Leveraging Conceptual Lexicon: Query Disambiguation using Proximity Information for Patent Retrieval

Parvaz Mahdabi[†] Shima Gerani[†] Jimmy Xiangji Huang[‡] Fabio Crestani[†]
[†]University of Lugano, Faculty of Informatics, Lugano, Switzerland
{parvaz.mahdabi, shima.gerani, fabio.crestani}@usi.ch
[‡] School of Information Technology, York University, Toronto, Canada
jhuang@yorku.ca

ABSTRACT

Patent prior art search is a task in patent retrieval where the goal is to rank documents which describe prior art work related to a patent application. One of the main properties of patent retrieval is that the query topic is a full patent application and does not represent a focused information need. This query by document nature of patent retrieval introduces new challenges and requires new investigations specific to this problem. Researchers have addressed this problem by considering different information resources for query reduction and query disambiguation. However, previous work has not fully studied the effect of using proximity information and exploiting domain specific resources for performing query disambiguation.

In this paper, we first reduce the query document by taking the first claim of the document itself. We then build a query-specific patent lexicon based on definitions of the International Patent Classification (IPC). We study how to expand queries by selecting expansion terms from the lexicon that are focused on the query topic. The key problem is how to capture whether an expansion term is focused on the query topic or not. We address this problem by exploiting proximity information. We assign high weights to expansion terms appearing closer to query terms based on the intuition that terms closer to query terms are more likely to be related to the query topic. Experimental results on two patent retrieval datasets show that the proposed method is effective and robust for query expansion, significantly outperforming the standard pseudo relevance feedback (PRF) and existing baselines in patent retrieval.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Retrieval Models, Query Formulation

General Terms

Experimentation, Performance, Measurement

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

Keywords

Patent Search, Query Expansion, Proximity Information

1. INTRODUCTION

For any new invention, in order to be granted a valid patent, the owner of the idea needs to ensure that the invention is novel, i.e., it has not been previously patented by someone else or has not been described in a scientific paper. To this end, a search for documents relevant to the invention is carried out to see if the idea in the patent application is invalidating on some existing ideas. This type of search is called *prior art search*.

This search is executed by the author himself or (someone else hired by the author) using keyword-based searches from the claims of the patent application prior to filing the patent application. The same type of search is performed by the patent examiner in the patent office after the application is filed, to ensure if the filed patent application can be granted a valid patent.

These keyword-based searches are then completed using other metadata associated with the patent applications such as (International Patent Classification) IPC classes[1] and date tags. Bibliographic information such as citations, both backward and forward, 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 [11].

Prior art search can take a very long time (days or weeks) because the searcher needs to ensure he is not missing on any relevant documents as infringing on some existing patents might result in a multi-million dollar lawsuit. Therefore, prior art search is considered as a recall-oriented application.

Patent retrieval, similar to other domain specific IR tasks [17, 19], has the following three major problems: the frequent usage of non-standardized acronyms which are invented by patent applicants, the presence of homonyms (the same word referring to two or more different entities), such as bus[2] and closet[3], and synonyms (two or more words referring to the same entity) such as signal and wave.

[1]http://www.wipo.int/classifications/ipc/en/
[2]i) motor vehicle, ii) an electronic subsystem transferring plurality of digits bits in group.
[3]i)water closet (flush toilet), ii) a small cupboard used for storing things.

Previous research [15, 9] has tackled this term mismatch problem by first forming a keyword query from the patent application (query patent) based on the frequency information, and then enhancing the query through a knowledge base such as Wikipedia or WordNet, exploiting this enhanced query to disambiguate the occurrences of query terms in the documents.

Using external resources has been shown to be more effective compared to the performance of the initial query and pseudo relevance feedback (PRF). In fact, the retrieval effectiveness of PRF has been shown to be disappointing in patent retrieval mainly due to the low MAP of the initial rank list [5].

Patent examiners use term proximity heuristics in their searches in 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 model to address proximity. "ADJn" operator which searches for terms within n words proximity in the order specified, and"NEARn" operator, which searches for the terms within n words, in either order. This usage shows that proximity information plays an important role in patent searching.

Previous work [15, 9] which uses external resources for query expansion did not take into account proximity information between query terms and related expansion concepts to form high quality expansions. Expansion terms extracted from these external resources are often general terms. Thus, it is necessary to condition their occurrences on their neighboring query terms which are more precise and consequently ignore their appearance in isolation from a query term.

In this paper, our aim is to address the term mismatch problem in the patent retrieval through query expansion. Different from previous research [15, 9], we intend to use a domain-dependent resource. We believe that using a domain-dependent resource leads to the extraction of more relevant expansion concepts. To do this, we first constructed a lexicon from IPC definition pages[4]. Definition of IPC classes consists of the explanations regarding each IPC class which can be used to identify the important concepts and subtopics of the query. We extract expansion concepts specific to each query from this lexicon for query expansion. We then use term proximity information to calculate reliable importance weights for the expansion concepts.

We propose a proximity-based query propagation method to calculate the query term density at each point in the document. Our proximity-based framework incorporates positional information into the estimation of importance of expansion concepts so that we can reward expansion concepts occurring close to query terms. This way we can concentrate on the terms that are associated with the query terms and avoid the topic drift which is caused by taking into account irrelevant terms.

Our proposed model consists of four steps. In the first step, we shorten the query document by taking the first claim of the document itself. We use the first claim as the source for building the initial query since it contains the core of the invention and serves as a source for extracting precise initial query terms.

In the second step, a query-specific lexicon is built. In the third step, query expansion is performed by deriving expansion concepts from the query-specific lexicon and positional information is used to calculate weights for insuring high quality expansions. To this end, we utilize kernel functions to keep track of the distance of the expansion concepts from the query terms. Thus words appearing within the neighborhood of a given query term are more likely to be associated with that query term. In the fourth step, the initial query and feedback runs (made using proximity information) are combined together to generate a unique rank list.

Our contributions are:

- Presenting an approach to construct a domain-dependent lexicon for identifying expansion concepts.
- Presenting a proximity-based method for estimating the probability that a specific query expansion term is relevant to the query term.
- Investigating different strategies for extracting concepts from domain-dependent lexicon for query reformulation.

We evaluate our work on two patent retrieval corpora, CLEF-IP 2010 and CLEF-IP 2011, using baselines which employ external resources for query expansion. The experimental results show that our model achieves significant improvement over these baselines in terms of recall. The results show the advantage of deploying a domain-dependent resource for selecting expansion candidate terms in contrast to systems which deploy external resources that are not dependent to the domain.

The results also confirm that our model outperforms systems which perform query expansion via PRF. Besides, the results demonstrate that utilizing the proximity information leads to the calculation of reliable weights for the expansion concepts in the process of query expansion.

2. RELATED WORK

Patent retrieval faces many challenges among which we focus on the following two. The first challenge is to reduce the original query topic in order to find a focused information need and remove the ambiguous and noisy terms. In previous work, researchers explored different sections of the query patent to perform the query reduction [18, 4]. Some of the previous work reported that effective queries were built from the query patent [18], while others obtained better results using single fields of the query patent such as claims or description [4].

The second challenge is related to query disambiguation. Previous work used different external resources for query expansion such as Wikipedia [9] and WordNet [15] with the goal of query disambiguation. The goal of this task is to alleviate the term mismatch problem by expanding the query with all words related to the concept of the query or synonyms of the query terms.

Our proximity-based framework is inspired by the original work of Lv and Zhai on positional language model and positional relevance model [12, 13]. An advantage of Lv and Zhai's work is that they can capture passage level evidence in a "soft" way by modeling proximity information via density functions. Their experiments confirmed that such approach works better than applying a "hard" boundary of passages.

Proximity information has shown to be useful in different IR tasks such as opinion mining [6]. Authors investigate

[4]http://web2.wipo.int/ipcpub/

proximity information for capturing the opinion density at each point in the document.

Term position and proximity cues were mostly ignored in previous work in patent retrieval. Recently, Ganguly et al.'s work on reducing query patent using PRF captures term position and proximity evidence indirectly through the use of appropriate passages [5]. This work provides a general model for query reduction using PRF. In fact, this passage-based feedback model is orthogonal to our approach, in the sense that it can be used as a pre-processing for our query expansion method. Our proximity-based query expansion framework can be applied on their reduced query. This way we can ensure a more precise set of query terms as the initial query.

Another recent study, Bashir and Rauber's work on improving retrievability of patent documents [3], 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 [2] of individual patents using proximity heuristics compared to the standard PRF. However, they did not evaluate directly the performance of their approach in terms of retrieval effectiveness.

Finally, Mahdabi et al.'s work used a learning approach to predict the quality of a query [16]. This work uses proximity information loosely through the use of noun phrases. This work is complementary with our ideas as our proximity-based method for estimating the probability of query relatedness of an expansion term can be integrated as a feature in their learning-based model to improve their performance.

3. MULTIPLE INFORMATION SOURCES

We identify different information sources that can be used as additional knowledge for query reformulation in patent retrieval. In this section we summarize and categorize them.

- **Query patent**: This document is a structured document which is composed of the following sections: *title*, *abstract*, *description*, and *claims*. A claim which does not reference any other claim is called an *independent claim* and others are called *dependent claims* [10]. The independent items in the claims of the patent document comprise the kernel of the technical innovation of a patent document. Among the claims the most important claim is the first independent claim 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 [11].
- **IPC classification**: The International Patent Classification (IPC classification) provides a hierarchical categorization over different technological fields such as computer science, electronics, mechanics, materials science, and bio-chemistry. Such classes are language independent keywords 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 [11]. For each conceptual tag there is textual descriptions (IPC definition pages) available which is an additional source of information, providing contextual cues about different technical fields.
- **Retrieval corpus**: Our retrieval corpus can be used as another source for extracting expansion concepts via the procedure of PRF.

The above sources have different vocabulary usage. The query patent itself has an obscure style of writing (patentese) [11]. This characteristic might create term mismatch problem for finding relevant documents in the existing filed patents, such that a query term may not be a good indicator in referring to a technological concept as it is not frequently used in the given context by other authors.

However, the two other resources provide a more established vocabulary usage. The descriptions of IPC classification represent the standard vocabulary usage related to different domains. This source provides a general but accepted vocabulary usage for explaining technological concepts. Performing query expansion via PRF allows extracting a set of terms which are not used by the author but are frequent in the retrieval corpus. Thus, the vocabulary usage of the two latter sources are complementary to the query itself.

It is worth to mention that previous work [8] used the conceptual tags for filtering the patent documents. In this work, we propose to exploit the textual description of such conceptual tags in addition to the tags themselves.

4. BUILDING DOMAIN-DEPENDENT LEXICON

We now explain the process of building a lexicon from IPC definition pages. We refer to this lexicon as a *conceptual lexicon*. We first perform stop-word removal on the text of IPC definition pages. We then build a language model for each IPC class and detect terms with document frequency higher than 10 (based on experiments). We refer to these terms as patent-specific stop-words and we 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. These terms are extracted from the IPC definition pages as previously described. An example of an entry in the conceptual lexicon is presented in Table 1.

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

Table 1: An entry in the conceptual lexicon.

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

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 (a non-standard terminology), while conceptual lexicon provides a more standard terminology. We try to combine these two complementary vocabularies, as we believe this will help alleviate the term mismatch problem.

In the next sections we will explain with more detail how the lexicon is used for query expansion.

5. A PROXIMITY-BASED FRAMEWORK

Our hypothesis is that an author will use standard terminology from a conceptual lexicon for clarifying his invented concepts. We note that a query term might belong to the author terminology or some terminology that is commonly used in a domain but is not as conventional as the vocabulary of the conceptual lexicon.

We now focus our attention to identify expansion concepts in the document that are referring to the concepts in the query. These expansion concepts can be extracted from one of the sources explained in the previous section. We need to estimate the probability that an expansion term is referring to a query term. To do this, we rely on the structure of the document. We assume that an expansion term refer with higher probability to the query terms closer to its position. We thus regard the distance of an expansion term to the query term as a measure of relatedness.

5.1 Estimating the Query Relatedness

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

$$P(q|i,d) = \sum_{j=1}^{m} P(q|t_j)P(j|i,d) \quad (1)$$

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

$$P(j|i,d) = \frac{k(j,i)}{\sum_{j'=1}^{|d|} k(j',i)} \quad (2)$$

where $K(i,j)$ is the kernel function which determines the weight of propagated query-relatedness from t_j to t_i. We model the query relatedness by placing a density kernel function around query terms. We use the same notation throughout the paper.

In the following, we present different kernels used in our experiments. We study three different density functions, namely Gaussian, Laplace, and Rectangle kernel. We selected Gaussian and Laplace kernels as they have been shown to be the best performing kernels among the kernel functions tested in the previous work [12, 6]. We also chose Rectangle kernel to simulate the effect of imposing a hard boundary over passages in contrast to the soft boundary introduced by other kernels.

- **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 kernels which favor expansion term occurrence in close proximity of query terms or not.

5.2 Calculating Document Relevance Score

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

$$P(q|d,e) = \sum_{i=1}^{|d|} P(q,i|d,e) = \sum_{i=1}^{|d|} P(q|i,d,e)P(i|d,e) \quad (3)$$

We assume e and q are conditionally independent given the position in the document. 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:

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

by substituting this in Equation 3 we have:

$$P(q|d,e) = 1/|pos(e)| \sum_{i \in pos(e)} P(q|i,d) \quad (4)$$

- **Max Position Strategy**: As an alternative, we can assume that only the expansion term position where $P(q|i,d)$ is maximum is important:

$$P(q|d,e) = \max_{i \in pos(e)} P(q|i,d) \quad (5)$$

Example 1 (Topic ID: pac-1474) Patent title: "Optical information recording medium" Query terms extracted from first independent item of claims: optical, layer, record, lens, light, interlay, irradiation, wavelength Expansion concepts selected from the conceptual lexicon related to the query: organic, dielectric, sensitizing, record, reproduction
Retrieved docs for example 1 Number of retrieved relevant documents in the baseline run: 15/42 Number of retrieved relevant documents after using proximity-based method: 24/42
Example 2: (Topic ID: pac-552) Patent title: "Power supplying apparatus, design method of the same, and power generation apparatus" Query terms extracted from first independent item of claims: power, supply, boost, transformer, switch, resonance Expansion concepts selected from the conceptual lexicon related to the query: conversion, semiconductor, electrode, light, push-pull
Retrieved docs for example 2 Number of retrieved relevant documents in the baseline run: 6/15 Number of retrieved relevant documents after using proximity-based method: 12/15

Table 2: Examples where using proximity-based information with IEC method improves recall in patent retrieval.

6. QUERY REFORMULATION

In this section, given an input query $Q = \{q_1, q_2, ..., q_n\}$, we can identify a set of concepts $C_E = \{e_1, e_2, ..., e_m\}$ which are selected from the conceptual lexicon. The set of C_E is associated to the query Q as the conceptual lexicon contains explanations about the IPC classes to which a patent document is assigned. Once the set of concepts C_E is identified, we determine the concept 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 1 demonstrates the process for calculating the importance weight for expansion concepts. We can then re-rank documents in the original rank list $\mathbb{R}$ using a weighted combination of the matches of concepts in C_E and our original keyword query Q based on Equation 3.

We use four different strategies for the expansion concept selection process. We categorize and summarize these strategies below.

Explicit Expansion Concepts In this setting, we use the concepts in our conceptual lexicon which match against the IPC classification of the query. However, we restrict our attention to concepts that appear in D_Q. This provides a set of explicit expansion concepts (a subset of C_E) which serves as candidate expansion terms. We refer to this set as X_E. We utilize the proximity of query terms and expansion terms inside query document D_Q to assign importance weights to the explicit expansion concepts. These weights are then used to re-rank 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 the documents inside $\mathbb{R}$ for computing the importance weights associated with the expansion terms. The advantage of this strategy compared to the previous one is that we are able to make use of all terms available in the C_E and we are not limited to the query document.

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, ...

Table 3: Comparison between the list of expansion terms derived from the information sources for the query with title "inkjet recording ink".

Combining Search Strategies In this strategy, instead of query expansion, we first calculate a relevance score based on the original keyword query Q. We then calculate an IPC score based on the expansion concepts in C_E. We linearly combine the two scores together. Our goal is to compare whether having a unified query, as 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 professional searchers for retrieving relevant documents [11]. In such a search strategy, searchers perform separate searches based on different information sources, such as the query document and IPC classification, and then combine the results of the runs together to produce a unique rank list.

Proximity-based Pseudo Relevance Feedback As a comparison baseline we also used the retrieval corpus as a source for PRF where we used the feedback set for selecting expansion terms and identifying their weights. We use the distance between the query terms and the candidate expansion concepts inside the feedback documents to calculate the weight for the expansion concepts.

As an example, Table 3 shows the terms selected from different information resources. The terms from the query document is selected from the first item of the claims. The terms from the retrieval corpus are selected via the procedure of PRF.

7. EXPERIMENTAL SETUP

In this section we first explain our experimental setup for evaluating the effectiveness of our proposed approach.

Testing Collections We used Terrier Information Retrieval System[5] to index the collection with the default stemming and stop-word removal. We removed patent-specific stop-words such as "device" and "method".

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 and CLEF-IP 2011 consists of 3 million patent documents. In our experiments we used the English subsection of both collections. The English test set of CLEF-IP 2010 corresponds to 1348 topics. The English test set of CLEF-IP 2011 consists of 1351 topics. We used the training topics of CLEF-IP 2010 for tuning some parameters of our model. This training set consists of 300 topics.

Evaluation We used the relevance judgment for the English language provided by CLEF-IP for evaluation. We report the Recall, Mean Average Precision, and Patent Retrieval Evaluation Score [14] which combines MAP and Recall in one single score. We report the evaluation metrics on the top 1000 results. In the remainder of our experiments we used the Wilcoxon signed ranked matched pairs test with a confidence level of 0.05 level for testing statistical significance improvements.

8. EXPERIMENTAL RESULTS

8.1 Building the Initial Query

We built keyword queries by extracting distinguishing terms from the query patent document. To this end, we estimated the importance of each term according to a weighted log-likelihood based approach, as in [16]. We used the claims and the first independent claim for selecting query terms. Table 4 summarizes the results we obtained for the topics in the training and test set of CLEF-IP 2010 and the test set of CLEF-IP 2011. We used top-10 query terms with higher weights from the estimated query model in our experiments. Results marked with † and ‡ achieve statistically significant improvement in terms of MAP and recall, respectively. Note that this comparison is performed among runs belonging to the same experimental settings.

The results of Table 4 demonstrate that the performance of the runs obtained by issuing the query built from the first-claim is always stronger than the performance of the runs where the query is built from the claims in terms of MAP. However, the opposite holds for recall. The reason for the high MAP is because the first independent item of claims is focused on the core invention of the patent document. However, we are losing some information by ignoring the text of the rest of the claims and this explains the low recall of this setting.

To guarantee the assignment of reliable importance weights to the expansion concepts in our proximity-based framework, we need to start with a set of precise query terms. This is because we rely on the distance between query terms and expansion concepts to calculate importance weights for expansion concepts. Obviously starting with focused and less noisy query terms has a direct effect on the quality of calculated importance weights. Thus, in the remainder of this paper, we focus on selecting query terms from the first independent item of the claims.

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

CLEF-IP 2010 (training topics)				
Method	Run description	MAP	Recall	PRES
C10TR	claims	0.1211	0.6302 ‡	0.5492
FC10TR	first-claim	0.1530 †	0.6015	0.5479
CLEF-IP 2010 (test topics)				
Method	Run description	MAP	Recall	PRES
C10TE	claims	0.1293	0.6067 ‡	0.5140
FC10TE	first-claim	0.1445 †	0.5624	0.4911
CLEF-IP 2011 (test topics)				
Method	Run description	MAP	Recall	PRES
C11TE	claims	0.0823	0.5905 ‡	0.4850
FC11TE	first-claim	0.1198 †	0.5360	0.4538

Table 4: Choosing baseline on the two retrieval collections.

We used the Language Modeling approach with Dirichlet smoothing [20] to score documents from both collections and build the initial rank lists. We empirically set the value for the smoothing parameter μ to1500. We also used Language Modeling for the re-ranking of the results. We note that we do not use citation information in our experiments.

8.2 Choosing the Baseline

In terms of our comparison baseline, we chose the strongest configuration in terms of PRES from Table 4, the retrieval run where query terms are selected from the claims section of the patent document. C10TR is chosen as the baseline on the training topics of CLEF-IP 2010. C10TE is chosen as the baseline on the test topics of CLEF-IP 2010 and C11TE is chosen 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 C10TR in such comparisons.

Table 5 shows the performance of strong baselines of the previous work [5, 15] over CLEF-IP 2010. We presented their performance evaluation in terms of MAP and PRES as the recall values were not reported for the two baselines.

method	MAP	PRES
baseline1 [5]	0.1278	0.4604
baseline2 [15]	0.1399	0.4860

Table 5: Strong baselines of the previous work

As we can see from the results of Table 5, C10TE is as strong as the baseline1 and baseline2 in terms of PRES. This ensures the selection of a strong baseline which will be used in evaluating the performance of our proposed model in the rest of the paper.

8.3 Motivation for Using Proximity Information

In order to test if closeness of expansion concepts to the query terms is correlated with relevance, we carry out preliminary experiments on the CLEF-IP 2010 collection. In these experiments, we selected 100 random queries.

For each query, we first retrieved the top 100 documents

IEC				
kernel \ σ	25	75	125	150
Gaussian	**0.6443**	**0.6561** †	**0.6676** †	**0.6795** †
Laplace	0.6422	0.6556 †	0.6588 †	0.6709 †
Rectangle	0.6398	0.6523	0.6559 †	0.6678 †

EEC				
kernel \ σ	25	75	125	150
Gaussian	**0.6388**	**0.6418**	0.6669 †	**0.6637** †
Laplace	0.6362	0.6390	**0.6685** †	0.6516
Rectangle	0.6339	0.6375	0.6642 †	0.6497

Table 6: Recall results of different settings of the kernel functions using query reformulation methods (IEC and EEC) on the training topics of CLEF-IP 2010.

using a Language Modeling retrieval method. We separated relevant and non-relevant documents according to relevance judgements (qrels). We then looked at the average distance between query terms and expansion concepts inside the set of relevant documents, denoted by R, and the set of non-relevant documents, denoted by $\bar{R}$. The distance in each of the two mentioned sets is calculated as follows:

$$DIS(Q,R) = \sum_{D \in R} \frac{\sum_{q \in Q} min_{e \in E}(Dis(q,e))}{|Q||R|}$$

$$DIS(Q,\bar{R}) = \sum_{D \in \bar{R}} \frac{\sum_{q \in Q} min_{e \in E}(Dis(q,e))}{|Q||\bar{R}|}$$

where q denotes a query term drawn from the set of query terms, denoted by Q, e denotes an expansion term, and E denotes the set of expansion concepts selected from the conceptual lexicon. The distance between two terms is calculated in terms of the number of terms between them. The minimum is calculated among all the occurrences of each pair of query term and expansion term.

Figure 1 shows the average distance for relevant and non-relevant documents for each query topic. For clarity purposes, topics are sorted by the average distance DIS of their relevant documents.

It can be seen from Figure 1 that the minimum distance between an expansion term and a query term in relevant documents is less than their respective distance in non-relevant documents. Therefore we can use this proximity information to differentiate the relevant documents from non-relevant documents and to improve the ranking of relevant documents.

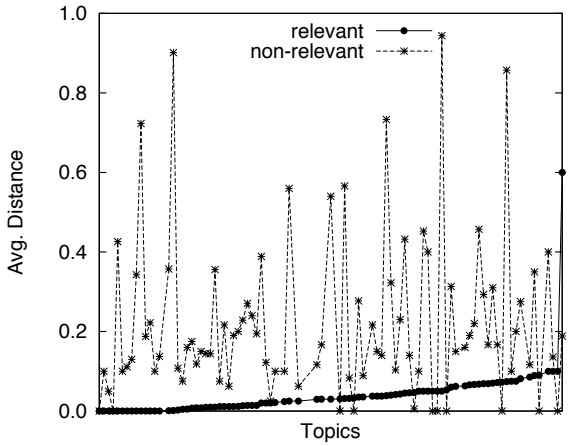


Figure 1: $AverageDIS$ value for each topic using qrels and top-retrieved non-relevant documents.

8.4 Effect of Density Kernel

We are interested to investigate the effectiveness of different query reformulation methods proposed in Section 6 for scoring documents in our proximity-based framework. The results of this comparison are summarized in Table 6.

In all the comparisons, our query expansion method which uses *explicit expansion concept* is denoted as EEC. The query expansion method which uses *implicit expansion concept* is referred to as IEC.

Since the performance of these methods is directly determined by the effectiveness of the kernel function used to estimate the propagated query relatedness probabilities for the expansion concepts, we first need to compare three different proximity-based kernel functions to see which one performs the best.

We place a density kernel around each occurrence of query term positions in the document as previously explained in Section 5. 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 which occurs at a position close to many query terms will receive high query relatedness and thus will obtain a higher importance weight.

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.

The results of comparing different kernel functions on the training topics of CLEF-IP 2010 are shown in Table 6. A † denotes statistical significant improvement over C10TR and the best result for each kernel type is highlighted. The results show that the performance of EEC and IEC with all kernel functions improve over C10TR.

It is also clear that among all the kernel functions, the Gaussian kernel outperforms other types of kernels in most cases. Since the Gaussian kernel performed the best in most of the carried experiments, we use this kernel function for our system evaluation 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 of [25, 200] with a step of 25 similar to the previous work [12, 13]. Table 6 reports the performance of different kernel functions using varying values of σ. The results show that selecting a value of 125 or 150 usually gives the best retrieval performance.

Overall, the results of Table 6 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. We thus use the σ value of 150 in the rest of our experiments. In Section 8.5 we further study the performance of the query reformulation

Collection	metric	IEC	EEC	CSS	PPRF
CLEF-IP 2010	MAP	0.1050	0.1026	0.0982	0.0705
	Recall	0.6595 †	0.6437 †	0.6241	0.5877
	PRES	0.5540	0.5498	0.5354	0.5023
CLEF-IP 2011	MAP	0.0772	0.0761	0.0738	0.0629
	Recall	0.6371 ‡	0.6254 ‡	0.6088	0.5632
	PRES	0.5288	0.5249	0.5127	0.4945

Table 7: The performance results of query reformulation approaches on two patent retrieval datasets on the test topics of CLEF-IP 2010 and CLEF-IP 2011.

methods.

Comparison of Max and Avg Strategy We are interested to evaluate the two strategies for calculating the probability of relevance of a document as proposed in Section 5.2. Table 8 shows the result of using avg and max strategies for different sigma values on the training topics of CLEF-IP 2010 using the IEC reformulation method.

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

method \ σ	25	75	125	150
max	0.6443 †	0.6561 †	0.6676 †	0.6795 †
avg	0.6164	0.6198	0.6207	0.6238

Table 8: Recall of the Max and Avg method using Gaussian kernel with IEC reformulation method on training topics of CLEF-IP 2010.

8.5 Effect of Query Reformulation

In this section, we present the evaluation results of our proposed approaches on the topics in the test set of CLEF-IP 2010 and CLEF-IP 2011.

Table 7 reports the retrieval performance of query reformulation methods described in Section 6. The symbols † and ‡ denote statistical significant improvements over C10TE and C11TE, respectively.

We now compare the performance of our query formulation methods. In addition to EEC and IEC which were introduced earlier, the results of the two other query reformulation methods (described in Section 6) are presented in Table 7. Our method which *combines search strategies* is denoted as CSS. The last method in our comparisons is the *positional-based pseudo relevance feedback*, which is denoted by PPRF.

The main observation from Table 7 is that IEC is always more effective than the other three methods. In addition, IEC improves the baseline in terms of recall on both collections significantly.

Table 7 shows that a method which uses a conceptual lexicon for selecting expansion terms outperforms a method which uses feedback documents for identifying expansion terms. This is evident by comparing the performance of EEC, IEC and CSS to the performance of PPRF, as 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 7 demonstrate that IEC obtains 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 itself. This diminishes the power of EEC in contrast to IEC, and explains the advantage of IEC. Results confirm that the unlimited usage of the conceptual lexicon is superior to the limited usage of it.

Another observation which can be made from Table 7 is that CSS attains worst results compared to both EEC and IEC. This is perhaps due to the fact that some information is lost during the combination of two separate runs made from the query terms and expansion terms. While, in EEC and IEC we use a unified query which is composed of query terms and expansion terms.

Overall, the results of Table 7 show that using the conceptual lexicon as a domain-dependent external resource is effective in terms of recall, although this improvement does not hold for precision. Table 2 shows some queries where using IEC reformulation method improved the recall.
We used 40 expansion terms (based on initial experiments) in each of the query reformulation methods. We studied the effect of the number of expansion terms on the performance of each method. We report the result of this study in Section 8.7.

8.6 Comparison to Standard PRF

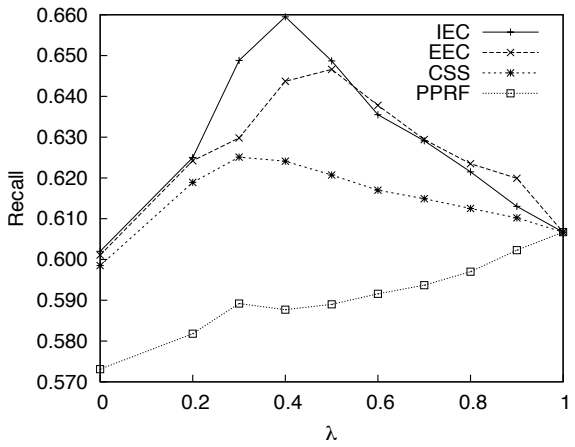
Table 9 reports the retrieval performance of PPRF compared to PRF. A ‡ indicates the statistical significant improvement over the baseline which is built from the first-claim presented in Table 4. A † denotes the statistical significant improvement over standard PRF in terms of recall.

As previously explained in Section 6, PPRF is similar to PRF since they both use feedback set for expansion term selection. However, PPRF uses proximity information inside feedback set to calculate weight for expansion terms in contrast to standard PRF.

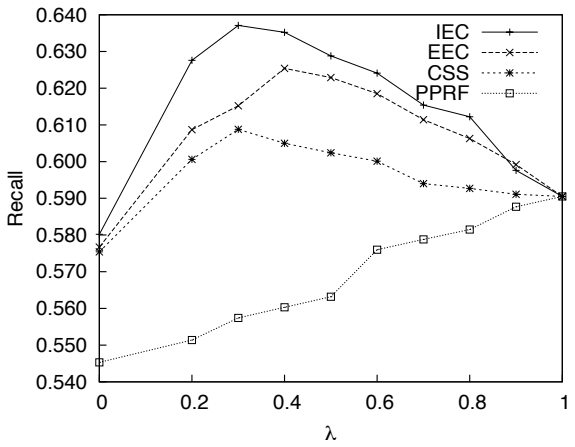
The results show that PPRF performs significantly better than the standard PRF. This result confirms the usefulness of proximity information for identifying importance weights for expansion terms as previously was shown in [13].

Note that PPRF and PRF does not achieve improvement over the baseline, but a fair comparison is to compare the retrieval effectiveness after query expansion with the retrieval effectiveness before query expansion. We thus need to compare the results of Table 9 with the results of FC10TE and FC11TE which correspond to the performance of the initial query built from the first claim.

Our results show that the performance obtained with PPRF method achieves statistical significant improvements in terms



a) CLEF-IP 2010



b) CLEF-IP 2011

Figure 2: Sensitivity to the λ coefficient in the linear combination of results made from the initial and the expanded query

of recall over the initial query (before expansion). This result is interesting as a recent study [5] pointed out that often standard PRF method fails on the patent prior art search. This poor performance is associated to the low MAP of the initial rank list from which feedback documents are selected. This comparison demonstrates the usefulness of aggregating the proximity information in the calculation of the expansion weights as performed in our proximity-based framework.

We fixed the number of feedback documents used in both PPRF and PRF to 10.

8.7 Influence of Different Parameter Settings

In this section, we are interested to study the influence of different parameters on the effectiveness of our proposed methods. We used the test topics of both corpora during the evaluations.

Collection	metric	PPRF	PRF
CLEF-IP 2010	MAP	0.0705	0.0650
	Recall	0.5877 ‡†	0.5630
	PRES	0.5023	0.4961
CLEF-IP 2011	MAP	0.0629	0.0617
	Recall	0.5632 ‡†	0.5346
	PRES	0.4945	0.4792

Table 9: The comparison of performance results of PRF and PPRF.

Number of Expansion Terms To see the effect of the number of expansion terms on the effectiveness of our proposed methods we plot the sensitivity of different query reformulation methods to the number of expansion terms over CLEF-IP 2010 test topics. We change the number of expansion terms from 1 to 50. The recall results are shown in Figure 3. We observe that all four methods achieve effective performance using around 40 expansion terms.

Effect of Combination In all configurations of our experiments we linearly combined the results we got from each

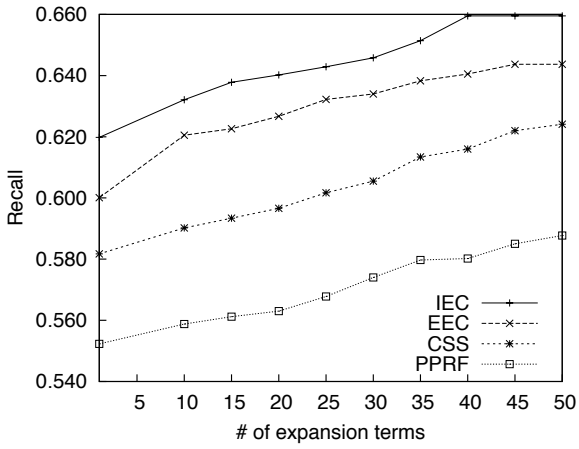


Figure 3: Sensitivity to the effect of number of expansion terms on CLEF-IP 2010

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

Figure 2 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 range around 0.4.

Effect of Normalization

We now compare the effect of different normalization methods prior to linear combination using two score normalization methods, MinMax [7] and HIS normalization [1], which are used in distributed information retrieval or meta-search.

MinMax normalization method shifts and scales scores to be between zero and one. While, 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 we applied HIS normalization. We refer to this method as MinMax-HIS throughout the experiments.

Table 10 shows the comparison among different normalization methods. These results correspond to the final performance of each run after the combination over test topics of CLEF-IP 2010. The results are obtained with the IEC method. The best results are highlighted although the difference is not statistically significant.

metric	MinMax	HIS	MinMax-HIS
MAP	0.0924	0.0991	**0.1050**
Recall	0.6520	0.6568	**0.6595**
PRES	0.5473	0.5522	**0.5540**

Table 10: The comparison of different normalization methods over CLEF-IP 2010 using IEC method

We observe that IEC achieves the best performance using MinMax-HIS normalization. The results of other methods were also confirming that applying normalization using MinMax-HIS was better compared to either MinMax or HIS alone, although not significantly. We thus presented the results of normalization using MinMax-HIS method throughout the paper.

9. CONCLUSION AND FUTURE WORK

In this paper we introduced a proximity based framework for query expansion which utilizes a conceptual lexicon for patent retrieval. To this end, we constructed a domain-dependent conceptual lexicon which can be used as an external resource for query expansion. Our proximity-based retrieval framework provides a principled way to calculate the importance weight for expansion terms selected from the conceptual lexicon. We showed that proximity of expansion terms to query terms is a good indicator of the importance of the expansion terms. In this paper we focused on performing query expansion with single terms to ensure the efficiency of the expansion concept selection process.

We have evaluated our proposed method on two patent retrieval corpora, namely CLEF-IP 2010 and CLEF-IP 2011. Our query formulation method, IEC, was shown to outperform the strong baselines of CLEF-IP and a standard pseudo relevance feedback method in terms of recall. Further analysis of the performance of the query reformulation methods proposed in this paper showed the high quality of expansion terms extracted from the conceptual lexicon.

10. ACKNOWLEDGMENTS

The work of Parvaz Mahdabi was funded by the Information Retrieval Facility, through the research project "Interactive Patent Search (IPS)". This research is also supported in part by the research grant from Natural Sciences and Engineering Research Council (NSERC) of Canada. We thank anonymous reviewers for their thorough review comments on this paper.

11. REFERENCES

[1] A. Arampatzis and J. Kamps. A signal-to-noise approach to score normalization. In *CIKM*, pages 797–806, 2009.

[2] L. Azzopardi and V. Vinay. Retrievability: an evaluation measure for higher order information access tasks. In *CIKM*, pages 561–570, 2008.

[3] S. Bashir and A. Rauber. Improving retrievability of patents in prior-art search. In *ECIR*, pages 457–470, 2010.

[4] S. Cetintas and L. Si. Effective query generation and postprocessing strategies for prior art patent search. *JASIST*, 63(3):512–527, 2012.

[5] D. Ganguly, J. Leveling, W. Magdy, and G. J. F. Jones. Patent query reduction based on pseudo-relevant documents. In *CIKM*, pages 1953–1956, 2011.

[6] S. Gerani, M. J. Carman, and F. Crestani. Aggregation methods for proximity-based opinion retrieval. *TOIS*, 30(4):26, 2012.

[7] J.-H. Lee. Analyses of multiple evidence combination. In *SIGIR*, pages 267–276, 1997.

[8] P. Lopez and L. Romary. Patatras: Retrieval model combination and regression models for prior art search. In *CLEF (Notebook Papers/LABs/Workshops)*, pages 430–437, 2009.

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

[10] M. Lupu and A. Hanbury. Patent retrieval. *Foundations and Trends® in Information Retrieval*, 7(1):1–97, 2013.

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

[12] Y. Lv and C. Zhai. Positional language models for information retrieval. In *SIGIR*, pages 299–306, 2009.

[13] Y. Lv and C. Zhai. Positional relevance model for pseudo-relevance feedback. In *SIGIR*, pages 579–586, 2010.

[14] W. Magdy and G. J. F. Jones. PRES: A score metric for evaluating recall-oriented information retrieval applications. In *SIGIR*, pages 611–618, 2010.

[15] W. Magdy and G. J. F. Jones. A study on query expansion methods for patent retrieval. In *PAIR 2011 - CIKM*, pages 19–24, 2011.

[16] P. Mahdabi, L. Andersson, M. Keikha, and F. Crestani. Automatic refinement of patent queries using concept importance predictors. In *SIGIR*, pages 505–514, 2012.

[17] P. Sondhi, V. G. V. Vydiswaran, and C. Zhai. Reliability prediction of webpages in the medical domain. In *ECIR*, pages 219–231, 2012.

[18] X. Xue and W. B. Croft. Automatic query generation for patent search. *CKIM*, pages 2037–2040, 2009.

[19] X. Yin, X. Huang, and Z. Li. Promoting ranking diversity for biomedical information retrieval using wikipedia. In *ECIR*, pages 495–507, 2010.

[20] C. Zhai and J. D. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR*, pages 334–342, 2001.