claims field of a patent application after its filing. These keyword-based searches are completed using other metadata associated with the patent application, such as IPC classes[1] (International Patent Classification) and date tags. Bibliographic information such as backward and forward citations[2] can also be used to perform prior art searches. These searches made from different sources are then merged to compose a unique rank list. The goal of combining these complementary searches is to solve the term mismatch problem which is due to the obscure style of writing a patent ("patentese") and often leads to low retrieval effectiveness [Lupu et al. 2011].

Prior art search is a *recall-oriented application*. This search can take a very long time, as the searcher needs to ensure that he is not missing any relevant document, because infringing on existing patents might result in costly lawsuits.

Patent search covers multiple subject areas, such as chemistry, mechanical engineering, electrical engineering, and practically all other domains of industry applicable human knowledge [Lupu and Hanbury 2013]. Thus, it can be seen as a generalization of other domain-specific IR tasks such as health-care, biomedical, and chemical domains [Sondhi et al. 2012; Yin et al. 2010]. Patent search inherits the following problems of domain-specific IR tasks: the frequent usage of nonstandardized acronyms which are invented by patent applicants, the presence of homonyms (the same word referring to two or more different entities, such as bus[3] and closet[4]), the presence of synonyms (two or more words referring to the same entity such as signal and wave), and paraphrasing ("picture-taking device" used to describe a camera).

Documents relevant to a given query may not contain the exact terms used by the patent author, which are given to a search system as specific query terms. This problem is called *term mismatch*. In this article, our aim is to address the term mismatch problem in patent retrieval through query expansion. In order to cope with this problem, we first expand the initial query with the terms appearing in the cited documents. Using citation information, we derive the word usage of the community of inventors related to the topic of the query. To this end, we use the citation links, term distribution, and temporal features of cited documents to expand our initial query.

Previous work used knowledge bases such as Wikipedia and WordNet [Lopez and Romary 2010; Magdy and Jones 2011] for query expansion. We think that using a domain-dependent resource might help to extract more relevant expansion concepts compared to Wikipedia and WordNet. Thus, we intend to use a domain-dependent resource. To do this, we take IPC definition pages[5] and construct a lexicon from them. We then extract expansion terms specific to the IPC classes of the query patent from the lexicon and use proximity information to calculate importance weights for them.

We propose a proximity-based query propagation method to calculate the query term density at each point in the document. Our proximity-based framework rewards

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

[2] Forward citations denote the citations to a query patent application from patents which are *forward in time* from the query patent. In contrast, backward citations indicate the citations made by query patent to patents which are *backward in time* with respect to it.

[3] (i) Motor vehicle, (ii) an electronic subsystem transferring plurality of digits bits in group.

[4] (i) Water closet (flush toilet), (ii) a small cupboard used for storing things.

[5] `http://web2.wipo.int/ipcpub/`.

expansion concepts occurring close to query terms by using positional information while estimating the importance of expansion concepts. We hypothesize that this way, we are able to focus on expansion terms associated with query terms and avoid topic drift.

Our proposed model consists of four steps. In the first step, we estimate a query model from textual fields of a query patent. In the second step, we build a topic-specific citation graph and use the term distribution and temporal features of documents in the citation graph to estimate a citation query model which is used to expand our initial query model. In the third step, a query-specific lexicon is built. In the fourth and last step, query expansion is performed by deriving expansion concepts from the query-specific lexicon, and positional information is used to calculate weights for ensuring high-quality expansions. To this end, we utilize kernel functions to keep track of the distance of expansion concepts from query terms.

In this article, we seek to answer the following research questions.

—*RQ1*. Could we use the citation links, the content, and the temporal features of the cited documents to expand the initial query model built from a query patent?
—*RQ2*. How can we leverage the IPC classifications to construct a domain-dependent lexicon for query expansion?
—*RQ3*. How does the final rank list perform compared to the state of the art of prior art search approaches?

Our contributions are as follows.

—We present an approach to construct a domain-dependent lexicon for identifying query expansion concepts.
—We describe an approach for expanding the initial query model using a topic-sensitive graph built from the citation links.
—We describe an approach for exploiting the temporal features of documents in the citation graph for building a query model.
—We present a proximity-based query expansion method for estimating the probability that an expansion term is relevant to a query term.
—We investigate different query reformulation strategies for extracting concepts from a domain-dependent lexicon.

This article extends our previous work on leveraging conceptual lexicon for query expansion in patent retrieval [Mahdabi et al. 2013] by presenting a deeper analysis on the proposed models and a more exhaustive set of experiments on the collections. This paper also complements our other previous work [Mahdabi and Crestani 2013] on citation analysis for patent retrieval. This paper extends the previous paper as follows: (1) It presents a unified framework for query formulation retrieval synthesizing different relevance evidence sources associated with the query patent such as patent classifications and bibliographic information. (2) It introduces a new technique for utilizing the term distribution of cited documents by modeling the time information (3) It presents a more extensive set of experiments investigating the synthesis of different sources of relevance evidence on two datasets. (4) It provides detailed explanations and analyses of the results.

We evaluate our work on two patent collections, CLEF-IP 2010 and CLEF-IP 2011. The experimental results demonstrate that combining different sources of relevance evidence in a unified framework improves over using them separately. The results also show that query expansion using the term distribution and temporal features of documents in the citation graph leads to improving the precision of the rank list. The results confirm the advantage of deploying a domain-dependent resource for selecting expansion terms in contrast to Wikipedia and WordNet. Besides, the results demonstrate that utilizing proximity information leads to the calculation of reliable weights

for the expansion terms, and show consistent improvements in terms of recall in the final rank list.

The rest of the article is organized as follows. Section 2 describes the literature survey. Section 3 provides definitions of relevance evidence sources, presents the architecture of our proposed model, and explains the construction of an IPC lexicon. Section 4 describes the details of constructing a citation graph and explains a query expansion method which uses the citation links, the content, and the publication dates of the cited documents. Section 5 presents a framework for query expansion using the IPC lexicon. Section 6 and 7 present experimental settings and experimental results. We conclude in Section 8 with a summary and an outline of the future work.

2. RELATED WORK

Patent prior art search is composed of a search over previously filed patents and non-patent data with the aim of retrieving relevant documents which may invalidate or at least describe the prior art work in a patent application (henceforth referred to as *query patent*). The challenges of patent prior art search are different from those of standard ad hoc text and Web search. The first distinguishing property of prior art search is that the information need is presented by a patent document rather than short queries [Xue and Croft 2009a]. Another property is related to the overwhelming vocabulary mismatch which is due to the intentional obfuscation of content. For example, one patent document may contain few or no keywords in common with the query patent, but the idea conveyed in it might be quite similar or even identical to the query patent [Atkinson 2008]. The last property is linked to the structure of patents. Patents are structured documents with different fields such as abstract, description, and claims. Patent writers use different writing styles for describing the invention in different fields. For example, abstract and description fields use technical terminology while claims field uses legal jargon [Xue and Croft 2009a].

Among the mentioned challenges, we focus on the following three. The first challenge is to reduce a query patent in order to find a single focused information need and to remove the ambiguous and noisy terms. In previous work, researchers explored different fields of the query patent to perform query reduction [Xue and Croft 2009a; Cetintas and Si 2012]. Some of the previous work reported that effective queries were built from the entire query patent [Cetintas and Si 2012], while others obtained better results using single fields such as "background summary" [Xue and Croft 2009a]. It is worth mentioning that the "background summary" field is specific to U.S. patents.

The second challenge is related to query disambiguation. Previous work used different external resources for query expansion, such as Wikipedia [Lopez and Romary 2010] and WordNet [Magdy and Jones 2011] with the goal of query disambiguation. The goal here is to alleviate the term mismatch problem by expanding the query with topically related words or synonyms of the query terms.

The third challenge is related to the term mismatch problem. The language of patents contains highly specialized or technical words not found in everyday language [Joho et al. 2010]. Patent retrieval is often cumbersome and distinct from other information retrieval tasks. This is due to the inherent properties of patent content, namely, exceptional vocabulary, curious grammatical constructions, regulatory, and legal requirements [Atkinson 2008]. Patent authors purposely use 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 search systems.

We now explain the related work to this article. We first survey different approaches for query formulation in Section 2.1 and describe how the patent text and different metadata such as classification are used to build a query. We then describe different

approaches which exploit knowledge bases and proximity information in Section 2.2. Finally, in Section 2.3, we present different techniques which consider citation information.

2.1. Query Formulation for Patent Retrieval

The main wave of research in patent retrieval started after the third NTCIR workshop in 2003 [Iwayama et al. 2003], where a few test collections were released. Starting from the fourth NTCIR workshop in 2004 [Fujii et al. 2004], a search task was presented called "invalidity search run."[6] The goal was to find documents before the filing date of the application in question that conflict with the claimed invention. The citation parts of the applications are removed and counted as ground truth. Participants used different term weighting methods for query generation from the claims field.

Takaki et al. [2004] studied the rhetorical structure of a claim (an item in the claims field). 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. Similar work was introduced in Mase et al. [2005], where the authors analyzed the structure of the claims field to enhance retrieval effectiveness. The structure of each item of claims usually consists of the *premise* and *invention* parts, which describes existing and new technologies, respectively. The authors proposed a two-stage process where they first extract a query from the premise 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 both stages.

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 [2009b], who conducted a series of experiments in order to examine the effect of different patent fields on query formulation and concluded with the observation that the best Mean Average Precision (MAP) is achieved using the text from the "background summary" field of the query patent.

The current developments in patent search are driven by the Intellectual Property task within the CLEF[7] initiative. Several teams participated in 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 framework and employed the document structure of the patent documents in various ways [Piroi 2010]. Here, we only discuss in detail the two best-performing approaches in CLEF-IP 2010. Lopez Lopez and Romary [2010] constructed a small corpus by exploiting the citation structure and IPC metadata. They then performed the retrieval over this initial corpus. Magdy and Jones [2010a] generated the query out of the most frequent unigrams and bigrams. In this work, the effect of using bigrams in query generation studied, but the improvement was not significant, perhaps because of the unusual vocabulary usage in the patent domain.

So far, one of the most comprehensive descriptions of the problems and possible solutions for prior art search has been presented by Magdy et al. [2010]. The authors showed that the best-performing run of CLEF-IP 2010 [Lopez and Romary 2010] used 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

[6]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.

[7]`http://ifs.tuwien.ac.at/ clef-ip/.`

sophisticated retrieval methods 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 these approaches and concluded that the second-best run achieves a statistically indistinguishable performance compared to the best run when initial citations are provided with the query patent.

Classification Information. Many CLEF-IP and NTCIR participants have used classification information as an extra feature besides the content of the patent. Thus, a different range of methods for combining text content and classification information has been proposed. A standard way of combining the classification information is to consider it as a metadata and use it to filter the search results [Takaki et al. 2004; Gobeill et al. 2009; Teodoro et al. 2010; Verma and Varma 2011; Harris et al. 2011; Mahdabi et al. 2011]. This helps to filter out classifications that are too general or not related to the subject area of the query patent. Conclusive results are reported with respect to the usefulness of filtering using classification information. Fujita [2004] integrated IPC codes into a probabilistic retrieval model, employing the IPC codes for estimating the document prior. A different usage of IPC classification has been performed in D'hondt et al. [2011]. They used the classification information to extract query terms from triples specific to an IPC class. To do this, they used LCS software [Koster et al. 2003] which builds class profiles representing the term distribution (word and dependency triples) per IPC class. They created a sub-corpus per query document that contains documents with at least one IPC class in common with the query document. Classification information has been successfully used by Salampasis et al. [2012] in a different manner. They used classification information to partition the collection into different subject areas, and with this partitioning, they simulate a federated search for patent documents.

2.2. Leveraging Knowledge Bases and Proximity Heuristics

Previous research [Magdy and Jones 2011; Lopez and Romary 2010] tackled the term mismatch problem by first forming a keyword query from the query patent using the frequency information. The initial query is then expanded using a knowledge base such as Wikipedia or WordNet, exploiting this enhanced query to disambiguate the occurrences of query terms. The use of external resources has shown to be more effective compared to the use of the initial query and pseudo relevance feedback (PRF). In fact, the retrieval effectiveness of PRF in patent retrieval has been shown to be disappointing mainly due to the low MAP of the initial rank list [Ganguly et al. 2011].

Patent examiners use term proximity heuristics in their searches using the Boolean retrieval model in order to reward a document where the matched query terms occur close to each other. Two forms of adjacency operators are used in Boolean retrieval to address proximity: the "ADJn" operator, which searches for terms within a window of n words in the order specified, and the "NEARn" operator, which searches for the terms within a window of n words, in either order. This usage shows that proximity information plays an important role in patent searching.

Previous work [Magdy and Jones 2011; Lopez and Romary 2010] did not consider proximity information between query terms and expansion concepts while employing external resources for query expansion. Expansion terms extracted from these resources are often general terms. Thus, it is useful to condition their occurrence on their neighboring query terms and ignore their occurrence in isolation from any query term.

Our proximity-based framework is inspired by the work of Lv and Zhai on positional language model and positional relevance model [2009, 2010]. Lv and Zhai's work can capture passage-level evidence in a "soft" way by modeling proximity information via density functions. Their experiments confirmed that this approach works better than applying a "hard" boundary of passages.

Term position and proximity cues were mostly ignored in previous work in patent retrieval. Recently, Ganguly et al.'s work captured term positions and proximity evidences indirectly through the use of appropriate passages [2011]. This work provides a general model for query reduction using PRF.

A different approach has been proposed by D'hondt et al. [2011] that rewrites the query using Natural Language Processing (NLP) techniques. They extracted textual relations as triple dependencies from the title, abstract, and the first 400 words of the description field to enhance the query. Such dependencies are representations of grammatical relations between words in a sentence. They observed that adding triples to the query did not improve MAP scores, in comparison to a bag-of-word baseline, but had a positive effect on recall scores.

Another recent study on improving retrievability of patent documents [Bashir and Rauber 2010] combined term proximity heuristics with other features to select good query expansion terms in the context of PRF. In this work, different distance functions were considered from different windows surrounding query term occurrences. They reported an increase in terms of retrievability [Azzopardi and Vinay 2008] of individual patents using proximity heuristics compared to standard PRF. However, they did not evaluate directly the performance of their approach in terms of retrieval effectiveness.

A different approach is introduced by Calegari et al. [2012] which addresses the patent retrieval as an XML retrieval task. The authors encapsulate proximity information by introducing flexible constraints on the document structure (*near* and *below*) which produce a numerical score based on tag positions in the XML structure of patent documents. They calculate the similarity of a document to a query by taking advantage of the XML structure of patent documents together with document content. They showed that their approach achieved high recall and high precision by employing structure-based constraints, as opposed to most of the existing patent retrieval approaches which have good recall but suffer from low precision.

2.3. Citation Analysis

We now explain previous approaches that used citation information. Fujii [2007] applied Page Rank algorithm [Brin and Page 1998] on a graph created based on the citation link structure of patent documents. He developed two distinct methods for measuring the influence of a patent document on the citation graph. In the first method, he calculated the Page Rank score for each document by considering a graph structure composed of all documents in the collection. This method is not specific to the query submitted to the system. In the second method, he computed the Page Rank score for a query-specific citation graph, which is composed of the top-k documents initially retrieved for a given query patent and their cited documents. His experimental results on the NTCIR-6 test collection demonstrated that the query-specific Page Rank score is more effective than the traditional Page Rank score. As a baseline for this article, we implemented the work of Fujii [2007]. Similar to his work, we used the Page Rank measure on a query-specific citation graph to calculate a score for quantifying the authoritativeness of each document.

Lopez and Romary used references in a patent document as a starting point for prior art search [2010]. They showed that extracting patent references using regular expression patterns resulted in missing at least 40% of references. In order to increase the accuracy of the extraction module, they identified patent reference blocks in the text of the patent using a Linear Chain CRF (Conditional Random Field) model. The reference block is then parsed to obtain a set of bibliographical attributes. They also used online bibliographical services to enrich the identified references. In order to extract characterizing key terms from a document to formulate a synthetic query, they extracted candidate phrases up to 5-grams from the text of the patent documents. They

estimated the potential of each phrase to serve as a key term with a bagged decision tree. This model is trained on the key terms annotated by authors and readers from a set of training documents. Our work here is different from these mentioned works, on one hand, we do not use references to cited patents; on the other hand, we do not have access to annotated key terms that form the text of the patent query to supervise the query formulation process.

In prior art search task of CLEF-IP, citation information of query patents (topics in the testset) was removed and used for building the relevance judgement (ground truth). However, references to cited patents in the text of the query patent were not removed; as a result, the usage of these references in the text of the query patent was not recommended by organizers, unless participants explicitly mention this usage.

3. POTENTIAL RELEVANCE EVIDENCE SOURCES FOR QUERY REFORMULATION

In this section, we categorize different information sources that can be used as additional knowledge for query reformulation in patent retrieval.

Query Patent. A query patent is a structured document which is composed of the following fields: *title*, *abstract*, *description*, and *claims*. The claims field comprises of multiple claims and they are numbered. A claim which does not refer to any other claim is called an *independent claim*, while others are called *dependent claims* [Lupu and Hanbury 2013]. The independent items in the claim field of the patent comprise the kernel of the technical innovation of the patent. Among the claims, the most important one is the first independent claim (the first item in the claims), which represents the essence of the technology of the patent document. The other parts of the patent document illustrate the reason, background, implementation, and advantages, of the invention being described [Lupu et al. 2011]. An example of a patent application is shown in Figure 1. According to this example, claim 1 is an independent claim while claims 2–5 are dependent claims.

IPC Classification. The International Patent Classification (IPC classification) provides a hierarchical categorization over different technological fields, such as computer science, electronics, mechanics, and biochemistry. Such classes are language-independent symbols assigned as metadata to the patent documents. They categorize the content of a patent document and describe the field of technology that a patent document belongs to. These IPC classes can be seen as conceptual tags assigned to the documents [Lupu et al. 2011]. For each conceptual tag, there are textual descriptions available (IPC definition pages) that provide contextual cues about different technical fields.

Citation Chain. Patents are issued with a list of other documents that were cited during the processing of the patent application either by the patent examiner or the inventor. A patent searcher has access to documents cited by each patent but also has access to documents that cite each patent. The process of searching both of these sets of documents is referred to as backward and forward citation searching, respectively.

These sources have different vocabulary usage. The query patent itself has an obscure style of writing (patentese) [Lupu et al. 2011]. This characteristic might create a term mismatch problem in finding relevant documents for a given patent. However, the other two resources provide a more established vocabulary usage. The descriptions of IPC classes represent the standard vocabulary usage related to different domains. The citation chain contains the language used by the community of inventors related to the subject of the invention of the query. Thus, the vocabulary usage of the two latter sources are complementary to the query itself.

Application Number	EP-1832953-A2
Title	Method and apparatus for managing a peer-to-peer collaboration system
IPC Classes	G06F1/00, G06F15/00, G06F21/00, G06F21/24, H04L29/06, H04L29/08
Abstract	Users and devices in a peer-to-peer collaboration system can join a management domain in which members are administered as a group by a centralized management server operated by an enterprise. In response to a administrator request to join the management domain, the user downloads an injectible identity file containing a definition of the managed user/device into the user system. The user then joins the managed domain by associating the injected identity with their actual identity. Once a user or device is part of a management domain, that user or device receives license rights and policy restrictions that are associated with the domain. In return, the management server interacts with the individual peer-to-peer collaboration systems to enable the enterprise to monitor the enterprise to monitor the usage of, and control the behavior of, that specific identity within the peer-to-peer collaboration system.
Description	This invention relates to peer-to-peer collaboration systems and, in particular to methods and apparatus for gathering usage statistics for managing such systems. New collaboration models have been developed which operate in a "peer-to-peer" fashion without the intervention of a central authority. One of these latter models is built upon direct connections between users in a shared private "space". In accordance with this model, users can be invited into, enter and leave a shared space during an ongoing collaboration session between other users. Each user has an application program called an "activity", which is operable in his or her personal computer system, communication appliance or other network-capable device which generates a shared "space" in that user's computer. The activity responds to user interactions within the shared space by generating data change requests, called "deltas." The activity also has a data-change engine component that maintains a local data copy and performs the changes to the data requested by the deltas. The deltas are distributed from one user to another over a network, such as the Internet, by a dynamics manager component. When the deltas are received by another user activity in the shared space, the local data copy maintained by that activity is also updated...
Claims	**1.** A method for managing a peer-to-peer collaboration system in which users having identities are directly connected to each other in a shared private space by client software operating in devices and wherein the users communicate with a management server using the client software, the method comprising: (a) sending a request from the management server to the user to become a managed entity; (b) downloading from the management server to the client software a definition file containing a definition of the managed entity; and (c) associating information in the definition file with user identities and device in the client software in order to create a managed entity. **2.** The method of claim 1 wherein the managed entity is a managed user and the definition information file is an injectible identity file. **3.** The method of claim 1 wherein the managed entity is a managed device and the definition information file is a device information file. **4.** The method of claim 3 wherein the device information file is a Windows REG file. **5.** The method of claim 1 further comprising: (d) sending at least one license file from the management server to the managed user; and (e) in response to information in the license file, enabling at least one function in the client software....

Fig. 1. An example topic in CLEF-IP 2011 testset (an excerpt).

3.1. The Architecture of our Proposed Model

Figure 2 illustrates the general scheme of our proposed method for query expansion. The system receives a full patent application (query patent) consisting of textual fields and classification information. Note that we do not have the citation information associated with the patent application. While, for the rest of the documents in the collection, we have access to both classification information and citation information.

In the first step, we estimate a query model from the textual fields of the patent. In step II, we build an initial rank list based on the query model estimated from the query patent. In step III, we take the initial ranked list and extract query-dependent citation links from the top-k ranked documents. We then build a query-specific citation graph. We perform influence analysis on the citation graph incorporating the temporal features of the cited documents into our model. In step IV, we build a citation query model. In step V, we build a query-specific lexicon from IPC definition pages. In step VI, we make a lookup in the lexicon using the IPC classes of the query document. In step VII, we extract the terms related to the IPC classes of the query from the IPC lexicon. In step VIII, we expand the citation query model with expansion concepts extracted from the lexicon. In step IX, query expansion is performed, and the positional information between query terms and expansion terms is used to calculate weights for ensuring high-quality expansion. The final rank list is generated as the result of this step.

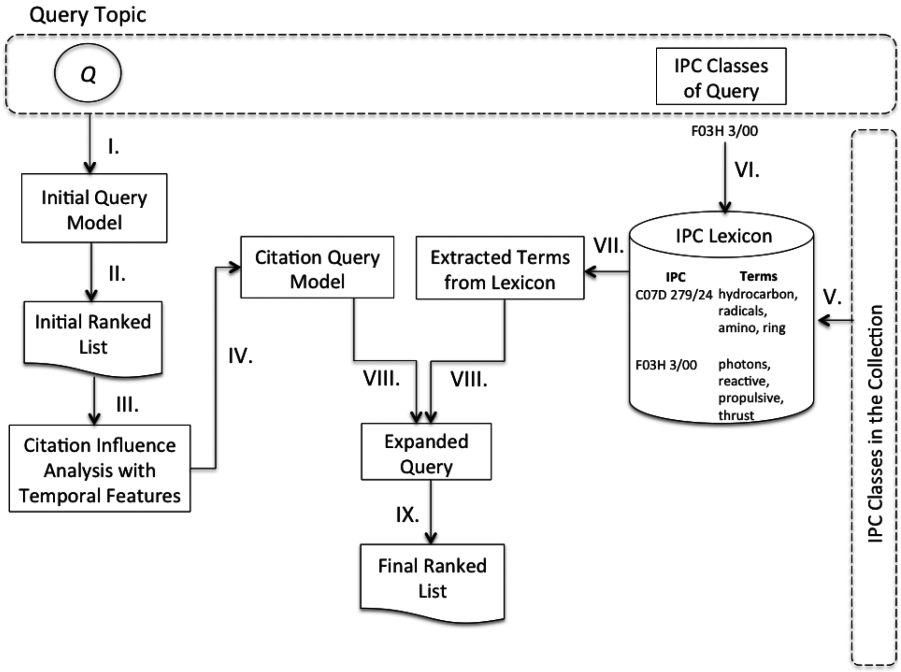


Fig. 2. The general scheme of our proposed method for query expansion using IPC lexicon and citation information. Numbers indicate the sequence flow of operations.

Table I. Entry in the Conceptual Lexicon

IPC Class	Representing Terms
C07D 279/24	hydrocarbon, radicals, amino, ring, nitrogen, atom

3.2. IPC Conceptual Lexicon

We now explain the process of building a lexicon from IPC definition pages. We refer to this lexicon as a *conceptual lexicon*. We consider the description of an IPC subgroup[8] as a text segment. We performed stop-word removal on these text segments. We then filtered the patent specific stop-words. The list for patent-specific stop-words is built as follows: we calculated document frequencies for each term in the collection. We selected terms with the top 10% highest document frequency and considered them as patent-specific stop-words. The threshold 10% was set experimentally. We then filter them out to increase the accuracy of our lexicon. Examples of these patent-specific stop-words are “method,” “device,” “apparatus,” “process.”

Each entry in our lexicon is composed of a key and a value. The key is an IPC class and the value is a set of terms representing the mentioned class. An example of an entry in the conceptual lexicon is presented in Table I.

The lexicon can be used to extract expansion concepts related to the information need of a given query patent. To this end, the IPC classes of the query patent are searched in the lexicon, and the matching terms are considered as expansion terms.

[8]IPC classification scheme is arranged in a hierarchical, tree-like structure. Subgroup is the lowest hierarchical level in the IPC hierarchy.

Query expansion using the lexicon will help us solve the two following problems. The first problem is related to the fact that the usage of words is sensitive to the topic domain; in different domains, the same word may be used to indicate different meanings. We aim at finding the correct sense of a word by associating relevant terms from the topic domain to the given query terms for each query patent.

The second problem is related to the term mismatch. The vocabulary of the query patent is tailored by the language usage of the author (who often uses a nonstandard terminology), while conceptual lexicon provides a standard terminology. We try to combine these two terminologies, as we think this might alleviate the term mismatch.

4. QUERY-SPECIFIC CITATION GRAPH

In this section, we present the basics of representing the patent collection as a directed unweighted graph. We then focus on the influence analysis aiming to find the important documents in the citation graph that could influence their domain terminology.

In the CLEF-IP collections, the citations of query topics (query patents) are removed by the organizers and used for building the query relevance judgements (qrels) which are later used for automatic evaluation of topics. However, we have access to the citations of all other documents apart from the query topics in the collection. A recent work [Lopez and Romary 2009] used a Web service offered by the European Patent Office[9] to retrieve the bibliographical attributes of documents in the collection. We also used this Web service to extract all the citations of the documents in the collection with the exception of the query documents.

With these, we can use the citation links and build a graph from the documents in the collection. The assumption is that if a patent is cited by a large number of documents, the cited patent is possibly a foundation of the citing patents and is considered important. Therefore, its language might be useful to bridge the gap between the query and its relevant documents.

As previous work suggests [Fujii 2007], computing Page Rank values as a measure of static quality of the patent documents in the collection (calculated independently of any query a system might receive) has a clear disadvantage compared to conditioning the computation of Page Rank 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 Page Rank values relative to particular queries.

To gather a subset of documents in the collection, we follow the two subsequent steps. We later define how these documents are used to build a topic-specific citation graph.

(1) Given a query patent, we perform the search and retrieve an initial rank list of documents. We take the top-k documents from this list and call it the *root set*.
(2) We construct the *base set* by expanding the root set with any document that either cites or is cited by a document in the root set.

The subset of selected documents can be considered as a directed unweighted graph $G = (V, E)$, where V is a set of $|V| = N$ patent documents and $E \in V \times V$ is a set of citation relationships between patent documents. Each citation link from document A to document B can be seen as an endorsement of document B. We now compute the topic-specific Page Rank values for all nodes in the citation graph.

4.1. Establishing a Baseline Query

We now explain our approach to estimate a unigram query model from the query patent document. This query will be used to retrieve an initial set of documents to form the

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

root set. We create a language model Θ_Q for the query patent:

$$P(t|\Theta_Q) = P_{ML}(t|D), \tag{1}$$

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

$$P_{ML}(t|D) = \frac{n(t, D)}{\sum_{t'} n(t', D)}. \tag{2}$$

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

$$P(t|Q_{Init}) = Z_t \; P(t|\Theta_Q) \; log\left(\frac{P(t|\Theta_Q)}{P(t|\Theta_C)}\right), \tag{3}$$

where $Z_t = \frac{1}{\sum_{t \in V} P(t|\Theta_Q) \, log \frac{P(t|\Theta_Q)}{P(t|\Theta_C)}}$ is a normalization factor. What we have in the denominator is the the Kullback-Leibler divergence between Θ_Q and Θ_C, as it is summed over all the terms in the vocabulary. Thus, the normalization factor can be written as $Z_t = \frac{1}{D_{KL}(\Theta_Q||\Theta_C)}$. 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. All fields of the query document are considered in this estimation.

4.2. Citation Analysis of the Graph Structure

The computation of Page Rank value for a document D is performed as follows:

$$PR(D) = \sum_{x \in d_{* \to D}} \frac{PR(x)}{d_{x \to *}}, \tag{4}$$

where $d_{* \to D}$ is a set of patent documents that cites D, and $d_{x \to *}$ is a set of patent documents cited by D. If D is cited by a large number of documents, a high score is given to D. However, if a document cites n documents, the value for each cited document is divided by n [Brin and Page 1998].

We calculated the Page Rank values for all the documents in the topic-specific citation graph. In the next section, we explain how this value is used to guide the priority assignment to documents while estimating a query model from citation graph.

4.3. Query Expansion Guided by the Page Rank Scores

Our approach for query expansion aims to improve the language model of the initial query model by using the term distribution of documents in the citation graph. The key assumption of this approach is that the term mismatch can be alleviated by using the term distribution of documents with higher Page Rank scores.

We identify and weigh the most distinguishing terms in the documents belonging to the citation graph, and we use the calculated Page Rank values as document prior in a language modeling framework. The term sampling is performed as follows:

$$P(t|Q_{cit}) = Z_t \sum_{D \in G_{cit}} P(t|D)P(D), \tag{5}$$

where G_{cit} denotes the citation graph and $P(D)$ indicates the Page Rank score of document D calculated according to Equation (4) after normalization. Z_t is a normalization factor.

4.4. Temporal Analysis of the Citation Graph

After conducting our citation analysis using Page Rank scores, we noticed that the Page Rank score is assigning a higher score to older documents. To investigate whether the language of the query is more susceptible to the terminology of older documents, we looked into the relationship between relevance and time and studied how relevance changes over time using time series.

For this analysis, we focused only on the result set of a query and not all the documents in the collection. We derived time series from the result set (which could be relevant to the query, thus referred to as pseudo relevant documents) and in parallel from the set of relevant documents (qrels). We then compared these two time series. We consider the publication date of the first kind-code of a patent application as the time tag. A patent document has different kind-codes (versions) which are used to denote its level of publication (e.g., first publication, second publication, or corrected publication). The unit of time granularity considered in our analysis is a year. We thus aggregated documents with publication dates in the same year in one bin.

After performing this analysis, we observed that for the majority of queries, the temporal distribution of true relevant documents (qrels) has a higher density of documents with recent publication dates, while our result set contains a higher number of documents with older publication dates. This means that the pseudo relevant set is lagging behind the qrels in the time dimension. Likewise, citation influence analysis using Page Rank scores is biased towards the terminology of older documents.

We are thus interested in taking into account the time dimension in order to improve the effectiveness of the retrieval. To do this, we need a query model that captures the established terminology (derived from older documents) but at the same time encodes the new vocabulary of the field (which is led by recent documents). The challenge is how to balance these two distinct terminologies and build a query model that combines both of these terminologies at once.

Modeling Decay over Time. Our aim is to capture the language change over time. We take into account the patent publication dates and prioritize recent documents while penalizing older documents.

In previous work, different functions have been used to model the decay over time in a retrieval setting [Amati et al. 2012; Peetz and de Rijke. 2013]. Exponential decay function has been used previously in IR tasks for modeling the time decay [Li and Croft 2003]. Recently, inspired by cognitive psychology, the Weibull function has been introduced as a time-aware prior and has been successfully employed on the blog and news collections for improving the query modeling of event-based queries. The Weibull function has been shown to be more effective compared to exponential decay according to the retrieval results obtained in Peetz and de Rijke. [2013]. In this work, we consider time-aware functions to discount the effect of older documents and capture the terminology of recent documents in the query model.

We describe two time-aware functions.

—Exponentail decay,

$$f_{\text{Exp-Decay}}(D, q, g) = \hat{\mu} e^{-\hat{\mu}\ \delta_g(q,D)} \tag{6}$$

—Weibull,

$$f_{\text{Weibull}}(D, q, g) = e^{-(\frac{\hat{\mu}\ \delta_g(D,q)}{\hat{d}})^{\hat{d}}}, \tag{7}$$

where $\delta_g(q, D)$ is the difference between the publication date of the query and the publication date of the document D. $\hat{\mu}$ determines the decay parameter, $\hat{d}$ indicates the steepness of the decay (forgetting) function, and g denotes the time granularity.

Table II. Comparing the Query Terms Selected from the Query Patentand the Citation Graph

Query Document	Citation Graph (Page Rank)	Citation Graph (Temporal)
manage, server, collaborate, client, soap, peer, ...	transact, handle, service, access, command, ...	network, permission, secure, request, collect, ...

We identify and weigh the most distinguishing terms in the documents in the citation graph, prioritizing recent documents. We consider a granularity of one year and employ time-aware functions as document priors.

$$P(t|Q_{innov-cit}) = Z_t \sum_{D \in G_{cit}} P(t|D)P(D), \tag{8}$$

where G_{cit} is the citation graph. The document prior component in Equation (8), $P(D)$, is proportional to the value calculated by the exponential decay function or Weibull function. The weight of each document, using the exponential decay function, is calculated as follows: $P(D) = \frac{f_{\text{Exp-Decay}}(D,q,g)}{\sum_{D \in G_{cit}} f_{\text{Exp-Decay}}(D,q,g)}$. Z_t is a normalization factor.

Our assumption is that recent documents have an innovative language, and using temporal priors allows us to capture the terminology of recent documents.

4.5. Query Expansion using Citation Graph and Temporal Features

We build a query model that has good coverage over different time intervals, utilizing the language usage of older documents (the established terminology of the domain) and the innovative language usage of recently published documents. This query is built from the linear combination of the initial query, the citation query model using the Page Rank scores, and the temporal query model of the citation graph. We interpolate the temporal query (as estimated in Equation (8)) with the citation query (as estimated in Equation (5)) and the initial query (as estimated in Equation (3)):

$$P(t|Q) = \alpha\ P(t|Q_{Init}) + \beta\ P(t|Q_{cit}) + (1 - \alpha - \beta)\ P(t|Q_{innov-cit}). \tag{9}$$

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

Table II shows a comparison between a list of terms derived from the patent application, "EP-1832953-A2," terms sampled from documents with high Page Rank scores belonging to the topic-specific citation graph, and terms derived from the temporal query model of the citation graph. This query topic belongs to the CLEF-IP 2011 topic set. The title of this patent topic is "Method and apparatus for managing a peer-to-peer collaboration system." By looking at this example, we see that we are able to select terms from documents in the citation graph which are relevant to the topic of the query but are not captured in the initial query model.

5. A PROXIMITY-BASED FRAMEWORK FOR QUERY EXPANSION

We now explain how the IPC lexicon is used for query expansion. To do this, we first describe strategies to identify expansion concepts that are referring to query concepts in Section 5.1. Then in Section 5.2, we explain how to estimate the probability that an expansion term is referring to a query term. Finally in Sections 5.3 and 5.4, we discuss calculating relevance scores for documents.

5.1. Query Reformulation

Let $Q = \{q_1, q_2, \ldots, q_k\}$ be a query composed of top-k query terms with highest weights according to a query model estimated from the query patent document D_Q (as explained in Equation (12)). Given the IPC classes assigned to D_Q, we select a set of concepts

Table III. Comparison between the List of Terms Derived from Three Information Sources for the Query with Title "Ink-Jet Recording Ink"

Query Document	Conceptual Lexicon	Retrieval Corpus
acrylate, ink, jet, acid, polymer, pigment, record, ...	light-sensitive, duplicate, printer, ink, sheet, mark, ...	record, liquid, surface, composition, polymer, cartridge, ...

$C_E = \{e_1, e_2, \ldots, e_m\}$ from the conceptual lexicon (as explained in Section 3.2). The set of C_E is associated to the query Q, since the IPC lexicon contains explanations about the IPC classes of D_Q. Once the set of concepts C_E is identified, we determine the importance weights according to their distance from the query terms based on the intuition that concepts closer to query terms are more related to the query. Equation (10) shows the process of calculating importance weights for expansion concepts. We can then rerank documents in the initial rank list $\mathbb{R}$ using a weighted combination of matches of concepts in C_E and our initial keyword query Q based on Equation (12).

We explain four different strategies for selecting expansion concepts in the following.

Explicit Expansion Concepts. In this strategy, we use the concepts in our conceptual lexicon which match against the IPC classes of D_Q. However, we restrict our attention to concepts that are present in D_Q. This provides a set of explicit expansion concepts (a subset of C_E) which serve as candidate expansion terms. We refer to this set as X_E. We utilize the proximity of query terms and expansion terms inside D_Q to assign importance weights to items in X_E. These weights are then used to rerank documents in the list $\mathbb{R}$.

Implicit Expansion Concepts. In this strategy, the expansion terms are not limited to the set of explicit expansion concepts X_E which were defined previously. Instead, our query expansion method includes all expansion concepts in C_E. In this setting, we extract proximity information from documents inside $\mathbb{R}$ to compute importance weights for expansion terms. This strategy is able to make use of all terms available in C_E and is not limited to the concepts that appear in D_Q.

Combining Search Strategies. In this strategy, instead of expanding the initial query, we calculate an IPC score based on the expansion concepts in C_E. We linearly combine this score with the initial scores calculated in $\mathbb{R}$. Our goal is to compare whether having a unified query, as it exists in the query expansion, is better than constructing two separate queries and combining their results at the end. We introduce this setting for the experiments in order to simulate the specific search strategies taken by searchers for retrieving relevant documents. In such a search strategy, searchers perform separate searches based on different information sources, such as the patent query document and IPC classes, and then merge the results of the runs together to produce a unique rank list [Lupu et al. 2011].

Proximity-Based Pseudo-Relevance Feedback. As a comparison baseline, we use the retrieval corpus as a source for PRF, and we use the feedback set for selecting expansion terms. The distance between query terms and expansion terms is used to calculate the weight for expansion terms.

As an example, Table III shows the terms selected from different information resources for the query patent "EP-1783182-A1" selected from CLEF-IP 2010 test topics. The terms from the retrieval corpus are selected via the procedure of PRF.

5.2. Estimating the Query Relatedness

In this section, we explain our method for estimating the probability that expansion term e at position i is related to query term q. We calculate this probability as follows:

$$P(q|i, D) = \sum_j P(q|j)P(j|i, D), \tag{10}$$

where D denotes a document, i denotes an expansion term position, and $j = \{1, 2, \ldots, k\}$ denotes a set of query term positions. $P(q|i, D)$ indicates the probability that the expansion term at position i in D is about the query term q. We refer to this probability as *query relatedness probability*. To find the query relatedness at position i, we calculate the propagated probability from all query positions at position i. For every position j in D, we consider the weight of query term at position j, denoted by $P(q|j)$, and weight it by the probability that the term at position j is about the expansion term at position i, denoted by $P(j|i, D)$. This probability is estimated as follows:

$$P(j|i, D) = \frac{k(j, i)}{\sum_{j'=1}^{|D|} k(j', i)}, \tag{11}$$

where $k(i, j)$ is the kernel function determining the weight of the propagated query-relatedness from j to i. We model the query relatedness by placing a density kernel function around query terms. The values are normalized to obtain the calculated weights in form of probabilities.

In the following, we present different kernels used in our experiments. We study three different density functions, namely, Gaussian, Laplace, and Rectangle kernels. We selected Gaussian and Laplace kernels as they have been shown to be the best performing kernels among the kernel functions tested in previous work [Lv and Zhai 2009; Gerani et al. 2012]. We also chose the Rectangle kernel to simulate the effect of imposing a hard boundary over passages in contrast to the soft boundary introduced by other kernels. The parameter σ controls the spread of kernel curves and restricts the propagation scope of each term.

—*Gaussian kernel*,

$$k(i, j) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(i-j)^2)}{2\sigma^2}\right].$$

—*Laplace kernel*,

$$k(i, j) = \frac{1}{2b} \exp\left[\frac{-|i-j|}{b}\right],$$

$$\text{where } \sigma^2 = 2b^2.$$

—*Rectangle kernel*,

$$k(i, j) = \begin{cases} \frac{1}{2a} & \text{if } |i-j| \leqslant a \\ 0 & \text{otherwise,} \end{cases}$$

$$\text{where } \sigma^2 = \frac{a^2}{3}.$$

Our aim is to investigate whether it is better to use kernel functions which favor expansion term occurrence in close proximity of query terms or not.

5.3. Calculating Document Relevance Score

In this section, we intend to calculate the overall probability that relevant expansion concepts (inside the document) are related to the technical concept of the query. This probability is denoted by $P(q|D, e)$, which is defined as

$$P(q|D, e) = \sum_{i=1}^{|D|} P(q, i|D, e) = \sum_{i=1}^{|D|} P(q|i, D, e)P(i|D, e). \quad (12)$$

We assume e and q are conditionally independent given their positions in document D. Thus, $P(q|i, D, e)$ reduces to $P(q|i, D)$, which can be estimated using the query relatedness probability. We now need to estimate the probability $P(i|D, e)$. We suggest two different methods for estimating $P(i|D, e)$.

—*Avg Position Strategy*. All positions of expansion concepts are equally important:

$$P(i|D, e) = \begin{cases} 1/|pos(e)| & \text{if } t_i \in e \\ 0 & \text{otherwise,} \end{cases}$$

by substituting this in Equation (12), we have

$$P(q|D, e) = 1/|pos(e)| \sum_{i \in pos(e)} P(q|i, D), \quad (13)$$

where $|pos(e)|$ denotes the number of occurrences of expansion term e in document D.

—*Max Position Strategy*. As an alternative, we only consider the expansion term position with the highest $P(q|i, D)$ as important:

$$P(q|D, e) = \max_{i \in pos(e)} P(q|i, D). \quad (14)$$

5.4. Normalization

Here we compare the effect of different normalization methods prior to linear combination using two score normalization methods: MinMax [Lee 1997] and HIS normalization [Arampatzis and Kamps 2009]. These methods are often used in distributed information retrieval. MinMax normalization method shifts and scales scores to be between zero and one. On the other hand, HIS normalization estimates a single cumulative density function (CDF) for every search engine based on historical queries.

We also experimented with a variation of score normalization where we first applied MinMax and then HIS normalization. We refer to this method as MinMax-HIS throughout the experiments.

6. EXPERIMENTAL SETUP

Here we describe the details of the experimental setup, reporting about the testsets and baselines.

6.1. Testsets

In this section, we explain the experimental setup for evaluating the effectiveness of our proposed approaches.

Testing Collections. We conducted our experiments over two years worth of CLEF Intellectual Property (CLEF-IP) task, including CLEF-IP 2010 and CLEF-IP 2011 datasets. CLEF-IP 2010 contains 2.6 million patent documents while CLEF-IP 2011 consists of about 3 million patent documents. In our experiments, we used the English subsection of both collections. The English testset of CLEF-IP 2010 corresponds to 1,348 topics. The English testset of CLEF-IP 2011 consists of 1,351 topics. We used the

Table IV. IPC Section Distribution over English Testset of CLEF-IP 2010 and CLEF-IP 2011

Category	Description	# of topics in CLEF-IP 2010	# of topics in CLEF-IP 2011
A	Human Necessities	154	250
B	Performing Operations and Transporting	307	213
C	Chemistry and Metallurgy	255	150
D	Textiles and Papers	10	23
E	Fixed Constructions	7	18
F	Mechanical Engineering, Heating, Weapons, and Blasting	90	143
G	Physics	289	263
H	Electricity	236	291
	total number of queries	1,348	1,351

training topics of CLEF-IP 2010 for the parameter tuning of our model. This training set consists of 300 topics.

We calculated statistics about the IPC classifications codes. In general, there are about 70,000 classes in the most fine-grained level of the IPC hierarchy.[10] The number of distinct classes in CLEF-IP 2010 and CLEF-IP 2011 are 62,183 and 63,495, respectively. On average, there are 3.4 IPC classes assigned to each documents in CLEF-IP 2010 and 3.9 in CLEF-IP 2011.

As previously mentioned, the relevance judgements for the CLEF-IP challenge are built based on the documents listed in the search report of a patent application which is written by a patent examiner. This report might share references with the initial citation list provided by the patent applicant. To remove the bias that might be introduced by the applicant, the initial citation of the query patent is removed by organizers of CLEF-IP. As a consequence, our model does not use the initial citation information of a query patent.

Pre-processing. We used the Terrier Information Retrieval System[11] to index the collection with the default stemming and stop-word removal. We considered the textual fields of entire patent documents while indexing. We then removed patent-specific stopwords such as "device" and "method." The list for patent-specific stop-words is built as follows. We calculated document frequencies for each term in the collection. We then selected terms with top 10% highest document frequency and considered them as patent-specific stop-words. The value 10% is experimentally set as a threshold.

Evaluation. We used the relevance judgement for the test topics with English language provided by CLEF-IP for evaluation purposes. We report recall, mean average precision (MAP), and patent retrieval evaluation score (PRES) [Magdy and Jones. 2010b], which combines MAP and recall in one single score.

In the remainder of our experiments, we used the randomization (permutation) test with a confidence level of 0.05 to report statistical significance test results, since this test has been shown to be more reliable than Wilcoxon and t-test [Smucker et al. 2007].

We also performed the evaluation per topics belonging to each technology class. Table IV represents the information about the field of technology of test topics. As shown in Table IV, IPC divides technology into eight sections.

6.2. Establishing a Baseline

We estimated an initial query model from the query patent document by calculating the importance of each term according to a weighted log-likelihood-based approach,

[10] http://www.wipo.int/classifications/ipc/en/general/statistics.html.
[11] http://ir.dcs.gla.ac.uk/terrier/.

Table V. Choosing Baselines on Two Retrieval Collections

CLEF-IP 2010 (training topics)			
Run identifier	MAP	recall	PRES
W10TR	0.1219	0.6367	0.5512
CLEF-IP 2010 (test topics)			
Run identifier	MAP	recall	PRES
W10TE	0.1295	0.6105	0.5150
CLEF-IP 2011 (test topics)			
Run identifier	MAP	recall	PRES
W11TE	0.0990	0.5935	0.4859

as explained in Section 4.1. The entire text of query patent documents is used in this estimation. Table V summarizes the results we obtained using the initial query model for the topics in the training and testset of CLEF-IP 2010 and the testset of CLEF-IP 2011. Throughout our experimental section, W10TE is used as the baseline on the test topics of CLEF-IP 2010, and W11TE is used as the baseline over test topics of CLEF-IP 2011. Note that the training set of CLEF-IP 2010 is only used for tuning the parameters of the model, thus we will refer to W10TR in such comparisons.

We used the language modeling approach with Dirichlet smoothing [Zhai and Lafferty 2001] to score documents from both collections and to build the initial rank lists. We empirically set the value for the smoothing parameter μ to 1500. We also used language modeling for the reranking of the results. Table V reports evaluation results after performing IPC filtering on the rank list. This means that documents in the rank list that do not share any IPC class with the query document are filtered out. IPC filtering improves the evaluation results.

At the end of the experimental section, we will compare the performance of our approach with the best official results of CLEF-IP 2010 and CLEF-IP 2011.

7. EXPERIMENTAL RESULTS

In this section, we explain the experiments conducted in order to evaluate the performance of the proposed models. We present the results and formulate answers to the following questions according to the results of our experiments.

(1) Do citation links together with the content of the cited documents improve the performance of the initial query built from the query document? Does employing the temporal features of the query and of the collection result in a more precise query? What type of document prior is more effective in modeling the decay over time?
(2) Is the IPC conceptual lexicon useful for query expansion? Is the proximity information between query terms and expansion terms, extracted from the IPC conceptual lexicon, helpful in identifying weights for expansion terms?
(3) What are the effects of the parameters of the model on the final performance?

In Section 7.1, we report the results of our experiments on the influence analysis over the citation graph. In Section 7.2, we show how the temporal query modeling from the citation documents improves over the citation query model. Sections 7.1 and 7.2 discuss and provide answers to the first question. In Section 7.3, we aim to answer the second question. To this end, we report the results of using IPC lexicon for query expansion, and we study the effects of using different density kernels to model the proximity information. In Section 7.4, we study the sensitivity of the proposed framework for query expansion using proximity information in order to answer the last question.

Table VI. Performance of Different Citation Analysis Methods with a CutOff Value of 1,000

CLEF-IP 2010 test set				
Method	Run description	MAP	recall	PRES
Score-cit1	citation depth level 1	0.102	0.567	0.449
Score-cit2	citation depth level 2	0.105	0.574	0.461
QM-cit1	citation depth level 1	0.118	0.580	0.469
QM-cit2	citation depth level 2	0.121	0.585	0.474

Table VII. Performance of Different Citation Analysis Methods with a Cut-Off Value of 1,000

CLEF-IP 2011 test set				
Method	Run description	MAP	recall	PRES
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

7.1. Query Expansion using Citation Graph

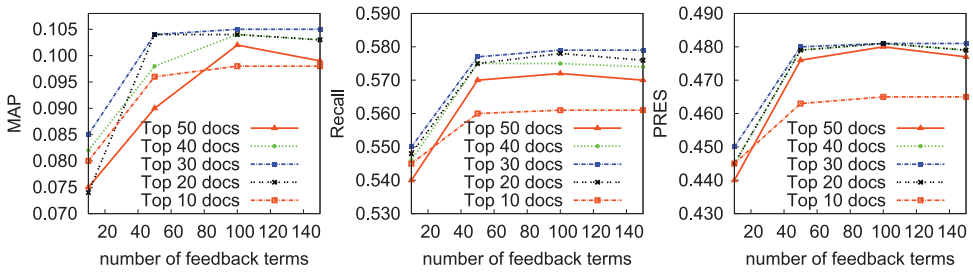
We now describe the structure of the experimental evaluations. We compare the two following methods with the baseline presented in Table V. The first method corresponds to our implementation of the work reported in Fujii [2007]. This method is focused on computing a composite score using the textual 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 second method is our proposed model which estimates a query model from the documents in the citation graph and expands the initial query using the estimated model from the term distribution of the documents in the citation graph. This method is referred to as *QM-cit*. Tables VI and VII show the evaluation results of different methods using the CLEF-IP corpora.

Results marked with † represents a statistical significant difference compared to Score-cit1 and Score-cit2. The reported results for QM-cit1 and QM-cit2 are obtained using the top-100 feedback terms selected from the expanded query model. The top-30 feedback documents are selected and used to generate the root set. The number of feedback terms and number of feedback documents are experimentally set with the goal of optimizing the performance of the method.

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 Section 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.

The results of Tables VI and VII suggest that the QM-cit method obtained better performance compared to Score-cit in terms of both recall and precision. QM-cit2 obtained statistical significant improvement in terms of recall over Score-cit1 and Score-cit2. This observation suggests that using the link-based structure as well as exploiting the term distribution of the citations (through estimation of a query model) is more useful than using the citation links alone. We can see from the results of Tables VI and VII that neither of the versions of Score-cit nor QM-cit achieved statistically significant improvement over the baselines W10TE and W11TE.

The results presented in Tables VI and VII show that increasing the depth of the citation graph (from depth 1 to depth 2) has a positive effect on the performance of



(a) MAP@1000. (b) Recall@1000. (c) PRES@1000.

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

both Score-cit and QM-cit methods. We also carried out experiments with a citation graph of depth 3, where three consecutive iterations of the steps described in Section 4 are considered. The obtained performance is statistically indistinguishable from the results for Score-cit2 and QM-cit2. We therefore did not present these results.

In Figure 3, we studied the effect of increasing feedback terms and feedback documents on the performance of QM-cit2. We notice that increasing the number of feedback terms has a consistent positive effect on all evaluation metrics. However, when we vary the number of feedback documents, we can see from the curve trends of QM-cit2 that there is a marked drop of performance in terms of MAP for values more than 30. We observe a less severe drop of performance in terms of recall. We can conclude that recall is less susceptible to the number of feedback documents than MAP. By looking at PRES values presented in Figure 3, we can see that the best performance of QM-cit2 is obtained with 30 feedback documents and 100 feedback terms.

In the next section, we show how we can obtain a better performance by incorporating temporal information into the model when performing citation influence analysis.

7.2. Enhancing Citation Analysis with Temporal Information

We study the impact of the temporal features for improving the citation query model, and we look into capturing the language change over time. Table VIII shows the results of the temporal query model (according to Equation (9)). Results marked with † achieved statistically significant improvement over the baselines W10TE and W11TE. The parameters of the model are tuned using five fold cross-validation to maximize PRES.

The results reported in Table VIII show that including temporal features into the citation query model led to statistically significant improvement in terms of MAP. Furthermore, reported results show that modeling the decay over time using the Weibull prior (Equation (7)) performed better than the exponential decay prior (Equation (6)).

To further investigate the effect of the temporal query modeling, we presented the evaluation results of methods in different technological fields in Tables IX and X.

The results of Table IX shows that TM-WB obtained a better performance compared to QM-cit2 in categories F, G, and H. These improvements hold for all three reported metrics. However, we observe that the performance of TM-WB is lower than QM-cit2 in category C. Our experiments on CLEF-IP 2010 showed that communities of inventors related to the topics in categories F, G, and H are more receptive to the linguistic changes over time, while categories A, B, C, D, and E are more resistant to the changes of the language. The decrease of precision on category C (which categorizes the patent documents related to “Chemistry” and “Metallurgy”) was counterintuitive, as we expected the language of this community to be evolving over time. However, we did not capture this effect in our query model. The results of Table X show the positive

Table VIII. Performance of Temporal Modeling with a Cut-Off Value of 1,000

CLEF-IP 2010 test set			
Method	MAP	recall	PRES
TM-ED	0.138	0.587	0.496
TM-WB	0.145 †	0.588	0.503
CLEF-IP 2011 test set			
Method	MAP	recall	PRES
TM-ED	0.124 †	0.580	0.487
TM-WB	0.128 †	0.582	0.490

Table IX. Evaluation Results over Testset of CLEF-IP 2010

method name	metric	A	B	C	D	E	F	G	H
QM-cit2	MAP	0.138	0.123	0.127	0.070	0.086	0.128	0.110	0.119
	recall	0.518	0.598	0.535	0.620	0.500	0.619	0.600	0.598
	PRES	0.433	0.489	0.448	0.540	0.427	0.517	0.476	0.482
TM-WB	MAP	0.134	0.122	0.121	0.075	0.081	0.148	0.117	0.134
	recall	0.518	0.596	0.533	0.620	0.500	0.638	0.624	0.599
	PRES	0.433	0.489	0.439	0.542	0.427	0.529	0.500	0.495

Table X. Evaluation Results over Testset of CLEF-IP 2011

method name	metric	A	B	C	D	E	F	G	H
QM-cit2	MAP	0.129	0.099	0.132	0.195	0.125	0.096	0.103	0.073
	recall	0.603	0.566	0.568	0.676	0.438	0.569	0.584	0.545
	PRES	0.509	0.471	0.485	0.588	0.370	0.490	0.483	0.440
TM-WB	MAP	0.127	0.102	0.130	0.195	0.137	0.120	0.105	0.101
	recall	0.603	0.568	0.565	0.679	0.446	0.579	0.585	0.558
	PRES	0.509	0.472	0.483	0.588	0.376	0.488	0.485	0.455

effect of our method in capturing the language change in categories E, F, and H, as opposed to other categories over the topics of CLEF-IP 2011.

A concluding remark for these experiments is that we obtained more accurate results when incorporating temporal features into our model. In the next section, we explain how the precise citation query model is expanded in our proximity-based framework.

7.3. Query Expansion using IPC Lexicon and Proximity Information

To guarantee the assignment of reliable importance weights to the expansion concepts, we need to start with a set of precise query terms, because we rely on the distance between query terms and expansion concepts in order to calculate importance weights for expansion concepts. Obviously starting with less noisy query terms has a direct effect on the quality of importance weights. According to our experiments, the run with the interpolated citation query model based on Equation (9) achieved the best MAP so far. Thus, in the remainder of the experiments, we focus on this query model. Note that Equation (7) is used as document prior for the temporal component.

We are interested in investigating the effectiveness of different query reformulation methods proposed in Section 5.1 for scoring documents in our proximity-based framework. The results of this comparison are summarized in Tables XI and XII. In all the comparisons, our query expansion method that uses *explicit expansion concept* is denoted as EEC, while the one that uses *implicit expansion concept* is referred to as IEC.

Table XI. Recall Results of Different Settings of the Kernel Functions using IEC Query Reformulation Method on the Training Topics of CLEF-IP 2010

IEC				
kernel\σ	25	75	125	150
Gaussian	0.6448	0.6564 †	0.6678 †	0.6805 †
Laplace	0.6429	0.6568 †	0.6583 †	0.6725 †
Rectangle	0.6401	0.6527	0.6563 †	0.6680 †

Table XII. Recall Results of Different Settings of the Kernel Functions using EEC Query Reformulation Method on the Training Topics of CLEF-IP 2010

EEC				
kernel\σ	25	75	125	150
Gaussian	0.6389	0.6420	0.6675 †	0.6640 †
Laplace	0.6365	0.6389	0.6688 †	0.6519
Rectangle	0.6340	0.6378	0.6646 †	0.6504

The performance of these methods is directly affected by the effectiveness of the kernel function used to estimate the query relatedness probabilities. Thus, we first compare three different proximity-based kernel functions.

As previously explained in Section 5, we place a density kernel function around each occurrence of query term positions in the document. The query relatedness at each expansion term position is then calculated by counting the accumulated query relatedness density from different query terms at that position. Therefore, an expansion term occurring at a position close to many query terms receives high query relatedness and thus obtains a greater weight in comparison to an expansion term which is located further away from query terms.

Our proximity-based framework has two parameters: the type of kernel function and its bandwidth parameter σ which controls the degree of query relatedness propagation throughout the entire document. To tune the parameters of our model, we used the training topics of CLEF-IP 2010.

Tuning the Parameters of the Kernel Functions. The results of comparing different kernel functions on the training topics of CLEF-IP 2010 are shown in Tables XI and XII. A † denotes statistical significant improvement over W10TR (presented in Table V). The results show that EEC and IEC achieved better performance over W10TR regardless of the choice of the kernel function.

It is also clear that among all the kernel functions, the Gaussian outperforms other types of kernels in most cases. Since the Gaussian kernel performed the best in most of the experiments, we use this kernel function in the rest of our experiments.

In order to find the best value for the parameter σ, we tried a set of fixed values in the range [25, 225] with a step of 25, similar to what has been done in previous work [Lv and Zhai 2009, 2010]. Tables XI and XII report the performance of different kernel functions using varying values of σ. We obtained the best result for IEC method using the σ value set to 150, and increasing the σ value to values more than 150 did not lead to an improvement in terms of retrieval effectiveness. For the EEC method, the best result is achieved using σ value set to 125. Overall, Tables XI and XII clearly demonstrate that the results obtained with the σ value of 150 achieved better performance in most cases, although the difference among different settings was not significant. Thus, we use the σ value of 150 in the rest of the experiments.

Comparison of Max and Avg Strategy. We now compare the max and avg strategies for calculating the probability of relevance of a document, as defined in Section 5.3.

Table XIII. Recall of the Max and Avg Method using Gaussian Kernel with IEC Reformulation Method on Training Topics of CLEF-IP 2010

method\σ	25	75	125	150
max	0.6448 †	0.6564 †	0.6678 †	0.6805 †
avg	0.6172	0.6205	0.6212	0.6249

Table XIV. Performance Results of Query Reformulation Approaches on Two Patent Retrieval Datasets on the Test Topics of CLEF-IP 2010 and CLEF-IP 2011

Collection	metric	baseline	IEC	EEC	CSS	PPRF
CLEF-IP 2010	MAP	0.1295	0.1434 †	0.1405 †	0.1301	0.1122
	recall	0.6105	0.6598 †	0.6452 †	0.6243	0.5890
	PRES	0.5150	0.5560 †	0.5510 †	0.5338	0.5029
CLEF-IP 2011	MAP	0.0990	0.1231 ‡	0.1225 ‡	0.1189	0.1022
	recall	0.5935	0.6369 ‡	0.6268 ‡	0.6094	0.5645
	PRES	0.4859	0.5290 ‡	0.5255 ‡	0.5141	0.4952

Table XIII shows the results of using avg and max strategies for different σ values on the training topics of CLEF-IP 2010 using the IEC reformulation method.

The results show that the max strategy is statistically better than the avg strategy. Thus, we use the max strategy in all configurations of our experiments throughout this article. A † denotes the statistical significant improvement over the avg method.

7.3.1. Effect of Query Reformulation. In this section, we present the evaluation results of our proposed approaches on the testset of CLEF-IP 2010 and CLEF-IP 2011. Table XIV reports the retrieval performance of query reformulation methods described in Section 5.1. The symbols † and ‡ denote statistical significant improvements over W10TE and W11TE (presented in Table V), respectively.

We now compare the performance of our query expansion methods which use IPC lexicon for extracting expansion candidates. In addition to EEC and IEC which were introduced earlier, the results of the other two query reformulation methods are presented in Table XIV. The method that *combines search strategies* is denoted as CSS. The last method in our comparison is the *positional-based pseudo relevance feedback*, which is denoted by PPRF.

The main observation from Table XIV is that IEC is always more effective than the other three methods. In addition, IEC improved significantly over the baseline in terms of recall on both collections.

Table XIV shows that a method which uses a conceptual lexicon for selecting expansion terms outperforms a method that uses feedback documents for identifying expansion terms. This is evident by comparing the performance of EEC, IEC, and CSS to the performance of PPRF, since the first three methods use the conceptual lexicon for query expansion. This result is consistent on both corpora used for evaluation.

In addition, the results of Table XIV demonstrate that IEC obtained improvement over EEC. In contrast to IEC, EEC extracts a limited set of expansion terms from the conceptual lexicon, the ones which are present in the query document. This diminishes the power of EEC in contrast to IEC. The results confirm that the unlimited usage of the conceptual lexicon is superior to its limited usage.

Another observation which can be made from Table XIV is that CSS achieved worse results compared to both EEC and IEC. This is perhaps due to the fact that information is lost during the merging of two separate runs made from the query terms and expansion terms. On the other hand, both EEC and IEC use a unified query which is composed of query terms and expansion terms. Overall, the results of Table XIV show

Table XV. Comparison with the Best Official Results on the English Subset of the Testset

Official best results of CLEF-IP 2010				
Method	Run description	MAP	recall	PRES
IEC	our method	0.1434	0.6598	0.5560
humb	rank 1	0.2264	0.6946	0.6149
dcu	rank 2	0.1807	0.616	0.5167
Official best results of CLEF-IP 2011				
Method	Run description	MAP	recall	PRES
IEC	our method	0.1231	0.6369	0.5290
nijm	rank 1	0.0582	0.6303	NA
hyder	rank 2	0.0593	0.5713	NA

Table XVI. Comparison of Performance Results of PRF and PPRF

Collection	metric	PPRF	PRF
CLEF-IP 2010	MAP	0.1122	0.0880
	recall	0.5890†	0.5630
	PRES	0.5029	0.4962
CLEF-IP 2011	MAP	0.1022	0.0842
	recall	0.5645†	0.5348
	PRES	0.4952	0.4794

that using the conceptual lexicon as a domain-dependent external resource is effective in terms of recall and precision. These findings deliver the answer to the third question listed in the beginning of Section 7.

We used 40 expansion terms (experimentally set) in each of the query reformulation methods. In Section 7.4, we studied the effect of varying the number of expansion terms and number of feedback documents on the performance of each method. We also presented the results of normalization using MinMax-HIS throughout the article. In Section 7.4, we investigated the effect of different normalization methods.

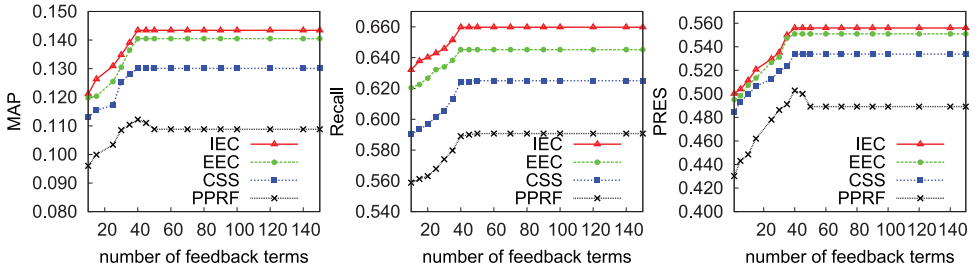
Table XV shows the performance of the IEC method along with the best official results of CLEF-IP 2010[12] and CLEF-IP 2011 [Piroi 2010; Piroi et al. 2011]. Note that PRES values were not reported for the best results of CLEF-IP 2011. It can be seen that the IEC method performed better than the best official results over CLEF-IP 2011. IEC can also be considered as the second best method on CLEF-IP 2010.

7.3.2. Comparison with Standard PRF. Table XVI reports the retrieval performance of PPRF and PRF. A † denotes statistical significant improvement over standard PRF.

As previously explained in Section 5.1, PPRF is similar to PRF since they both use a feedback set for selecting expansion terms. However, PPRF uses proximity information inside the feedback set to calculate the weight for expansion terms in contrast to standard PRF. The results show that PPRF performs significantly better than standard PRF. This result confirms the usefulness of proximity information for identifying importance weights for expansion terms, as previously shown in Lv and Zhai [2010]. PPRF and PRF did not achieve improvement over the baseline.

The number of feedback documents is set to 10 for both PPRF and PRF methods. We study the effect of this parameter on the performance of our methods in Section 7.4 and show that 10 is the optimal value.

[12]http://www.ifs.tuwien.ac.at/∼clef-ip/pubs/CLEF-IP-2010-IRF-TR-2010-00003.pdf.



(a) MAP@1000. (b) Recall@1000. (c) PRES@1000.

Fig. 4. Sensitivity to the number of expansion terms and number of feedback documents on CLEF-IP 2010.

7.4. Parameter Study

In this section, we study the influence of different parameters on the effectiveness of our proposed methods. The reported sensitivity results provide answers to the final question stated in the beginning of Section 7.

Number of Expansion Terms and Number of Feedback Documents. We plot the sensitivity of different query reformulation methods in relation to the number of expansion terms over CLEF-IP 2010 testset in Figure 4.

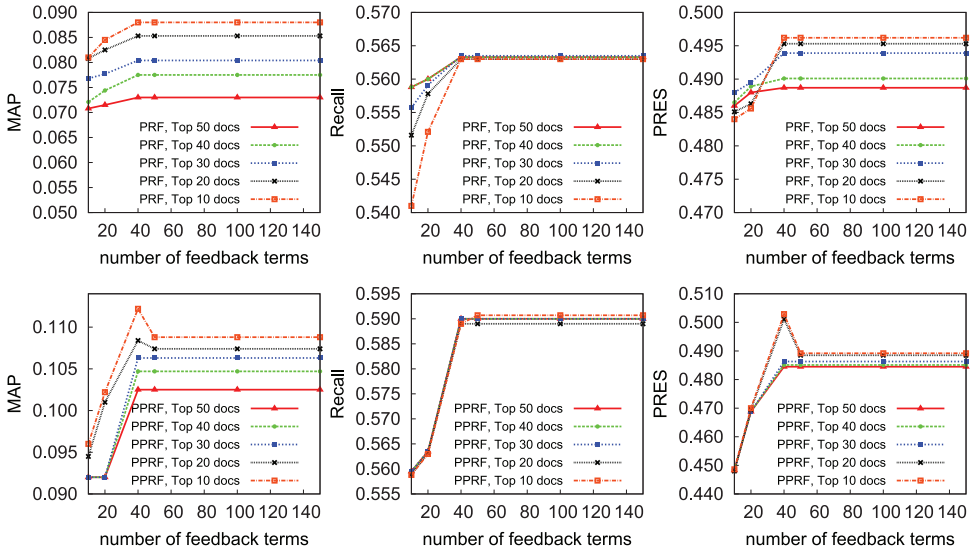
According to Figure 4, IEC is the clear winner among the four methods given the three evaluation metrics, and PPRF achieved inferior results compared to the other methods. We observe some variations in the performance of PPRF with different numbers of expansion terms. The best performance of PPRF is achieved with 40 expansion terms. Another observation is that IEC, EEC, and CSS seem to be less susceptible to the numbers of expansion terms. We can see that IEC, EEC, and CSS need 40 expansion terms to exhibit their best performance according to PRES values. IEC, EEC, and CSS continue to maintain a stable performance using higher numbers of expansion terms.

Since PRF and PPRF share the number of feedback documents, we are interested to understand how this parameter affects the retrieval performance of these two methods. We draw the sensitivity curves of PRF and PPRF with respect to the number of feedback documents and expansion terms on CLEF-IP 2010 in Figure 5. Since IEC, EEC, and CSS do not share the number of feedback documents as a parameter, we did not include them in this analysis.

Figure 5 shows that PPRF achieved better results compared to PRF. The best performance values for both PRF and PPRF are obtained with 10 feedback documents according to PRES values. The sensitivity curves for both PRF and PPRF show that using more than 10 feedback documents does not improve the performance. We hypothesize that this is because when we select feedback documents with higher rank positions in the rank list, more noisy terms are also selected, and this hampers the performance of PRF and PPRF.

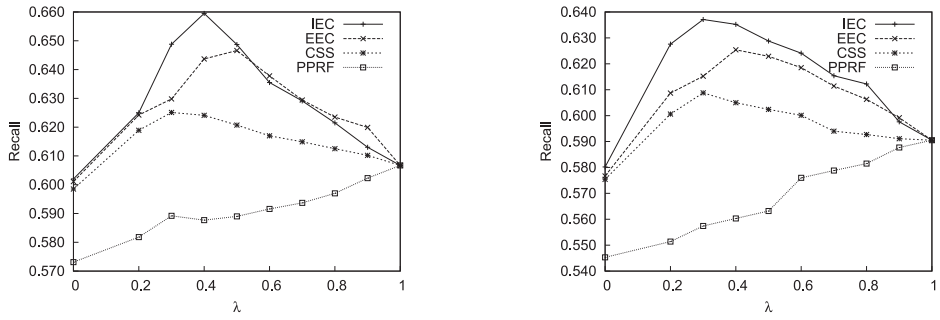
We can observe that retrieval effectiveness of methods presented in Figures 4 and 5 seem to stabilize after about 50 expansion terms. We hypothesize that this is because weights calculated for the expansion terms are small, and thus after a while, they do not play a powerful role in improving the retrieval effectiveness. We also think that this is because in an expansion setting, after a while, the query expansion reaches a saturation point, meaning that increasing the number of expansion terms does not lead to an improvement in retrieval effectiveness. We need to design further experiments to validate this hypothesis, and we leave this to the future work.

Effect of Combination. In all configurations of our experiments, we linearly combined the results from each of the reformulation methods with the initial query. The weight



(a) MAP@1000. (b) Recall@1000. (c) PRES@1000.

Fig. 5. Sensitivity analysis of PRF and PPRF on CLEF-IP 2010.



(a) CLEF-IP 2010. (b) CLEF-IP 2011.

Fig. 6. Sensitivity to the λ coefficient in the linear combination of results

of the interpolation λ controls the weight of the initial query. When $\lambda = 0$, the query expansion model is used, while when $\lambda = 1$, the initial query is used; λ was tuned based on the training topics of CLEF-IP 2010.

Figure 6 shows the results of the sensitivity analysis over the coefficient λ on the test topics of CLEF-IP 2010 and CLEF-IP 2011. We notice that IEC is more effective than other query reformulation methods for different λ values. The optimal value for the parameter λ seems to be in a small range around 0.4.

Effect of Normalization. Table XVII shows the comparison among different normalization methods. These results correspond to the final performance of each run after the combination on the testset of CLEF-IP 2010. The results are obtained with the IEC method. We observe that IEC achieved the best performance using MinMax-HIS normalization. The results of other methods confirm that applying normalization using MinMax-HIS is better than either MinMax or HIS alone. The improvements are not statistically significant.

Table XVII. Comparison of Different Normalization Methods over CLEF-IP 2010 using IEC method

metric	MinMax	HIS	MinMax-HIS
MAP	0.1312	0.1358	0.1434
recall	0.6534	0.6553	0.6598
PRES	0.5490	0.5525	0.5560

8. CONCLUSION AND FUTURE WORK

In this article, we presented a unified framework for query expansion which incorporates bibliographic information, IPC classifications, and temporal features to improve the initial query built from the query patent. We used the link-based structure of the citation graph together with the term distribution of cited documents and built a query model from the citation graph. We used the publication dates associated with the patents to adapt our query model to the change of vocabulary over time. The results showed the advantage of using the term distribution of the cited documents together with the publication dates. In particular, our citation influence analysis using temporal features improved the precision. It is worth mentioning that the positive effect of capturing the language change using the temporal query was more visible for patents belonging to domains such as "Mechanical Engineering" and "Electricity," while we observed a decrease in precision for topics belonging to the "Chemistry" category. These findings answered our first research question (RQ1) regarding employing the citation information for improving query modeling.

We then constructed an IPC lexicon which can be used as an external resource for query expansion. The IPC lexicon is built using the IPC descriptions available for each IPC class. Each entry in the conceptual lexicon is composed of a key and a value. The key is an IPC class and the value is a set of terms representing the mentioned class. We introduced a query expansion method leveraging the IPC lexicon by extracting expansion terms related to IPC classes of a given query document. We observed that the distance of expansion terms from query terms is a good indicator of the importance of expansion terms. We also noticed that the query expansion method using IPC lexicon has a recall enhancing effect. These observations provided answer to our second research question (RQ2).

We evaluated our proposed method using two patent datasets, namely, CLEF-IP 2010 and CLEF-IP 2011. The IEC query formulation method achieved similar performance as the state-of-the-art methods on CLEF-IP 2010 and was able to improve over the official best results of CLEF-IP 2011. We answered our third research question (RQ3) based on the evaluations performed over CLEF-IP datasets.

As a future direction, it would be interesting to capture the vocabulary change in domains such as "Chemistry," for which the proposed approach was not successful. One possible solution is to build different temporal query models, representing the language usage of different time intervals, in order to capture the gradual language change in a domain.

ACKNOWLEDGMENTS

We would like to thank Jimmy Huang and Shima Gerani for their help on an early stage of this work and Morgan Harvey for his feedback on parts of this study. Last but not least, we would like to thank the anonymous reviewers for their valuable feedback that helped us improve the quality of this article.

REFERENCES

G. Amati, G. Amodeo, and C. Gaibisso. 2012. Survival analysis for freshness in microblogging search. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*. 2483–2486.

A. Arampatzis and J. Kamps. 2009. A signal-to-noise approach to score normalization. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*. 797–806.

K. H. Atkinson. 2008. Toward a more rational patent search paradigm. In *Proceedings of the ACM Workshop on Patent Information Retrieval (PaIR)*. 37–40.

L. Azzopardi and V. Vinay. 2008. Retrievability: An evaluation measure for higher order information access tasks. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*. 561–570.

S. Bashir and A. Rauber. 2010. Improving retrievability of patents in prior-art search. In *Proceedings of the European Conference on Information Retrieval (ECIR)*. 457–470.

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

S. Calegari, E. Panzeri, and G. Pasi. 2012. PatentLight: A patent search application. In *Proceedings of the 2nd Symposium on Information Interaction in Context (IIiX)*.

S. Cetintas and L. Si. 2012. Effective query generation and postprocessing strategies for prior art patent search. *J. Am. Soc. Inf. Sci. Technol.* 63, 3 (2012), 512–527.

E. D'hondt, S. Verberne, W. Alink, and R. Cornacchia. 2011. Combining document representations for prior-art retrieval. In *Proceedings of CLEF (Notebook Papers/LABs/Workshops)*.

A. Fujii. 2007. Enhancing patent retrieval by citation analysis. In *Proceedings of the ACM SIGIR Conference on Research and Developement in Information Retrieval*. 793–794.

A. Fujii, M. Iwayama, and N. Kando. 2004. Overview of patent retrieval task at NTCIR-4. In *Proceedings of the 4th NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization*.

S. Fujita. 2004. Revisiting the document length hypotheses- NTCIR-4 CLIR and patent experiments at Patolis. In *Proceedings of the 4th NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization*.

D. Ganguly, J. Leveling, W. Magdy, and G. J. F. Jones. 2011. Patent query reduction based on pseudo-relevant documents. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*. 1953–1956.

S. Gerani, M. J. Carman, and F. Crestani. 2012. Aggregation methods for proximity-based opinion retrieval. *ACM Trans. Inf. Syst.* 30, 4 (2012), 26.

J. Gobeill, E. Pasche, D. Teodoro, and P. Ruch. 2009. Simple pre and post processing strategies for patent searching in CLEF intellectual property track 2009. In *Proceedings of CLEF (Notebook Papers/LABs/Workshops)*. 444–451.

C. G. Harris, R. Arens, and P. Srinivasan. 2011. Using classification code hierarchies for patent prior art searches. In *Current Challenges in Patent Retrieval*. The Information Retrieval Services, Vol. 9, Springer, Berlin, 287–304.

M. Iwayama, A. Fujii, N. Kando, and A. Takano. 2003. Overview of the 3rd NTCIR Workshop. In *Proceedings of the ACL Workshop on Patent Corpus Processing*. 24–32.

H. Joho, L. A. Azzopardi, and W. Vanderbauwhede. 2010. A survey of patent users: An analysis of tasks, behavior, search functionality and system requirements. In *Proceedings of the 3rd Symposium on Information Interaction in Context (IIiX)*. 13–24.

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

C. H. A. Koster, M. Seutter, and J. Beney. 2003. Multi-classification of patent applications with Winnow. In *Proceedings of the 5th International Ershov Memorial Conference on Perspectives of System Informatics.* Vol. 2890, Lecture Notes in Computer Science. Springer, 546–555.

J. H. Lee. 1997. Analyses of multiple evidence combination. In *Proceedings of the ACM SIGIR Conference on Research and Developement in Information Retrieval*. 267–276.

X. Li and B. Croft. 2003. Time-based language models. In *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM)*. 469–475.

P. Lopez and L. Romary. 2009. PATATRAS: Retrieval model combination and regression models for prior art search. In *Proceedings of CLEF (Notebook Papers/LABs/Workshops)*. 430–437.

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

M. Lupu and A. Hanbury. 2013. Patent retrieval. *Found. Trends Inf. Retr.* 7, 1 (2013), 1–97.

M. Lupu, K. Mayer, J. Tait, and A. J. Trippe. 2011. *Current Challenges in Patent Information Retrieval*. Springer.

Y. Lv and C. Zhai. 2009. Positional language models for information retrieval. In *Proceedings of the ACM SIGIR Conference on Research and Developement in Information Retrieval*. 299–306.

Y. Lv and C. Zhai. 2010. Positional relevance model for pseudo-relevance feedback. In *Proceedings of the ACM SIGIR Conference on Research and Developement in Information Retrieval*. 579–586.

W. Magdy and G. J. F. Jones. 2010a. Applying the KISS principle for the CLEF-IP 2010 prior art candidate patent search task. In *Proceedings of CLEF (Notebook Papers/LABs/Workshops)*.

W. Magdy and G. J. F. Jones. 2010b. PRES: A score metric for evaluating recall-oriented information retrieval applications. In *Proceedings of the ACM SIGIR Conference on Research and Developement in Information Retrieval*. 611–618.

W. Magdy and G. J. F. Jones. 2011. A study on query expansion methods for patent retrieval. In *Proceedings of the ACM Workshop on Patent Information Retrieval (PaIR)*. 19–24.

W. Magdy, P. Lopez, and G. J. F. Jones. 2010. Simple vs. sophisticated approaches for patent prior-art search. In *Proceedings of the European Conference on Information Retrieval (ECIR)*. 725–728.

P. Mahdabi, L. Andersson, A. Hanbury, and F. Crestani. 2011a. Report on the CLEF-IP 2011 experiments: Exploring patent summarization. In *Proceedings of CLEF (Notebook Papers/LABs/Workshops)*.

P. Mahdabi and F. Crestani. 2013. The effect of citation analysis on query expansion for patent retrieval. *J. Inf. Retr.*

P. Mahdabi, S. Gerani, J. X. Huang, and F. Crestani. 2013. Leveraging conceptual lexicon: Query disambiguation using proximity information for patent retrieval. In *Proceedings of the ACM SIGIR Conference on Research and Developement in Information Retrieval*. 113–122.

P. Mahdabi, M. Keikha, S. Gerani, M. Landoni, and F. Crestani. 2011b. Building queries for prior-art search. In *Proceedings of the Information Retrieval Facility Conference (IRFC)*. 3–15.

H. Mase, T. Matsubayashi, Y. Ogawa, and M. Wayama. 2005. Proposal of two-stage patent retrieval method considering the claim structure. *ACM Trans. Asian Lang. Inf. Process.* 4, 2 (2005), 190–206.

M. H. Peetz and M. de Rijke. 2013. Cognitive temporal document priors. In *Proceedings of the European Conference on Information Retrieval (ECIR)*. 318–330.

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

F. Piroi, M. Lupu, A. Hanbury, and V. Zenz. 2011. CLEF-IP 2011: Retrieval in the intellectual property domain. In *Proceedings of CLEF (Notebook Papers/Labs/Workshop)*.

M. Salampasis, G. Paltoglou, and A. Giahanou. 2012. Report on the CLEF-IP 2012 experiments: Search of topically organized patents. In *Proceedings of CLEF (Notebook Papers/LABs/Workshops)*.

M. D. Smucker, J. Allan, and B. Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM)*. 623–632.

P. Sondhi, V. G. V. Vydiswaran, and C. Zhai. 2012. Reliability prediction of webpages in the medical domain. In *Proceedings of the European Conference on Information Retrieval (ECIR)*. 219–231.

T. Takaki, A. Fujii, and T. Ishikawa. 2004. Associative document retrieval by query subtopic analysis and Its application to invalidity patent search. In *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM)*. 399–405.

D. Teodoro, E. Pasche, D. Vishnyakova, C. Lovis, J. Gobeill, and P. Ruch. 2010. Automatic IPC encoding and novelty tracking for effective patent mining. In *Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access.* 309–317.

M. Verma and V. Varma. 2011. Exploring key-phrase extraction and IPC classification vectors for prior art search. In *Proceedings of CLEF (Notebook Papers/LABs/Workshops)*.

X. Xue and W. B. Croft. 2009a. Automatic query generation for patent search. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*. 2037–2040.

X. Xue and W. B. Croft. 2009b. Transforming patents into prior-art queries. In *Proceedings of the ACM SIGIR Conference on Research and Developement in Information Retrieval*. 808–809.

X. Yin, X. Huang, and Z. Li. 2010. Promoting ranking diversity for biomedical information retrieval using Wikipedia. In *Proceedings of the European Conference on Information Retrieval (ECIR)*. 495–507.

C. Zhai and J. D. Lafferty. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the ACM SIGIR Conference on Research and Developement in Information Retrieval*. 334–342.

Received October 2013; revised March, July 2014; accepted July 2014