On Term Selection Techniques for Patent Prior Art Search

Mona Golestan Far

mona.golestanfar@anu.edu.au

A thesis submitted for the degree of
Masters of Philosophy
The Australian National University

May 2015

mona.golestanfar@anu.edu.au 2015

Except where otherwise indicated, this thesis is my own original work.

Mona Golestan Far

mona.golestanfar@anu.edu.au

12 May 2015

Declaration

I hereby declare that this thesis is my original work which has been done in collaboration with other researchers and me as the main author. This document has not been submitted for any other degree or award in any other university or educational institution. **Chapter 4** of this thesis has been published in collaboration to other researchers to international conference **SIGIR 2015**:

> Golestan Far, M.; Sanner, S.; Bouadjenek, R.; Ferraro, G.; and Hawking, D., 2015. On term selection techniques for patent prior art search. In *Proceedings of the 38th international ACM SIGIR conference on Research & development in information retrieval*. ACM.

To my Mom and Dad who have been the symbols of intelligence, wisdom, honesty, love, and devotion for me.

Acknowledgments

Upon the accomplishment of this thesis, I would like to thank various people: professors, colleagues, researchers, and friends. First, I would like express my gratitude to the main director of this research: Scott Sanner. I have been quite fortunate to work on another research advised by him before he leaves NICTA/ANU to Oregon State University. In my weekly meetings and Scott's reading group, I learned (1) to work hard, but with enthusiasm, (2) to feel responsibility to the contribution we make to the science, (3) to be an independent and efficient thinker (4) to collaborate with other researchers, and (5) to deal with a research problem by proposing to-the-point research questions and brave enough to examine new ideas and possible solutions. I hope I have learned enough to continue my academic career without his ultimate and unselfish support. I believe that the science world needs more people like Scott. The next scientist who I owe a debt of gratitude to is Bob Williams — the leader of the Machine Learning group of NICTA. He created such a positive research atmosphere with a collection of excellent researchers at NICTA, which made my tenure at NICTA a fruitful experience. I also thank my other committee members, Tom Gedeon, Gabriela Ferraro, and Hanna Suominen for their supports over the course of my MPhil.

During my one-year MPhil program, I met many successful IR researchers like Paul Thomas (and his IR & friends seminars), Milad Shokuhi, and Leif Azzopardi who left a strong impact on my career. I would also thank David Hawking for his comments on my work. Dave and Scott's support encouraged me to have a submission to SIGIR, which got accepted ◡̈. I appreciate Reda Bouadjenek, for his contribution to the baseline IR framework, as well Ehsan, for patiently answering my technical questions (Of course, Scott taught me a valuable lesson, which will be with me forever: "Google is your best friend!" ◡̈).

And finally, I am grateful to all of my family and friends for being there for me whenever I needed. An special thanks to my dearest Ali for always being my main support after my parents and respecting my attempts to follow my goals and values.

Abstract

A patent is a set of exclusive rights granted to an inventor to protect their invention for a limited period of time. Patent prior art search involves finding previously granted patents, scientific articles, product descriptions or any other published work to a new patent application. Many well-known Information Retrieval (IR) techniques, which are proven effective for web search such as typical query expansion methods, are unsuccessful for patent prior art search. In this thesis, we mainly investigate the reasons that generic IR techniques are not effective for prior art search on the CLEF-IP test collection. First, we analyse the errors caused due to data curation and experimental settings like applying International Patent Classification codes assigned to the patent topics to filter the search results. Then, we investigate the influence of term selection on retrieval performance on the CLEF-IP prior art test collection, starting with the Description section of the reference patent and using Language Models (LM) and BM25 scoring functions. We find that an oracular relevance feedback system, which extracts terms from the judged relevant documents far outperforms the baseline and performs twice as well on Mean Average Precision (MAP) as the best competitor in CLEF-IP 2010. We find a very clear term selection value threshold for use when choosing terms. We also noticed that most of the useful feedback terms are actually present in the original query and hypothesise that the baseline system could be substantially improved by removing negative query terms. We tried four simple automated approaches to identify negative terms for query reduction but we were unable to improve on the baseline performance with any of them. However, we show that a simple, minimal feedback interactive approach where terms are selected from only the first retrieved relevant document outperforms the best result from CLEF-IP 2010, suggesting the promise of interactive methods for term selection in patent prior art search.

Contents

List of Figures

List of Tables

Introduction

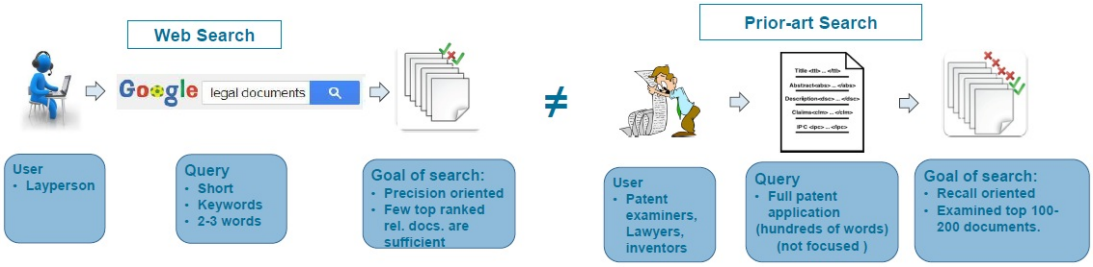


Figure 1.1: The main differences between patent prior art search and an standard web search are: (i) queries are reference patent applications, and (ii) patent prior art search is a recall-oriented task.

1.1 Motivation

Patent prior art search involves finding previously granted patents, or any published work, such as scientific articles or product descriptions that may be relevant to a new patent application. The objective and challenges of standard formulations of patent prior art search are different from those of standard text and web search [Magdy, 2012].

The main characteristic of prior art search is that queries are reference patent applications, which consist of documents with hundreds or thousands of words organised into several sections, while typical queries in text and web search constitute only a few words. In addition, in contrast to scientific and technical writers, patent writers tend to generalise and maximise the scope of what is protected by a patent and potentially discourage further innovation by third parties, which further complicates the task of formulating queries. Searching based on patent queries helps patent examiners to save time and avoid formulating appropriate search queries out of long and difficult patent applications. In general, patent examiners spend about 12 hours to complete an invalidity task by examining approximately 100 patent documents retrieved by 15 different queries in average [Joho et al., 2010].

Another important characteristic of patent prior art search is being a recall-oriented

task, where the primary focus is to retrieve all relevant documents at early ranks, in contrast to text and web search that are precision-oriented, where the primary goal is to retrieve a subset of documents that best satisfy the query intent (according to [Zhang and Kamps, 2010], 44.5% of web search users examine only one retrieved document). In prior art search, missing relevant documents is unacceptable because of the highly commercial nature of patents coupled with the high costs involved in creating a patent and infringing patented material [Joho et al., 2010].

The main users of patent prior art search are patent analysts, who are employed to determine the patentability of applications. We do not consider inventors, who want to determine whether their ideas are novel, as users of patent prior art search, because the prior art search starts with querying by a patent application that is not written when the author is going to check its novelty. Patent searchers have to perform an exhaustive and comprehensive search.

Users can save time if they start searching for prior art with a patent document as a query. However, this approach is less effective than web search [Lupu et al., 2013a]. In this thesis, we study query reformulation to transform an initial query (i.e., patent document) to another query to improve retrieval effectiveness. We mainly emphasise on query term selection techniques to formulate a query, which achieves the highest performance.

1.2 Summary

In this work, we focus on the task of query reformulation [Baeza-Yates and Ribeiro-Neto, 2011] specifically applied to patent prior art search [Mahdabi and Crestani, 2014; Piroi, 2010; Xue and Croft, 2009b]. While prior work has largely focused on specific techniques for query reformulation, we first build an oracular query formed from known relevance judgments for the CLEP-IP 2010 Prior Art test collection [Piroi, 2010] in an attempt to derive an upper bound on performance of standard Okapi BM25 and Language Models (LM) retrieval algorithms for this task. Since the results of this evaluation suggest that query reduction methods can outperform state-of-the-art prior art search performance, we proceed to analyze four simple automated methods for identifying terms to remove from the original patent query. Finding that none of these methods seems to independently yield promise for query reduction that strongly outperforms the baseline, we evaluate an alternative interactive feedback approach where terms are selected from only the first retrieved relevant document. Observing that such simple interactive methods for query reduction with a standard LM retrieval model outperform highly engineered patent-specific search systems from CLEF-IP 2010, we conclude that interactive methods offer a promising avenue for simple but highly effective term selection in patent prior art search.

1.3 Contributions

This thesis proposes term selection techniques for patent prior art search. The main contributions of this work are summarized as follows:

- We developed an oracular relevance feedback system which extracts terms from the judged relevant documents that far outperformed the the baseline and around twice as well on MAP as the best competitor in CLEF-IP 2010. Experiments related to oracular system suggests the necessity of precise query reduction and term selection techniques to improve the effectiveness for patent prior art search.
- We examined four simple query reduction methods to select useful words and prune out noisy words. We illustrated that these approaches are inefficient because they cannot discriminate between useful and noisy words. Since our system is over-sensitive to the existence of noisy words, we could not achieve high performance via these simple methods.
- We showed that a simple minimal interactive relevance feedback approach can perform as well as a highly engineered patent specific search system for CLEF-IP 2010.

1.4 Thesis Outline

The rest of the thesis is organised as follows: Chapter 2 reviews previous work from a number of related research areas, and Chapter 3 describes baseline and experimental Settings, test collections (i.e., queries and relevance judgments), etc. In Chapter 3, we also describe data curation and IPC filter errors. Chapter 4 contains the discussion and results related to our main analysis, where we figure out the main cause of low effectiveness of prior art search and then we propose and test the possible term selection methods. In Chapter 5, we finally conclude this thesis by summarising the results and discussing future work.

Chapter 2

Background and Related Work

In this chapter, first, we briefly explain the structure of patents, then we cover both general IR methods and patent-specific IR methods.

2.1 Structure of Patents

A patent is a structured document, which consists of title, abstract, description, citations, inventors, and several other sections. The text of a patent document is saved as an XML file with specific fields corresponding to each section or subsection in the patent document and some additional meta-data about the patent document itself (Figure 2.1). However, users usually get access to a text document — not an XML document [Magdy, 2012]. In this section, we briefly explain the main sections and meta-data of a patent document that are commonly used in a patent retrieval system as follows:

Title
The title of the patent appears in three languages — this is a feature in European Patent Office[1] (EPO) patents, where the title is stated in English, French, and German.

Abstract
Abstract is a short paragraph that contains a summary of the invention. This section is not always present in EPO patents since it is an optional section.

Description
This section of the patent document represents the core of the invention, since it contains all the technical details of the invention. It consists of a set of paragraphs that describe all the aspects of the invention in detail. The description section can contain tables, experimentation on the performance of the invention, and description of figures relating to the invention. The first paragraph of the description section usually contains information about the topical field of the invention. The references to other patent documents are very important information within the description

[1] http://www.epo.org/

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<patent-document ucid="UN-EP-1826951">
        <bibliographic-data>
                <technical-data>
                        <classifications-ipcr>
                                <classification-ipcr>H04L  12/28
                                    20060101AFI20070723BHEP
                                </classification-ipcr>
                                <classification-ipcr>H04L  12/28
                                    20060101CFI20070723BHEP
                                </classification-ipcr>
                        </classifications-ipcr>
                        <invention-title  lang="DE">Nahtlose Roaming-
                            optionen in einem Netz</invention-title>
                        <invention-title  lang="EN">Seamless roaming
                            options in a network</invention-title>
                        <invention-title  lang="FR">Options d&apos;
                            itinerance sans coupure dans un reseau</
                            invention-title>
                </technical-data>
        </bibliographic-data>
        <abstract lang="EN">
A communication protocol that provides load balancing and/or test
pattern information between devices is described. A first embodiment of
the protocol provides such information via a data frame that is
transmitted a definitive time after a special DTIM beacon is transmitted
. This protocol provides full compliance with IEEE 802.11. The second
embodiment of the protocol modifies the 802.11 beacon data structure
with additional information elements.
        </abstract>
        <description load-source="ep" status="new" lang="EN">
                <p num="1">
The present invention relates to the field of networking. In particular,
this invention relates to a protocol for providing load balancing and
test pattern signal evaluation information to wireless units in
accordance with Institute of Electrical and Electronics Engineers (IEEE)
802.11 constraints.
                </p>
                .
                .
                .
        </description>
        <claims load-source="ep" status="new" lang="EN">
                <claim num="1">
A method comprising:
modifying a beacon configured in accordance with a selected
communication protocol to produce a modified beacon, the modified beacon
comprising a plurality of additional information elements including at
least one of an access point name, an access point internet protocol
information and a load balancing information; andtransmitting the
modified beacon.
                </claim>
                .
                .
                .
        </claims>
</patent-document>
```

Figure 2.1: A sample patent XML file.

text. These references are part of the citations that a patent examiner would be interested to examine in order to measure the contribution of the invention against prior art.

Claims
The claims section of the patent document lists what aspects of the invention that the patent is going to protect. A successful patent does not have to have all its claims accepted, but at least one of them must be. The examination can lead to dropping some of the claims by showing that they are not novel. This usually happens because patent applicants try to generalize their invention as much as possible, which can lead to the novelty of some of the very general claims being found to be invalid. The claims section in EPO patents contains the list of claims in three languages (English, French, and German). However, this is not the situation for the initial patent application, where the claims are submitted in one language only, which is the language of the document. The claims translations are only provided for the granted patent.

Nonetheless, patents contain additional material, such as: tables, mathematical and chemical formulas, citations, technical drawing, meta-data, e.g., applicant, inventor, Intentional Patent Classification (IPC) codes, and publication date, that can be used to improve the retrieval effectiveness. IPC codes and citations has been widely applied in patent retrieval.

International Patent Classification (IPC) Code
In 1971, the Strasbourg Agreement established the International Patent Classification (IPC) under the World Intellectual Property Organization (WIPO), which divides technology into eight discrete Sections. The goal of this Agreement was to overcome the difficulties caused by using diverse national patent classification systems. [Harris et al., 2010]

A patent is assigned to one or more of the 71,000 IPC codes that indicate the related technical field or fields the patent covers. These codes are arranged in a hierarchical, tree-like structure with five distinct components. Figure 2.2 illustrates the components of an IPC classification.

The highest hierarchical level contains the eight sections of the IPC corresponding to very broad technical fields, labeled A through H. For example, Section C deals with"Chemistry and Metallurgy". Sections are subdivided into classes. The eighth edition of the IPC contains 120 classes. Class C07, for example, deals with "Organic Chemistry". Classes are further subdivided into more than 600 subclasses. Subclass C07C, for example, deals with "Acyclic or Carbocyclic Compounds". Subclasses are then further divided into main groups and subgroups. Main group symbols end with "/00". Ten percent of all IPC groups are main groups. For example, main group C07C 35/00 deals with "Compounds having at least one hydroxy or O-metal group bound to a carbon atom of a ring other than a six-membered aromatic ring". In some versions of the IPC, a series of numbers will follow the subgroup, reflecting the enactment date of the IPC version. '20060101' following the Subgroup indicates

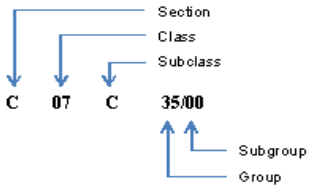


Figure 2.2: An example illustrating the main components of an International Patent Classification code.

a date of January 1, 2006, which is the date that the eighth version of the IPC took effect.

Patents are assigned at least one classification code indicating the subject to which the invention relates, this is called Main Code. They may also be assigned further classification and indexing terms to give further details of the contents of the invention, which are called Further Codes.

Citations
This section of the patent document contains the list of older patents that are related to the invention by describing the relevant parts of the prior-art of the invention, or these citations can be for patents that have been located by the patent examiners and were found to invalidate parts of the invention in the initially submitted patent application, where the final version of the patent get these parts modified or removed.

2.2 General Information Retrieval (IR)

An information retrieval (IR) system assists users in finding the information they need. Figure 2.3 illustrates the general IR process; First, a repository of indexed documents is created from a collection of documents to be searched for. Users formulate the information they need as a query and the IR system answers the query intent. In the matching process, the query and documents representations are compared and the result would be a ranked list of documents. The first attempt at formulation of a query with a particular information need in mind is often inaccurate and can result in an answer set that does not satisfy the user's information need. After reading some of the documents in the initial result set, the query can be reformulated in order to shift the result set toward the information need.

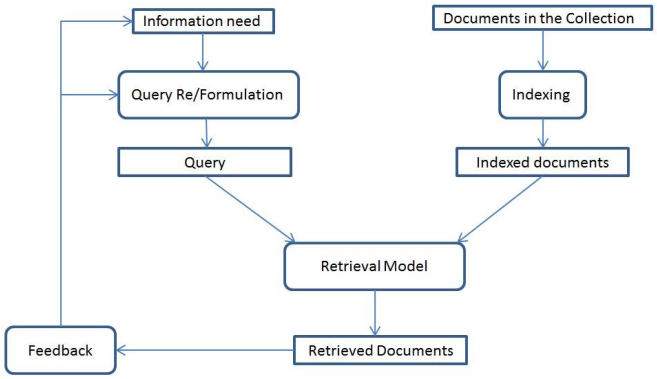


Figure 2.3: Simple illustration of the process in a general IR system.

2.2.1 Retrieval Models

Having constructed an index on a document collection, queries need to be matched to documents and a list of answers returned. We need a ranking algorithm based on good mathematical retrieval models to retrieve relevant documents at top of the ranking, consequently we will have high effectiveness. Three well-known retrieval models [Croft et al., 2010] are: (1) vector space models such as TF-IDF, (2) probabilistic models such as BM25, and Language Models.

Vector Space Model: TF-IDF
In vector space model, documents and queries are represented by a vector of term weights, and the collection is represented by a matrix of term weights:

$$D_i = (d_{i1}, d_{i2}, \ldots, d_{it}) \qquad Q = (q_1, q_2, \ldots, q_t)$$

$$\begin{array}{c} \\ Doc_1 \\ Doc_2 \\ \vdots \\ Doc_n \end{array} \begin{array}{c} \begin{array}{cccc} Term_1 & Term_2 & \ldots & Term_t \end{array} \\ \begin{bmatrix} d_{11} & d_{12} & \ldots & d_{1t} \\ d_{21} & d_{22} & \ldots & d_{2t} \\ \vdots & & & \\ d_{n1} & d_{n2} & \ldots & d_{nt} \end{bmatrix} \end{array}$$

where , D_i is each document in the collection C ($D_i \in C$), d_{ik} is the weight of term k in document D_i, and q_k represents a term in the query Q. The weighting function multiplies the occurrence of each query term in the document by the idf measure, after pivoted normalisation [Bache and Azzopardi, 2010]:

$$d_{ik} = \sum_{q \in Q \cap D} \frac{c(q_k, Di).idf_k(q)}{(1-b) + b.\frac{|D|}{avdl}} \tag{2.1}$$

where, $|D|$ is the size of the document and $avdl$ is the average document length, $tf_{ik}(q) = c(q_k, Di)$ is the number of occurrence of k-th query term in a document D_i,

and , and inverse document frequency measures importance in the collection:

$$idf_k(q) = \log \frac{N+1}{df(q_k)} \tag{2.2}$$

where, $df(q_k)$ is the number of documents in the collection which contain at least one occurrence of q, and N is the number of documents in the collection. This model scores a document higher if more query terms are present or these terms are rarer in the collection. The parameter b is set to 0.75 to be the same as the BM25 model below.

Probabilistic Models: BM25

BM25 is popular and effective ranking algorithm based on binary independence model. Equation (2.2) can be improved by factoring in the frequency of each term and document length - it is called Okapi weighting:

$$d_{ik} = \sum_{q \in Q \cap D} \log \frac{N+1}{df(q)} . \frac{(k_1+1)c(q,D)}{k_1((1-b)+b.\frac{|D|}{avdl})+c(q,D)} \tag{2.3}$$

The variable k_1 is a positive tuning parameter that calibrates the document term frequency scaling. A $k_1 = 0$ corresponds to a binary model (no term frequency), and a large value corresponds to using raw term frequency. b is another tuning parameter ($0 \leq b \leq 1$) which determines the scaling by document length: $b = 1$ corresponds to fully scaling the term weight by the document length, while $b = 0$ corresponds to no length normalization.

If the query is long, then we might also use similar weighting for query terms. This is appropriate if the queries are paragraph long information needs, but unnecessary for short queries.

$$d_{ik} = \sum_{q \in Q \cap D} \log \frac{N+1}{df(q)} . \frac{(k_1+1)c(q,D)}{k_1((1-b)+b.\frac{|D|}{avdl})+c(q,D)} . \frac{(k_3+1)c(q,Q)}{k_3+c(q,Q)} \tag{2.4}$$

with $c(q,Q)$ being the frequency of term q in the query Q, and k_3 being another positive tuning parameter that this time calibrates term frequency scaling of the query. In the equation presented, there is no length normalization of queries because retrieval is being done with respect to a single fixed query. The tuning parameters of these formulas should ideally be set to optimize performance on a development test collection. That is, we can search for values of these parameters that maximize performance on a separate development test collection (either manually or with optimization methods such as grid search or something more advanced), and then use these parameters on the actual test collection. In the absence of such optimization, experiments have shown reasonable values are to set k_1 and k_2 to a value between 1.2 and 2, and b = 0.75 [Manning et al., 2008].

Language Models with Terms Smoothing
The basic idea behind the *Language Modelling* approach is to estimate a language model for each document, and rank documents by the likelihood of the query according to the estimated language model. Here terms are assumed to occur independently, and the probability is the product of the individual query terms given the document model M_D of document D:

$$P(Q|M_D) = \prod_{q \in Q} P(q|M_D) \tag{2.5}$$

$$P(q|M_D) = \frac{c(q, D)}{|D|} \tag{2.6}$$

The overall similarity score for the query and the document could be zero if some of query terms do not occur in the document. However, it is not sensible to rule out a document just because a single query term is missing. For dealing with this, language models make use of smoothing to balance the probability mass between occurrences of terms in documents, and terms not found in the documents.

Jelinek-Mercer smoothing. Jelinek-Mercer smoothing language model [Zhai and Lafferty, 2004] combines the relative frequency of a query term $q \in Q$ in the document D with the relative frequency of the term in the collection (C) as a whole. With this approach, the maximum likelihood estimate is moved uniformly toward the collection model probability $P(q|C)$:

$$P(q|M_D) = (1 - \lambda)\frac{c(q, D)}{|D|} + \lambda P(q|C) \tag{2.7}$$

$c(q, D)$ represents the frequency of term q in document D. The optimal value of λ depends on both the collection and the query. It is normally suggested as ($\lambda = 0.1$) for title queries and ($\lambda = 0.7$) for long queries.

Dirichlet (Bayesian) smoothing (DirS). As long documents allow us to estimate the language model more accurately, Dirichlet smoothing [Zhai and Lafferty, 2004] smooths them less. If we use the multinomial distribution to represent a language model, the conjugate prior of this distribution is the Dirichlet distribution. This gives:

$$P(q|M_D) = \frac{c(q, D) + \mu P(q|C)}{|D| + \mu} \tag{2.8}$$

The formula assign negative score to documents that contain the term, but with fewer occurrence than predicted by the collection language model. As μ gets smaller, the contribution from the collection model also becomes smaller, and more emphasis is given to the relative term weighting. Precision is more sensitive to μ for long queries, especially when μ is small. When μ is sufficiently large, long queries perform better than short queries. The optimal value of μ varies from collection to

collection, though in most cases, it is around 2000. The performance is more sensitive to smoothing for verbose queries. Long queries also require more aggressive smoothing to achieve optimal performance.

2.2.2 The Study of Retrievablity

Retrievability measures indicate how easily a document could be retrieved using a given IR system, while findability measures indicate how easily a document can be found by a user with the IR system [Azzopardi and Vinay, 2008]. Some documents are retrieved by many queries while others may never show up within the top-n ranked results via any query terms that they are relevant for [Lupu et al., 2013a]. When a document is difficult or impossible to retrieve in a particular retrieval model, it is difficult or impossible to retrieve when relevant and this leads to a low recall.

Essentially, it is desirable that the retrieval system consider all documents with similar retrievability (Gini-Coefficient is used to measure the retrievability) because documents become less retrievable when others become more retrievable. However, two aspects can affect findability: the inherent bias favouring some types of documents over others introduced by the retrieval model, and the failure to correctly capture and interpret the context [Bashir and Rauber, 2009b, 2011]. There are certain features that increase access to the corpus by making the retrievability of documents more equal [Bache and Azzopardi, 2010]:

1. Sensitivity to term frequency: A higher frequency of a given query term makes the document more relevant.

2. Length normalization: Incorporation term frequency into a model make it biased to score longer documents higher than shorter documents, so there is a tendency to over-score longer documents. Shorter documents are not penalised when length normalisation is used.

3. Convexity: An IR model will have convexity if it ranks document d_3, which has both query words w_1 and w_2, higher than documents d_1 and d_2, which just have one of the query words twice.

Bias of retrieval systems is the characteristic of a system to give preference to certain features of documents, when it ranks results of any given query. For example, *PageRank* favours popular documents by evaluating the number of in-links of web pages in addition to pure content features while *TFIDF* and *OKAPI-BM25* favour large terms frequencies [Bashir and Rauber, 2011].

Retrievablity Measurement

Retrievability measures how likely each document d inside a collection D can be retrieved within the top c ranked results for all queries in Q. $r(d)$ defines as follows:

$$r(d) = \sum_{q \in Q} f(k_{dq}, c)$$

where k_{dq} is the rank of d in the result set of query $q \in Q$, c denotes the maximum rank that a user is willing to proceed down the ranked list. The function $f(k_{dq}, c)$ returns a value of 1 if $k_{dq} \leq c$, and 0 otherwise. *Retrievability* inequality can be analysed using the *Lorenz* Curve. Documents are sorted according to their retrievability score in ascending order, plotting a cumulative score distribution. If the retrievability of documents is distributed equally, then the Lorenz Curve will be linear. The more skewed the curve, the greater the amount of inequality or bias within the retrieval system. The Gini coefficient G is used to summarize the amount of bias in the Lorenz Curve, and is computed as follows:

$$G = \frac{\sum_{i=1}^{n}(2.i - n - 1).r(d_i)}{(n-1)\sum_{j=1}^{n} r(d_j)} \tag{2.9}$$

where $n = |D|$ is the number of documents in the collection sorted by $r(d)$. If $G = 0$, then no bias is present because all documents are equally retrievable. If $G = 1$, then only one document is retrievable and all other documents have $r(d) = 0$. By comparing the Gini-Coefficients, we can analyse the retrieval bias imposed by the underlying retrieval functions on a given document collection.

2.2.3 Query Expansion (QE)

One solution to the significant term mismatch between the query and the relevant documents is query expansion, which has been effective in many retrieval tasks. The idea of QE is to add more terms to the original user's query to increase the probability of matching of the query terms with relevant documents, with the objective of improving retrieval effectiveness. The expansion terms can be selected from a feedback process [Cao et al., 2008], or from external sources such as Wikipedia, or dictionaries. Original queries should be expanded by good terms, unless it can lead to retrieval of non-relevant documents.

Feedback-based QE
An initial query can be expanded using a feedback from users-*relevance feedback*-or automatically from top k ranked retrieved documents, assuming they are relevant to the query-*pseudo relevance feedback*. Getting feedback from users needs user studies and interaction that is far from our research, so we just explain pseudo relevance feedback in this section.

Pseudo Relevance Feedback (PRF)
PRF is a standard technique to enrich the initial query with additional terms from the top ranked documents from an initial retrieval run under the assumption that these documents are relevant.

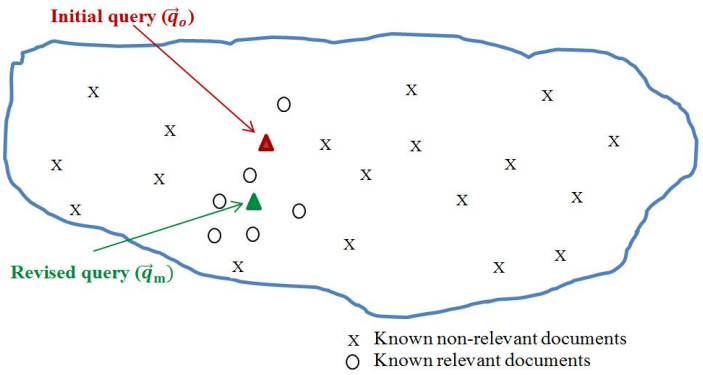


Figure 2.4: Rocchio algorithm for relevance feedback. Some documents have been labelled as relevant and non-relevant and the initial query vector is moved in response to this feedback [Manning et al., 2008].

The *Rocchio* algorithm is used to modify the query by the partial knowledge of known relevant and non-relevant documents; the goal is to make the query closer to the centroid of the relevant documents but further from irrelevant documents (figure 2.4). The modified query, $\vec{q}_m$ is:

$$\vec{q}_m = \alpha\vec{q}_0 + \beta\frac{1}{|D_r|}\sum_{\vec{d}_j \in D_r}\vec{d}_j - \gamma\frac{1}{|D_{nr}|}\sum_{\vec{d}_j \in D_{nr}}\vec{d}_j \tag{2.10}$$

where q_0 is the original query vector, D_r and D_{nr} are the set of known relevant and non-relevant documents respectively, and α, β, and γ are weights attached to each term. These control the balance between trusting the judged document set versus the query: if we have a lot of judged documents, we would like a higher β and γ [Manning et al., 2008].

Predicting the effectiveness of feedback documents or terms
Distinguishing between good expansion terms and bad ones only based on their distribution (extract the most frequent terms) in the feedback documents and in the whole collection (extract the most specific terms in the feedback documents) is not sufficient. Supervised learning methods for term selection-it is considered as a term classification problem to separate good expansion terms from others directly according to their potential impact on the retrieval effectiveness-helps in this regard. Any classifier can be used to classify terms or feedback documents such as: Suport Vector Machines(SVM) [Cao et al., 2008], Naïve Bayes and Logistic Regression [He and Ounis, 2009].

QE by External Resources
The most common form of query expansion is global analysis, using some form of thesaurus such as dictionaries, WordNet, Wikipedia, and etc. For each term t in a query, the query can be automatically expanded with synonyms and related

words of t from the thesaurus. Use of a thesaurus can be combined with ideas of term weighting: for instance, one might weight added terms less than original query terms [Manning et al., 2008].

2.2.4 Query Reduction (QR)

In general, retrieval effectiveness for long queries is often lower than retrieval effectiveness for shorter keyword queries because the additional information provided in verbose queries is more likely to confuse current search engines rather than help them. Query reduction, a technique for dropping unnecessary query terms from long queries, improves performance of retrieval. Two main approaches have proposed in previous works for long queries: selecting a subset of the verbose query (or sub-query) and weighting query words in the verbose query.

- **Selecting of Subsets.** A search engine do not retrieve related documents at top of the list for some long queries, but the same retrieval system perform more precisely when just the key concepts are used as a query. So, the identification of the key query concepts will have a significant positive impact on the retrieval performance for verbose queries. Extracting the key query concepts can be done by learning to identify key concepts in long queries using a variety of features [Bendersky and Croft, 2008]. The other approach, to choose effective subsets in a query, involves analysing all the subsets of terms from the original query (sub-queries), and identifying the most promising sub-query to replace the original long query. For ranking sub-queries, an algorithm based on the Support Vector Machines (SVM) classification can be used [Kumaran and Carvalho, 2009]. The quality of query reduction depends on the performance of the predictor and ranking algorithm [Balasubramanian et al., 2010].

- **Weighting Query Words.** Query term ranking approaches are used to select effective terms from a verbose query by ranking terms. A vast number of rankings are possible given different settings of individual term weights, for example, it is possible to train a regression model to weight all query words of a verbose query [Lease et al., 2009]. It is also possible to assign weights to concepts by learning the importance of concepts underlying the verbose query [Bendersky et al., 2010].

2.2.5 IR Evaluation Metrics

A retrieval system is evaluated considering a set of relevance judgements, a binary assessment of either *relevant* or *irrelevant* for each query-document pair. An ideal retrieval system can retrieve all relevant documents. Table 2.1 is the contingency table, where:

True Positive (tp): documents which are relevant and the system retrieves them.
False Negative (fn): documents which are relevant but the system does not retrieve

them.
False Positive (fp): documents which are non-relevant but the system retrieves them.
True Negative (tn): documents which are non-relevant and the system does not retrieve them.

	Relevant	Non-relevant
Retrieved	true positive (tp)	fale positive (fp)
Not-retrieved	false negative (fn)	true negative (tn)

Table 2.1: Contingency table.

Precision and Recall
Precision and recall are the most frequent and basic measures for information retrieval effectiveness. They are calculated with respect to what IR system returns as a set of documents for a query.
Precision (P) is the fraction of retrieved documents that are relevant:

$$Precision(P) = \frac{\sharp(relevant\ items\ retrieved)}{\sharp(retrieved\ items)} = \frac{tp}{tp+fp} = P(relevant|retrieved)$$

Recall (R) is the fraction of relevant documents that are retrieved:

$$Recall(R) = \frac{\sharp(relevant\ items\ retrieved)}{\sharp(relevant\ items)} = \frac{tp}{tp+fn} = P(retrieved|relevant)$$

For many prominent applications, particularly web search, good results on the first page or the first three pages are important than all relevant documents. So, they wish to look at precisions and recalls over a series of different rank cut-offs rather than to look at the entire retrieved set. This is referred to as "Precision/Recall at k", for example "Precision/Recall at 10".

$$precision@k = \frac{\sharp(documents\ retrieved\ and\ relevant\ up\ to\ rank\ k)}{k} \tag{2.11}$$

$$recall@k = \frac{\sharp(documents\ retrieved\ and\ relevant\ up\ to\ rank\ k)}{\sharp(documents\ relevant)} \tag{2.12}$$

Average Precision and Mean Average Precision (MAP)
We can measure MAP by calculating Average Precision on retrieval results. Average Precision is the average of precision at each point where a relevant document is found is computed as:

$$Avg.\ precision = \frac{\sum_{r=1}^{N}(P(r) \times rel(r))}{n} \tag{2.13}$$

where r is the rank, N the number of documents retrieved, $rel(r)$ a binary function of the document relevance at a given rank, $P(r)$ is precision at a given cut-off rank r, and n is the total number of relevant documents.

Then, for a given set of queries, Q, MAP can be calculated by:

$$MAP(Q) = \frac{\sum_{q \in Q} Avg.\ precision(q)}{|Q|} \tag{2.14}$$

where q is a query in Q.

2.3 Patent-specific IR

2.3.1 The Study of Retrievablity for patents

Rerievability is specifically critical in recall oriented application, such as patent retrieval, or legal settings. In these cases, the focus of a system is not so much on providing the best document to answer a specific information need (as e.g. in Web search settings), but to retrieve all documents that are relevant. Thus, all documents should at least potentially be retrievable via correct query terms. Designing retrieval systems for recall oriented tasks has been emphasized in recent years [Fujii et al., 2007; Kontostathis and Kulp, 2008], but before designing a new or using an existing retrieval system for recall oriented applications one needs to analyse the effects of the retrieval system bias as well as the overall retrievability of all documents in the collection using the retrieval function at hand.

Analysing retrievability of documents specifically with respect to relevant and irrelevant queries to identify whether highly retrievable documents are really highly retrievable, or whether they are simply more accessible from many irrelevant queries rather than from relevant queries, revealed that 90% of patent documents which are highly retrievable across all types of queries, are not highly retrievable on their relevant query sets [Bashir and Rauber, 2009a].

Experiments with different collections of patent documents suggest that query expansion with pseudo-relevance feedback can be used as an effective approach for increasing the findability of individual documents and decreasing the retrieval bias. Pseudo-relevance feedback documents were identified using cluster-based [Bashir and Rauber, 2009b] or terms-proximity-based methods [Bashir and Rauber, 2010].

Another study analysed the relationship between retrievability and effectiveness-based measures (Precision, Mean Average Precision) [Bache and Azzopardi, 2010]. Results showed that the two goals of maximizing access and maximizing performance are quite compatible. They further concluded that reasonably good retrieval performance can be obtained by selecting parameters that maximize retrievability (i.e., when there is the least inequality between documents according to Gini-Coefficient given the retrievability values). Their results support the hypothesis that retrieval functions can be effectively tuned using retrievability-based measure without recourse to relevance judgments, making it an attractive alternative for automatic evaluation.

2.3.2 Query Formulation

The patent prior-art search scenario begins with the full patent application as a query. A full text as a query is a challenge compared to a classical IR, since it is not focused on the information that the user needs. In order to achieve good retrieval results, it is important to extract the best representative text with the proper weights. Therefore, query generation based on query document is essential to reduce the difficulty of formulating effective queries by users.

Terms Selection

Selection based on terms which are frequent in query but rare in the collection
Identifying useful query terms and giving them higher weight is important to build an effective query. The simplest proposed approach was weighting terms in the query based on their perceived significance in the target corpus, combined with their significance in the query [Itoh et al., 2003]. The problem with their method was that they did not take into account the fact that some terms, while being important to the definition of the request for information, may not necessarily appear in the target set at all. For query term selection purposes, it would seem more useful to weight them based only on the genre to which the query belongs, rather than the genre of the target collection. The enhanced version of selecting the most discriminative terms for each topic patent is to compute Kullback-Leibler divergence (KLD) between the language model of the query and the whole collection:

$$KLD(P_Q(t)||P_C(t)) = P_Q(t).log(\frac{P_Q(t)}{P_C(t)}) \tag{2.15}$$

Where, P_Q is the probability of each term t within the patent topic q, and P_C is the probability of the same term t within the whole collection. By applying the equation (2.15), it is possible to rank all the terms from the patent topic according to their importance within the query. After ranking the terms by their divergence, only terms with divergence above an specific threshold are selected. Thus, we can build queries that contain the most discriminative terms in different fields of query, which appear frequently in the query, but not so frequently in the collection. So, it helps to retrieve the most relevant patents to a given topic [Pérez-Iglesias et al., 2010]. It is possible to exploit the knowledge of IPC meta-data into the query model [Mahdabi et al., 2011b]:

$$P_Q(t) = \lambda \frac{TF(t,Q)}{|Q|} + \frac{(1-\lambda)}{N} \sum_{d \in IPC_Q} \frac{TF(t,D)}{|D|} \tag{2.16}$$

Where, $TF(t,Q)$ is the term frequency of the term t in the query patent document, $|Q|$ is the length of the query patent, N is the size of the relevant cluster with the same IPC code as the query, and λ is the smoothing parameter.

Which fields in patent application are more effective to extract query terms

A special characteristic of patent documents is their structural information. They mainly have different fields such as title, abstract, description, and claims. Different fields use different type of language for describing the invention. Abstract and description use more technical terminology while claim field usually uses a legal jargon. Structured indexing keeps the field structure in the index, which allows searching specific fields instead of searching in full document. Separate fields for meta-data (section 2.3.5) like IPC code and author can help to retrieval effectiveness [Magdy et al., 2010].

Early patent search tasks mainly considered claims to build the query, the same as what examiners start the novelty process [Konishi, 2005; Takaki et al., 2004; Mase et al., 2005; Fujii, 2007a], whereas recent works have showed that building queries from description field is more useful in patent retrieval (considering background summary in US patents equivalent to description field in European patents.) [Xue and Croft, 2009b,a; Mahdabi et al., 2011b]. Another research showed that extracting terms according to $log(tf)idf$ scores from every field of the query patent, and giving higher importance to terms extracted from the abstract, claims, and description fields than to terms extracted from the title field, is an effective way of constructing a search query [Cetintas and Si, 2012]. The other experiment showed that discarding Description from query improves 'MAP' up to 30% because the description contains more noise than information [Gobeill et al., 2010]. They also showed that claims are more informative and title is poorly informative in retrieval.

Using Phrases instead of Terms

Most of query formulation techniques rely on terms, but encouraging results have been obtained using phrases recently [Becks et al., 2010]. Early results demonstrated that an NLP-based grouping of terms can increase the performance compared to the bag-of-words approach, though the increase is smaller than in a non-patent collection [Osborn et al., 1997]. Another task could improve retrieval effectiveness by adding syntactic phrases in the form of dependency triples, to a bag-of-words representation [D'hondt et al., 2011]. Key Phrase Extraction (KPE) algorithms is another way to form a query based on phrases. A list of phrases, generated by a KPE algorithm, can succinctly represent a complex and lengthy patent. [Verma and Varma, 2011a].

Diverse Query Generation

In this approach, the focus is on generating diverse queries that can improve overall retrieval effectiveness in sessions rather than generating a single best query that can retrieve more relevant documents from a single retrieval result (i.e., more relevant documents in aggregated retrieval results obtained by multiple queries in a session). Diverse query generation is important because query documents typically contain several different aspects (or topics) and different types of relevant documents may be related to these aspects. To identify aspects, 500 top terms based on their tf-idf rank, are clustered into n sets with respect to their similarity. Each distinct sets of

terms represents one query aspect, then top k retrieved documents for each sub-query consider as pseudo-relevant documents (PRD) and those ranked below the top k are non-relevant documents (NRD). Then the query is generated by decision tree. [KIM, 2014; Kim and Croft, 2014].

2.3.3 Query Expansion for Patents

In patent domain, query is very long and there is a significant mismatch between queries and relevant documents [Roda et al., 2010; Magdy et al., 2010]. Most of 'QE' techniques often did not demonstrate any significant improvement in effectiveness for patent search [Kishida, 2003; Konishi, 2005]. Therefore, for an effective query expansion in patent domain, as it will be discussed in this section, specific techniques have been exploited.

Query Expansion by Pseudo Relevance Feedback (PRF)
PRF, in patent domain, is not as effective as in other applications because of the poor effectiveness in initial retrieval. So, the assumption that top k documents are relevant is wrong and we might add noise in the query, so, the improvement is insignificant. The solutions proposed to cope with this problem are as follows:

- **Selecting documents for PRF based on cluster analysis:** a document that can cluster lots of high similar documents considers relevant and a document that has no nearest neighbour or some neighbours with low similarity is irrelevant [Lee et al., 2008]. In patent domain, where there is a large vocabulary diversity for expressing an invention, the idea can be improved by intra-cluster similarity rather than only on the basis of their size [Bashir and Rauber, 2009b].

- **Selecting patents for PRF based on their similarity with query patent via specific terms:** In this approach patents for PRF are identified based on their similarity with query patents over a subset of terms, rather than the overall document similarity. The succession of this approach highly depends on selecting appropriate terms from query patent, which produce the best PRF candidates that can help in improving retrievability during 'QE' [Bashir and Rauber, 2010]. This set of experiments showed significant improvement for Gini coefficient, which is used to measure retrievability, but there is no report on patent retrieval effectiveness measures.

- **Identyfying expansion terms:**
  Term proximity information can be used to identify expansion terms. Given a query patent, first an initial query is generated by taking ,for example, claim terms, then a query-specific lexicon that includes the terms from the same IPC patents is built. Among many terms in the lexicon, only expansion terms identified by two adjacency operators used in patent examination (i.e., "ADJn" and "NEARn") [Mahdabi et al., 2013].

- **Predicting the effectiveness of feedback documents:**
  Regression can be used to predict the effectiveness of a feedback document. Different features can also be used to capture the effectiveness of a feedback document in terms of its performance in query expansion [Mahdabi and Crestani, 2012].

Random indexing to identify terms to use for query expansion [Sahlgren et al., 2002], and expansion using noun phrases [Mahdabi et al., 2012] are the other works to improve the effectiveness of standard query expansion for prior-art search.

Query Expansion by External Resources
Some external resources like WordNet [Miller et al., 1990], which were reported to improve retrieval effectiveness in several IR research investigations, showed insignificant change to overall retrieval effectiveness, but a degree of improvement for some topics in patent domain [Magdy and Jones, 2011]. They also applied the idea of automatically generating the synonyms set (SynSet) using parallel manual translations to create possible synonyms sets (In CLEF-IP patent collection, some of the sections in some patents are translated into three languages: English, French, and German). Although this idea presented better results than WordNet, there was no considerable improvement in retrieval effectiveness. The only QE task that achieved the best results, used a combination of PRF and QE with translation of terms and phrases from German and French [Jochim et al., 2011].

2.3.4 Query Reduction for Patents

Query reduction is a solution for problems with using a full verbose patent as a query: it is not focused on information needed by the user, and a verbose query may cover more than one topic.

- **Query Segmentation :** Decomposing each patent query into coherent subtopics segments-using TextTiling [Hearst, 1997]-is a solution to make long ambiguous queries focused on the information need. Sub-topic segments can be used as separate queries (query stream) for initial retrieval, then the retrieval results from each of the individual streams are merged to construct the final ranked list for the whole original query. Using each sub-topic as a query stream enables a retrieval model to retrieve related documents from the collection in a more precise way and also allow the PRF algorithm to work on a more focused set of pseudo-relevant documents [Takaki et al., 2004; Ganguly et al., 2011a]. Another work adapted pseudo relevance feedback for query reduction by decomposing a patent application into constituent text segments and computing the Language Modelling (LM) similarities by calculating the probability of generating each segment from the top ranked documents. The least similar segments from the query removed from the query, hypothesizing that removal of segments most dissimilar to the pseudo-relevant documents can increase the

precision of retrieval by removing non-useful context, while still retaining the useful context to achieve high recall as well [Ganguly et al., 2011b].

- **Patent Summarization :** This approach assumes that the patent summary (using TextTiling) reflects the main topic as well as the subtopics of a patent document in a concise manner. Then, language model for the query, collection, and each summary are generated [Mahdabi et al., 2011a].

2.3.5 The Use of Metadata

The main textual content of patent documents is known to be difficult to process with traditional text processing and text retrieval techniques, but patents contain additional material, such as: tables, mathematical and chemical formulas, citations, technical drawing, meta-data, e.g., applicant, inventor, Intentional Patent Classification (IPC) codes, and publication date, that can be used to improve the retrieval. In this section, the use of patent meta-data to improve the retrieval has been explained.

The Use of Citation

The most successful use of metadata to date is the citation lists in order to learn patterns of relevance [Lupu et al., 2013a]. The patent collection is a very dense network of citations creating a set of interrelations particularly interesting to exploit during a prior art search. The large majority of patents are continuations of previous works and patents. The citation relations make this development process visible. Similarly, fundamental patents which open new technologies sub-fields are exceptional but tends to be cited very frequently in the whole sub-field during years. Citation graph of a patent collection is used for identifying patent thickets, i.e. the patent portfolios of several companies overlapping on a similar technical aspect. Related patents can be inferred from the overall citation network of a patent collection. If a new patent applicant belonging to this patent ticket appears, it is very likely that the most relevant prior art documents are already present in this patent thicket [Lopez and Romary, 2009].

The patents cited in the description of the topic patent are used as relevant documents, because citations are usually prior arts for a citing patent. Only citations which are in the collection can be helpful in the retrieval process. The idea of *PageRank*-identifying authoritative pages by analysing hyperlink structure on World Wide Web_ can be used for citations. A patent, which is cited by a large number of other patents, is more important. Text-based and citation-based scores combined to compute the ranking score for documents [Fujii, 2007a,b].

Citation texts for patents are a whole paragraph. Therefore, for each patent document presented and cited in the collection, the entire paragraph of citation can be appended to the textual material of the cited patent. A boolean feature uses to indicate weather a cited patent in query patent has retrieved, then this document can get a higher weight at any future post-ranking process. Due to the limited number of citation texts, this approach showed just a trivial improvement [Lopez and Romary,

2009]. However, citation information does not always presented in the patent application and this method can not be used in real life patent search and initial citations by the applicants may not consider relevant by patent examiners [Magdy and Jones, 2010a; Magdy et al., 2011].

Similar tasks also indicated improvement in 'MAP' and 'Recal' using citations in patent retrieval [Gobeill et al., 2010; Gurulingappa et al., 2010].

The Use of IPC Codes

Patents are classified by the patent offices into large hierarchical classification schemes based on their area of technology. The use of patent classification has two major benefits. The first is that the classifications provide access to concepts rather than words, such that even if the same word or phrase is commonly used in two technology areas, patent classifications will provide the context of its use. In effect, they allow the search space of patents to be reduced, by allowing the user to exclude from the search process patents in classes not related to the search topic at hand [Lopez and Romary, 2010]. The second major benefit is the language independence provided by classifications, as classification symbols can be mapped to multiple languages [D'hondt and Verberne, 2010]. This allows patent searchers to conduct reasonably effective retrieval even in languages that they do not understand. All previous works, considered IPC code in their search, reported improvement in retrieval effectiveness [Harris et al., 2010, 2011, 2009; Fujita, 2005; Graf et al., 2010; Herbert et al., 2010; Kang et al., 2007; Verma and Varma, 2011a]. It has also reported that using complete IPC code leads in better results than just 4-digit code [Gobeill et al., 2010].

The Use of Images

For the purposes of the search for innovation, we are interested in all forms of information. Some technology areas rely information present in images (flowcharts and diagrams), so, beyond text data, image processing tasks also can contribute to the search. Graph-based measure has a higher discriminative power, but higher computational costs than the text-based measures [Lupu et al., 2013b].

2.3.6 Multilinguality

The interest in multilingual patent search arises from their international and multilingual nature (the European patent office-EPO-makes patent text available in three languages: English, French, and German). Patents on the same topic may be published in different countries in different languages, and it is important for patent examiners to be able to locate relevant existing patents whatever language they are published in. Therefore an important topic in patent retrieval is Cross-Language Information Retrieval (CLIR), where the topic is a patent application in one language and the objective is to find relevant prior-art patents in another language [Lupu et al., 2013a; Joho et al., 2010; Roda et al., 2010; Piroi et al., 2012]. In recent years machine translation (MT) has become established as the dominant technique for translation in

CLIR, which usually achieve better CLIR effectiveness than dictionary-based translation (DBT) methods. However, translation using MT is time consuming and resource intensive for cross language patent retrieval (CLPR), where the query text can often take the form of a full patent application running to tens of pages. Applying IR text pre-processing like stop word removal and stemming to the MT training corpus prior to the training phase can lead to a significant decrease in the MT computational [Magdy and Jones, 2013].

2.3.7 Multi-stage Retrieval

It is common to use patent meta-data and non-textual features as pre and post processing steps of text-based retrieval techniques [Lopez and Romary, 2009]. Many patent retrieval tasks re-rank the top retrieved documents from initial retrieval stage based on additional patent feature [Lopez et al., 2010], claim structure [Mase et al., 2005], and considering IPC information of patent and its neighbours to retrieve similar patents [Verma and Varma, 2011b].

2.3.8 Evaluation Metrics for Patent Retrieval

The simplest solution to measure the performance in a recall focused IR task is to evaluate the recall, however, it fails to reflect how early a system retrieves the relevant documents. Although recall is the objective for such applications, the score should be able to distinguish between systems that retrieve relevant documents earlier than those that retrieve them later. For recall-oriented IR applications, the problem is viewed as a ranking problem with a cut-off for a maximum number of documents to be checked N_{max}.

Patent Retrieval Evaluation Score (PRES)
PRES is a novel metric for evaluating recall-oriented IR applications, which derived from the normalized recall measure (R_{norm}). It measures the ability of a system to retrieve all known relevant documents earlier in the ranked list. Unlike MAP and Recall, PRES is dependent on the relative effort exerted by users to find relevant documents. This is mapped by N_{max} (Equation 2.17), which is an adjustable parameter that can be set by users and indicates the maximum number of documents they are willing to check in the ranked list. PRES measures the effectiveness of ranking documents relative to the best and worst ranking cases, where the best ranking case is retrieving all relevant documents at the top of the list, and the worst is to retrieve all the relevant documents just after the maximum number of documents to be checked by the user (N_{max}). The idea behind this assumption is that getting any relevant document after N_{max} leads to it being missed by the user, and getting all relevant documents after max leads to zero Recall, which is the theoretical worst case scenario. PRES is the area between the actual and worst cases (A_2) divided by the area between the best and worst cases ($A_1 + A_2$). N_{max} introduces a new definition to the quality of ranking of relevant results, as the ranks of results are relative to the value

of N_{max}. For example, getting a relevant document at rank 10 will be very good when $N_{max} = 1000$, good when $N_{max} = 100$, but bad when $N_{max} = 15$, and very bad when $N_{max} = 10$. Systems with higher Recall can achieve a lower PRES value when compared to systems with lower Recall but better average ranking. The PRES value varies from R to $\frac{nR^2}{N_{max}}$, where R is the Recall, according to the average quality of ranking of relevant documents.

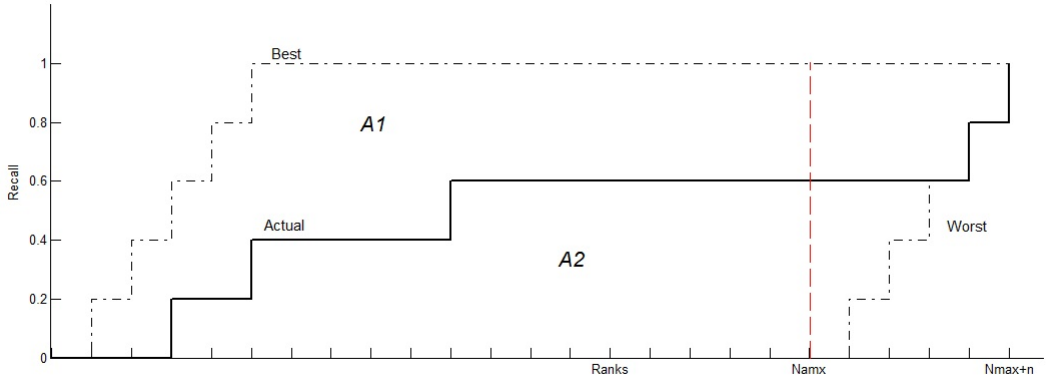


Figure 2.5: PRES curve is bounded between the best case and the new defined worst case [Magdy and Jones, 2010b].

$$PRES = \frac{A_2}{A_1 + A_2} = 1 - \frac{\frac{\sum r_i}{n} - \frac{n+1}{2}}{N_{max}} \tag{2.17}$$

where r_i is the rank at which the ith relevant document is retrieved, N_{max} is the maximum number of retrieved documents to be checked by the user, i.e. the cut-off number of retrieved documents, and n is the total number of relevant documents [Magdy and Jones, 2010b].

Chapter 3

Baseline IR Framework

In this chapter, we first briefly explain the baseline system and the experimental settings, and we describe the data collection we used — CLEF-IP. Then we cover two main errors caused by the data curation and experimental settings.

3.1 Baseline and Experimental Settings

We developed a baseline IR system for patent prior art search on the top of the Lucene search engine[1], which processes queries using both BM25 [Robertson et al., 1993] and LM (Dirichlet smoothing, and Jelinek-Mercer smoothing) [Zhai and Lafferty, 2004] scoring functions. We used Lucene to index the English subset of CLEF-IP 2010 dataset[2] (We will describe CLEF-IP in Section 3.2) that contains 2.6 million patent documents and 1,303 topics (queries) for the English test set. We used the default Lucene settings with the Porter stemming algorithm Porter [1980] and English stop-word removal. We also removed patent-specific stop-words as described in Magdy [2012]. In our implementation, each section of a patent (title, abstract, claims, and description) is indexed in a separate field. However, when a query is processed, all indexed fields are targeted with an equal weight, since this generally offers best retrieval performance. We also used the International Patent Classification (IPC) codes assigned to the topics to filter the search results by constraining them to have common IPC codes with the patent topic as suggested in previous works [Lopez and Romary, 2010]. Although this IPC code filter may prevent retrieval of relevant patents — as it will be explained in Section 3.3.2 — we keep it for the following reasons: (i) more than 80% of the patent queries share an IPC code with their associated relevant patents, and (ii) it makes the retrieval process much faster. The accuracy of the results is evaluated using three popular metrics — Mean Average Precision (MAP), Average Recall, and Patent Retrieval Evaluation Score [Magdy, 2012] (PRES) — on the top-100 results for each query, assuming that patent examiners are willing to assess the top 100 patents [Joho et al., 2010].

We achieved the best performance while querying with the description section as in previous work [Xue and Croft, 2009b] and using either the LM or the BM25

[1] http://lucene.apache.org/

[2] http://www.ifs.tuwien.ac.at/~clef-ip/

Table 3.1: Comparing performance metrics for different IR models and query formulation.

IR model	Metric	Patent section					
		Title	Abstract	Description	DescP5	Claims	Claims1
BM25	PRES	0.3700	0.488	0.539	0.476	0.504	0.474
	MAP	0.0567	0.101	0.131	0.097	0.109	0.094
	A. Recall[3]	0.4848	0.594	0.6342	0.585	0.610	0.582
TF-IDF	PRES	0.364	0.481	0.521	0.483	0.520	0.482
	MAP	0.056	0.097	0.121	0.098	0.115	0.097
	A. Recall	0.478	0.590	0.621	0.591	0.628	0.590
LMDIR	PRES	0.361	0.498	0.547	0.478	0.500	0.472
	MAP	0.049	0.100	0.133	0.095	0.101	0.090
	A. Recall	0.475	0.611	0.638	0.588	0.610	0.580
LMJ	PRES	0.060	0.040	0.038	0.040	0.039	0.040
	MAP	0.002	0.001	0.001	0.001	0.001	0.001
	A. Recall	0.110	0.079	0.075	0.078	0.075	0.078

scoring function. We call this initial query the Patent Query and use it as our main baseline.

Table 3.1 compares the system performance for different IR models (BM25, TF-IDF, LM with Dirichlet, and Jelinek-Mercer smoothing — Section 2.2.1) with Lucene default settings and different sections of the patent query. It can be seen that the results for BM25 and LM with Dirichlet (LMDIR) are very similar, therefore in the this thesis, we report our results based on LM.

In addition, we compare our results to *PATATRAS*, a highly engineered system developed by Lopez et al. [2010], which achieved the best performance in the CLEF-IP 2010 competition. This system uses multiple retrieval models (especially Kullback-Leibler divergence [Baeza-Yates and Ribeiro-Neto, 2011] and Okapi BM25) and exploits patent meta-data and citation structures. While our evaluation excludes 22 of the 1,303 topics for which no relevant English documents were available, the difference in the MAP score between our evaluation and the full 1,303 topic evaluation of PATATRAS is negligible. We exclude 22 queries because the focus of our research has been on term analysis and errors related to term matching process of ranking functions. Therefore, we eliminated data curation errors and IPC filter errors — as they will be described in Section 3.3.1 and Section 3.3.2 — to increase the accuracy of our data analysis results.

[3] Average Recall

3.2 Data Collection

We target CLEF-IP 2010 and 2011 collections[4] for our research. CLEF-IP 2010 is a set of patents from the 'European Patent Office' (EPO), and CLEF-IP 2011 includes the same patent data collection from EPO, but with the addition of a new set of patents from 'World Intellectual Property Organization' (WIPO). CLEF-IP 2010 contains 2.6 million patent documents and CLEF-IP 2011 consists of 3 million patent documents. The English test sets of CLEF-IP 2010 and CLEF-IP 2011 correspond to 1,303 and 1,351 topics respectively.

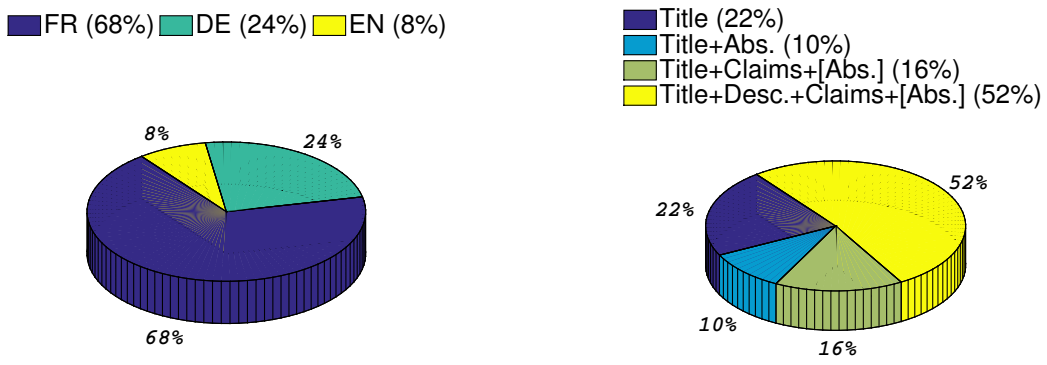
Each patent in the collection consists of multiple versions of documents in the XML format, labelled as A1, A2, ..., B1, and B2. The letter 'A' refers to different versions of patent applications. The 'B' versions refer to granted patents. Each of these versions contains some updates to the text, citations, and claims of previous one. As recommended by Magdy [2012], we merge different versions of a single patent into one single document. The content of each section in the merged document is taken from the latest available versions of documents, which leads to presence of some patents in the collection with many missing content fields. The problem of missing data is in some cases so significant that some of these patents consist only of title.

The patent collections contain material in three different languages: English, German, and French. Granted published version of a patent (i.e., the 'B' version) by the EPO should contain the claims section manually translated into all three languages. In addition, all patents have the title in the three languages. The description section of all patents is always provided in the original submission language only.

Test topics provided are English, German, and French patent applications, which are used as a query for the retrieval system. Topics for CLEF-IP 2010 are patent applications rather than granted patents as in 2009. Therefore, non-English patent applications did not contain any English translations in any section except the title and the abstract. However, in CLEF-IP 2010, 1,302 out of 1,303 test topics were English and only one was German.

Figure 3.1(a) shows the percentage of the English, German, and French patents in the CLEF-IP 2010 collection. Since some patents in the collection do not contain all sections, and since some of the non-English patents do not contain translations into English, Figure 3.1(b) presents the distribution of the missing English sections in the patents. Figure 3.1(b) shows the amount of the English content present in the patents in the 2010 collection, where only 52% of the patents in the collection were complete English documents. 16% of the collection included the titles and claims sections only, while some of them contained the abstract section as well. These patents are not complete patent documents, but at the same time, they are not short because of the presence of the claims section which contains most of the important information about the invention disclosed. 32% of the patents do not include the description or the claims sections in English, while most of them included the titles only, which means that the retrievability of these patents is expected to be very low, since they

[4] http://www.ifs.tuwien.ac.at/~clef-ip/



(a) (b)

Figure 3.1: (a) Percentage of English, German, and French patents in CLEF-IP 2010 collection. (b) Completeness of the presence of English text in the CLEF-IP 2010 patent collection. [Magdy, 2012].

contain only a very small number of words. The overall aim of Figure 3.1(b) is to show that the documents in the patent collection are not homogeneous since many of them are in some respect incomplete [Magdy, 2012].

3.3 Errors Caused by Baseline Settings

Data curation and IPC filter used in baseline settings are two sources of errors; in this section, we will discuss these two origins of the errors.

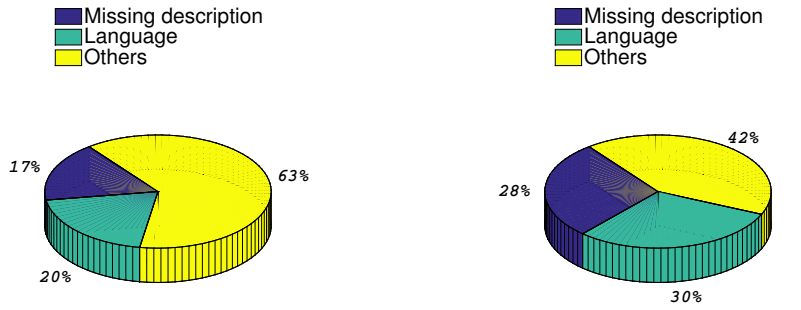
3.3.1 Data Curation Errors

Our baseline system cannot retrieve some relevant patent documents because of two main characteristic of CLEF-IP data collection:

1. **Missing description:** As we described in Section 3.2, some patents in the union collection lose the contents of some sections due to merging different versions of patents. Therefore, relevant patents with missing description are not retrieved by our system.

2. **Non-English relevant patents:** CLEF-IP data collection has been designed for a multilingual patent search and it consists of patents in three different languages: English, German, and French. However, our baseline *IR* system is not designed for multilingual search and it cannot retrieve non-English relevant patents.

We calculate the percentage of errors caused by data curation in this experiment. As it has been illustrated in Figure 3.2(a) , overall, 37% of errors are due to CLEF-IP data



(a) Cut-off rank (k) = 100 (b) Cut-off rank (k) = 1,000

Figure 3.2: Average percentage of errors due to missing description, language. Overall, 37% of errors are because of data curation while 63% of English complete patent documents cannot be retrieved. Increasing k from 100 to 1,000 reduces the errors of the yellow area, but the value of 42% is still notable.

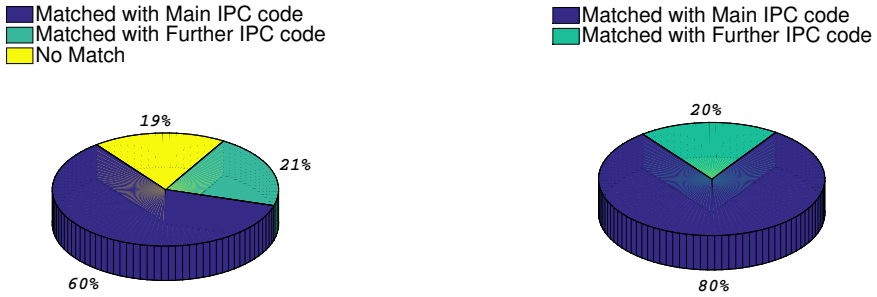
curation (missing description and non-English relevant patents) while the majority of relevant patents, which are not retrieved (63%), are full English patent documents (Figure 3.2). These results indicates that the baseline retrieval system is ineffective to retrieve the majority of the relevant patents because of other reasons. In this research, we are interested in the other reasons that result in low effectiveness of general IR techniques in patent domain. Figure 3.2(b) shows that by increasing the cut-off rank to 1,000, still considerable percentage of full English relevant patents — about 42% — are not retrieved.

3.3.2 Classification Code Mismatch

As we mentioned in Section 3.1, IPC codes (Section 2.1) are assigned to patent queries to filter the search results by constraining them to have common IPC codes with the patent query. In this section, we investigate the errors caused by classification code mismatch between topics (queries) and relevant documents for three different level of hierarchy.

(I) Applying three first IPC components for filtering (filter type I)
First, we examine the effect of filtering out the patents, which their three first symbols of IPC code, including section, class, and subclass (e.g., *C07C* in Figure 2.2), do not match with the patent query. We have applied this filter to our baseline system. As a consequence, relevant patents, which do not share these three symbols of the IPC code with the patent query, are not retrieved by the system.

Our experiments show that around 19% of the not-retrieved relevant patents do not share any IPC code with the patent query, but the majority of them have main IPC code of the query, and about 21% have, at least, one of the further IPC codes



(a) Not-retrieved relevant patents (False Negative (FN) patents)

(b) Retrieved relevant patents (True Positive (TP) patents)

Figure 3.3: Classification code overlap between the query and non-relevant retrieved patents (False Negative (FN) patents).

of the query (Figure 3.3(a)). We repeat the experiments for the true positive (TP) patents; as it has been shown in Figure 3.3(b), 80% of TP patents have an overlap with the main IPC code of the query and 20% with, at least, one of the query further IPC codes. Although we cannot retrieve around 19% of relevant patents as a result of applying the IPC filter, we still keep using the filter in our experiments for the following two main reasons:

1. CLEF-IP 2010 collection contains 2.6 million patent documents. If we do not use the IPC filter, it will take long time to compare each patent in the whole collection with the query. Nonetheless, if we apply the filter, this process will take faster because the matching process is done on only the portion of the collection, which shares an IPC code with the patent query not the whole collection. Since only less than 19% of errors are due to a classification mismatch, we continue our analysis by keeping the filter on. The matching process is computationally much faster when we apply the IPC filter. In trade off between losing the percentage of the relevant patents and faster computation, we choose the efficient computation. The computational time is critical in patent prior art search because the query is the description of the the patent query, consisting of thousands of words.

2. The precision in the top $k(= 100)$ significantly drops, when we rank the whole collection versus only a subset of patents that have the same classification code with the patent query.

We conduct the following experiment to justify the first above-mentioned reason. First, we calculate the number of documents that should be processed during the ranking process per query after applying the filter. Then we plot the distribution of this number over all test topics. Figure 3.4 illustrates that the matching process should be only done over $25,000$ documents for the majority of queries. On average,

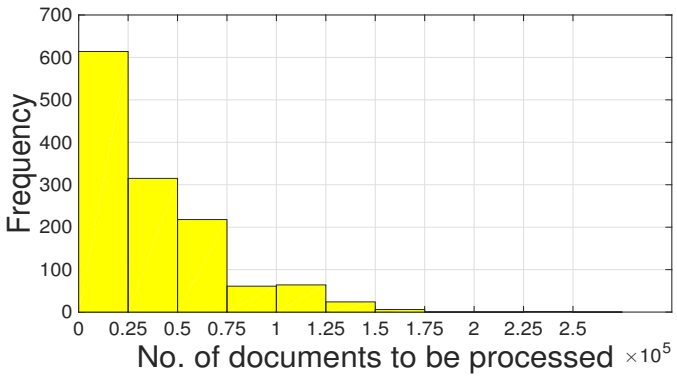


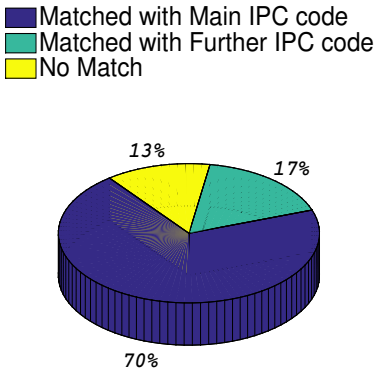
Figure 3.4: The distribution of the number of patents that should be ranked for each query over all test topics (1,303), after applying the IPC filter (filter type I). On average, the matching process for each query is done over 36,254 documents instead of the whole collection (2.6 million documents), which dramatically reduces the computational time.

this number is 36,254, which indicates that the system just needs to look into 36,254 documents per query on average instead of the entire collection that contains 2.6 million patent documents. Therefore, applying the IPC filter computationally saves us considerable amount of time.

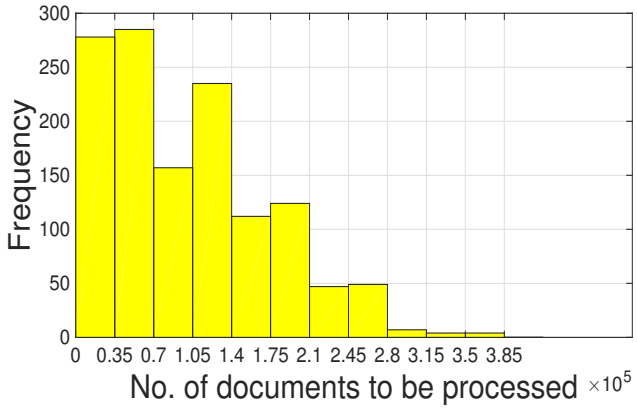
In trade off between losing 19% of relevant patents and making the ranking process faster, we choose faster computation. In addition, we notice that the histogram falls down by increasing the number of documents that should be processed; this means that for the majority of queries the matching process is done over less number of patent documents.

(II) Applying first two IPC code components for filtering (filter type II)
We hypothesise that the errors will be reduced, if we broaden the filter by selecting two first components of the query IPC, namely, section, and class (e.g., C07). We repeat the experiments for filter type II. The results have been illustrated in Figure 3.5. Figure 3.5(a) shows that we can reduce the errors related to filtering from 19% to 13% by omitting the subclass component. However, the number of documents that should be ranked increases from 36,254 to 99,754 on average. As it can be seen in Figure 3.5(b), the distribution of the number of documents that should be compared in matching process does not follow the falling trend as filtering with three first components. We conclude that this filter is not appropriate since we only reduce the error by 6% whereas the average number of documents, which should be processed, triples.

(III) Applying the first IPC code component for filtering (filter type III)
We can even make the filter more general by choosing only the first component, namely, section (e.g., C), corresponding to very general technical fields. Figure 3.6(a)



(a) The portion of patents in the collection which are matched with the query IPC code.

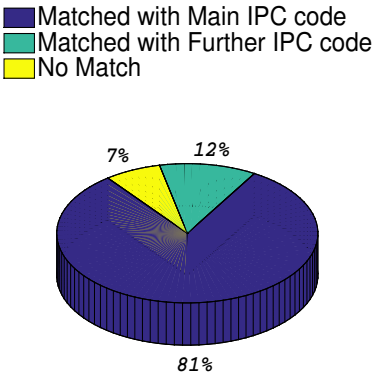


(b) The distribution of the number of patents should be ranked for each query over all test queries (1,303). In average, the matching process for each query is done over 99,754 documents instead of the whole collection (2.6 million documents), which dramatically reduce the computational time.

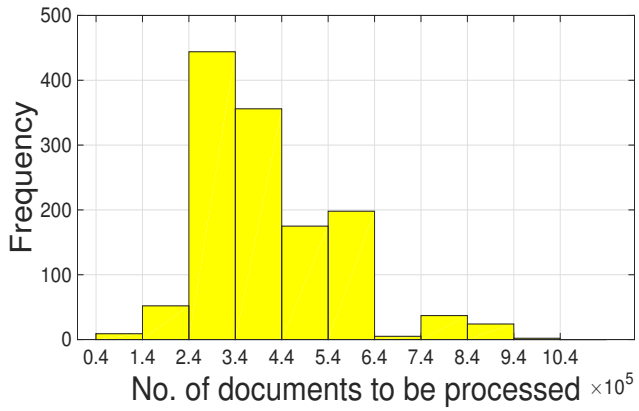
Figure 3.5: Applying first two IPC code components (Section and Class) for filtering

shows that about 7% of relevant patents do not share the most general component of the query IPC Code. Figure 3.6(b) shows the distribution of the number of patents should be ranked for each query after applying the IPC filter. The results show that the matching process for each query is done over 415,828 documents, on average, instead of the whole collection (2.6 million documents). This number is much higher than the number for previous filters, which shows that using only the first component of the IPC code is not computationally efficient because it does not reduce the computational time as well as it still causes 7% of the errors.

To recap our experiments related to the IPC code filtering, we showed that, in



(a) The portion of patents in the collection which are matched with the query IPC code. Filter: The first two components



(b) The distribution of the number of patents should be processed for each query after applying the IPC filter. In average, the matching process for each query is done over 415,828 documents instead of the whole collection (2.6 million documents). This number is much higher than using more restricted filters, so it is not computationally efficient.

Figure 3.6: Applying the first IPC code component for filtering (Section)

trade off between the errors related to applying IPC code filter and computationally efficient matching process, we got the best results when we applied the first three IPC code (section, class, and subclass) of the reference query as a filter. The filter reduced the number of documents to be ranked from the whole collection to 36,254 documents on average, so using the IPC filter saved a considerable amount of computational time for us.

The last point we discuss in this chapter is about the changes we made on rele-

Table 3.2: System performance after changing in relevant patents.

	Patent Query Not-Filtered	Patent Query Filtered
PRES	0.3910	0.5355
MAP	0.1214	0.1618
A. Recall	0.4008	0.5491

vant patents. We showed that some relevant patents are not retrieved because of the following three reasons: (1) their original language is not English; (2) they have a missing description, or (3) they do not have any of query IPC codes. Since our system is not designed for multilingual search and also we keep the IPC code filter due to the computational benefits, mentioned errors are fixed in our IR system. We exclude relevant patents with three above-mentioned errors for the accuracy of our experiments, analysing other errors, in the next chapter. This ended in 22 queries with no relevant complete English patent, sharing, at least, one IPC code with the query.

Table 3.2 indicates the performance of the system after filtering out relevant patents with above-mentioned errors. We can see that the poor performance of the baseline system is not mainly because of data curation and IPC filter. In the next chapter we investigate the errors caused by specific characteristic of the patents and prior art search. These errors consider the main reasons of the retrieval low effectiveness failure because as we showed in this chapter, they constitute 63% of the whole errors.

Chapter 4

Towards Optimal Query Term Selection

In this chapter, we will investigate the problem from term analysis perspective for both Patent Query and relevant documents because we are interested in finding what is wrong in term matching process between the Patent Query and relevant patents. We start with experiments which show sufficient term overlap between the Patent Query and relevant documents, then we introduce an Oracular Relevance Feedback scoring criteria to discriminate useful terms from noisy terms. We formulate two Oracular Query, based on this score, that give us an upper-bound performance. In addition, our experiments demonstrate the sufficiency of terms in the Patent Query to achieve a high performance. We try four simple query reduction approaches and we discuss the reasons that they are not efficient. Finally, we show that we can get improved using a simple, minimal feedback interactive approach.

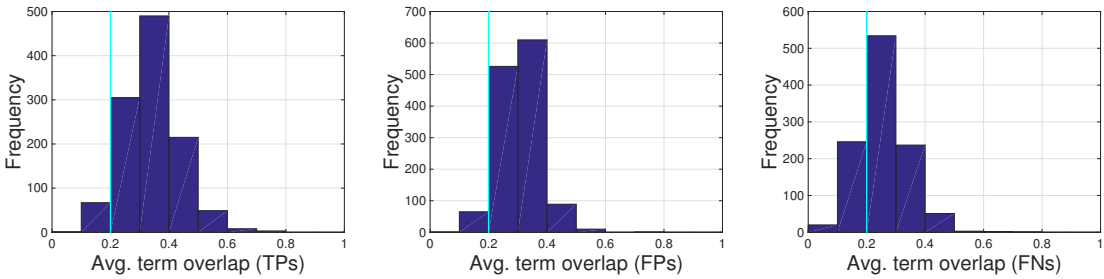
4.1 Term Mismatch

Standard retrieval models rank documents based on term matching between the query and documents. A significant term mismatch between the query patent and relevant patents has been mentioned the main cause for low effective patent prior art search in previous works [Roda et al., 2010] [Magdy, 2012]. We examine term overlap between Patent Query and three different patent documents: (i) retrieved relevant patents (TPs), (ii) retrieved non-relevant patents (FPs), and (iii) non-retrieved relevant patents (FNs), respectively, by calculating the average term overlap per query as follows:

$$TO(Q) = \frac{1}{|D|} \sum_{d \in D} \frac{|terms_{Q \cap d}|}{|Q|} \tag{4.1}$$

where $TO(Q)$ is the average term overlap per query — we calculate this score for TP, FP, and FN patents respectively, D is a collection of TP patents, FP patents, or FN patents for each query respectively, $|D|$ is the number of TP patents, FP patents, or FN patents for each query, $|terms_{Q \cap d}|$ is the number of query terms appear in each TP, FP, or FN patent, $|Q|$ is the size of the query.

The results has been illustrated in Figure 4.1. We conclude two main facts:



(a) Query and TP patents. (b) Query and FP patents. (c) Query and FN patents.

Figure 4.1: The distribution of term overlap between the query and documents over 1,303 test queries.

1. For the majority of queries (around 94% of queries), patent documents are retrieved (TPs and FPs) when they have a term overlap with a query above 0.2.

2. We can also see sufficient term overlap with the query for FN patents, whereas, compared to TPs and FPs, more queries can be seen with the term overlap less than 0.2 (about 24% of queries).

In summary, this experiment shows that low or zero term match is not the main cause of low effectiveness for patent prior art search.

4.2 Oracular Relevance Feedback System

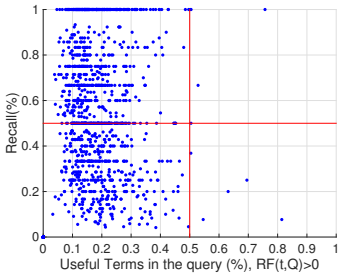
A query is optimal if it ranks all relevant documents before those that are not relevant, that is, it would lead to a ranking with an average precision of 1.0. A query is most likely to achieve a ranking that is as close to optimal as possible if it contains all terms that appear in all relevant documents, but explicitly discounts all terms that occur in non-relevant documents [Manning et al., 2008]. Inspired by this fact, we develop an oracular relevance feedback system, which extracts terms from the judged relevant documents to derive an upper bound on performance of standard Okapi BM25 and Language Models (LM) retrieval algorithms for patent prior art search.

4.2.1 Selecting Useful Terms

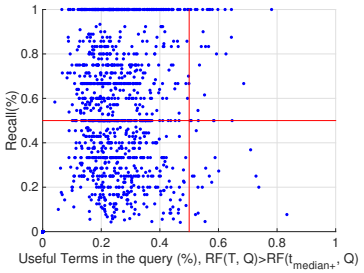
We aim at identifying the terms, which are considered useful in the query to achieve a ranking that is as close to optimal as possible. For this purpose, after an initial run of reference Patent Query, we calculate an Oracular Relevance Feedback (RF) score for each term in the top-100 retrieved documents as follows:

$$RF(t, Q) = Rel(t) - Irr(t) \tag{4.2}$$

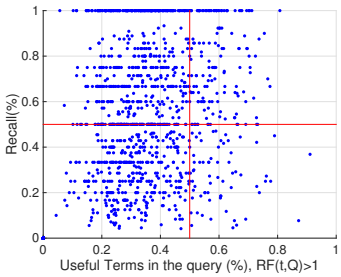
$$t \in \{\text{terms in top-100 retrieved documents}\}$$



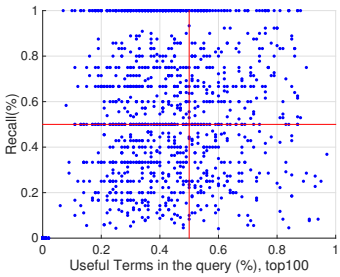
(a) Useful Terms: $\{t|RF(t,Q) > 0\}$



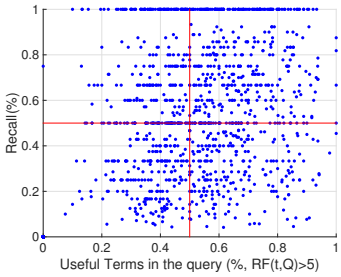
(b) Useful Terms: $\{t|RF(t,Q) > RF(t_{+median})\}$



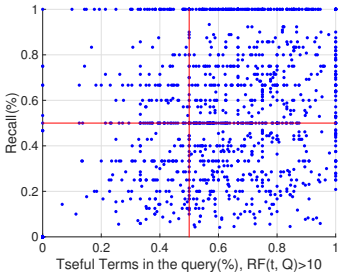
(c) Useful Terms: $\{t|RF(t,Q) > 1\}$



(d) Useful Terms: top 100 high-scored terms



(e) Useful Terms: $\{t|RF(t,Q) > 5\}$

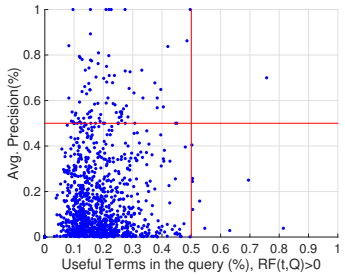


(f) Useful Terms: $\{t|RF(t,Q) > 10\}$

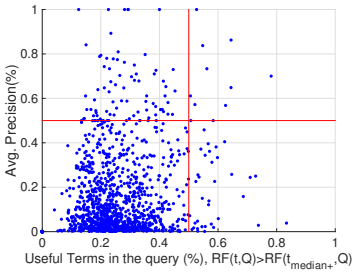
Figure 4.2: Scatter plot of Recall versus the existence of Useful Terms in query.

where $Rel(t)$ is the average term frequency in retrieved relevant patents and $Irr(t)$ is the average term frequency in retrieved irrelevant patents. We assume words with a positive score are Useful Terms since they are more frequent in relevant patents, while words with a negative score are Noisy Terms as they appear more frequently in irrelevant patents.

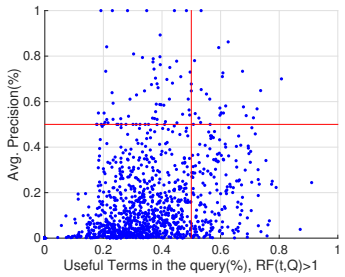
We yield the Oracular Relevance Feedback score to: (i) find a pattern for the system performance versus Useful Terms; (ii) show the term overlap with Useful Terms and Noisy Terms for TP, FN patents; (iii) examine the existence of Useful Terms in different sections of Patent Query.



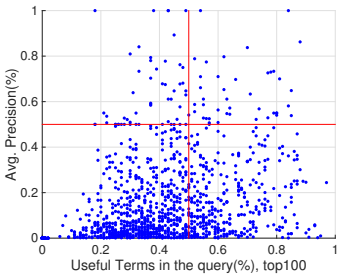
(a) Useful terms: $\{t|RF(t,Q) > 0\}$



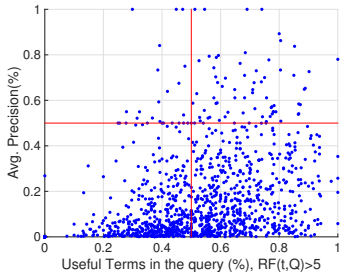
(b) Useful terms: $\{t|RF(t,Q) > RF(t_{+median},Q)\}$



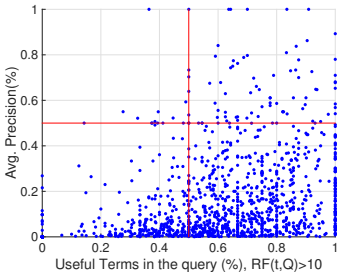
(c) Useful terms: $\{t|RF(t,Q) > 1\}$



(d) Useful terms: top 100 high-scored terms



(e) Useful terms: $\{t|RF(t,Q) > 5\}$



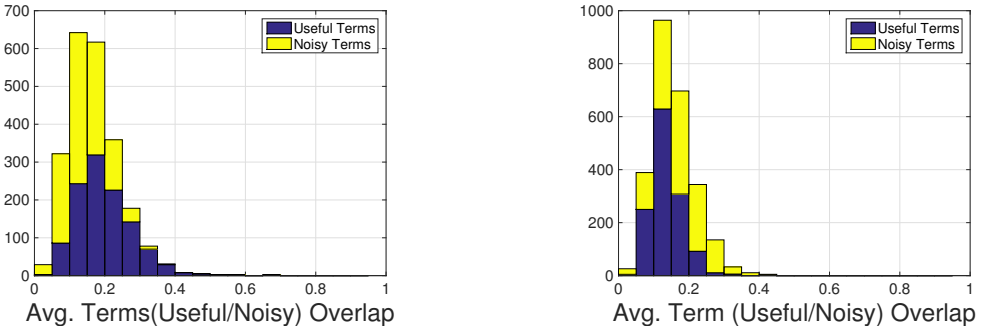
(f) Useful terms: $\{t|RF(t,Q) > 10\}$

Figure 4.3: Scatter plot of Average Precision versus the existence of Useful Terms in query.

4.2.1.1 Performance versus Useful Terms

In our first experiment, we investigate whether the existence of more Useful Terms in the reference Patent Query means achieving a higher performance. In other words, we seek for a pattern between the performance and the existence of Useful Terms in initial Patent Query. We define four different criteria to select Useful Terms:

1. Terms with positive RF scores ($RF(t,Q) > 0$)).
2. Terms with the score higher than the positive median score ($RF(t,Q) > RF(t_{+median},Q)$).
3. Terms with the score higher than a constant: 1, 5, and 10 ($RF(t,Q) > 1,5,10$)).



(a) TPs (b) FPs

Figure 4.4: The distribution of the term overlap between the query and Useful Terms/Noisy Terms in TPs and FPs. Relevant patents have higher term overlap with Useful Terms while irrelevant patents have higher term overlap with Noisy Terms.

4. Top-100 high-scored terms.

Figures 4.2 and 4.3 show the scatter plot of the performance (Average Precision, and Recall) versus the existence of Useful Terms in query. We expected to see a higher performance for the queries which contain more Useful Terms and a lower performance for the ones with less Useful Terms. However, unlike our first assumption, we do not see any correlation between the performance and the presence of Useful Terms in the query. The pattern for the recall is irregular while there is a very weak correlation between Average Precision and Useful Terms for top-scored words ($RF(t, Q) > 10$). This experiment explicitly indicates that term mismatch is not the main reason for low effectiveness of prior art search.

4.2.1.2 Term Overlap with Useful Terms and Noisy Terms

In the second experiment, we check the term overlap with Useful Terms and Noisy Terms for TP and FP patents. Figure 4.4 shows that relevant patents have a higher term overlap with the Useful Terms while irrelevant patents have a higher term overlap with the Noisy Terms. This experiment shows that Noisy Terms are the main reason that irrelevant patents are retrieved at top of the list.

4.2.1.3 Useful Terms in Different Sections of Patents

Patents are structured documents containing Title, Abstract, Description, and Claims (section 2.1). In this experiment, we investigate Useful Terms in different sections of patents. Table (4.1) shows the average number of Useful Terms in different sections of Patent Query. As it can be seen, Description has the highest number of useful terms in both cases where RF score threshold (τ) is '0' and '1'. When $\tau = 0$, the average number of the Useful Terms in Description is quite twice of when $\tau = 1$. Compared to other sections, Description contains more Useful Terms, which proves

Table 4.1: Average number of Useful Terms in the different sections of Patent Query

	Title	Abstract	Description	Claims
$\tau = 0$	2	12	164	26
$\tau = 1$	2	9	80	19

Table 4.2: Average percentage of Useful Terms in the different sections of Patent Query

	Title	Abstract	Description	Claims
$\tau = 0$	0.4	0.37	0.27	0.33
$\tau = 1$	0.36	0.29	0.14	0.25

why we achieved higher performance querying with Description (Section 3.1). Table (4.2) shows the average percentage of Useful Terms in different sections of Patent Query. It shows that, overall, Useful Terms constitute less than 50% of the whole words in each section of Patent Queries. For example, we can see that only 27% of the whole Patent Query in average are Useful Terms and the rest are irrelevant terms.

4.2.2 Oracular Query Formulation

As it explained in section 4.2.1.1, we could not find a pattern for the performance and the existence of the Useful Terms in Patent Query. In this section, we examine the system effectiveness for queries formulated by terms selected by Oracular Relevance Feedback (*RF*) system. We formulate two different Oracular Queries.

The first query is formulated by selecting terms in the top-100 retrieved documents using Oracular Relevance Feedback score and we call it Oracular Query:

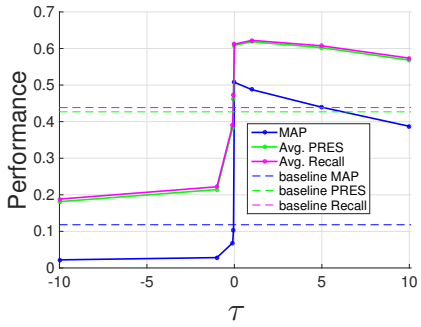
$$Oracular\ Query = \{t \in top - 100 | RF(t, Q) > \tau\} \tag{4.3}$$

We empirically seek to evaluate the threshold τ on $RF(t, Q)$ and query size yielding the best oracular query. Table 4.3 and Figure 4.5(a) show that the Oracular Query far outperform the baseline query (reference Patent Query), and it approximately performs twice as well on the PATATRAS system, the best competitor in CLEF-IP 2010 system. In Table 4.3, we also compare the influence of weighed and unweighed terms for both baseline query and Oracular Query. It can be seen that weighing terms in Patent Query with their frequency helps the performance while weighing terms in Oracular Query with $RF(t, Q)$ harms the performance. Figure 4.5(a) shows how the performance changes by the values of τ. We remark two important facts:

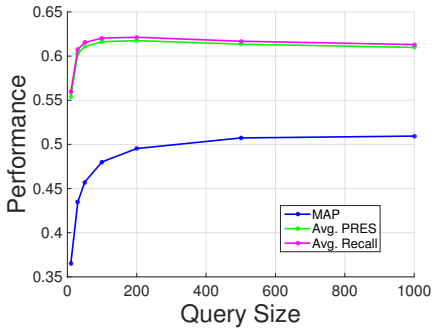
1. Including slightly Noisy Terms (i.e., τ just slightly less than 0) leads in an unexpected steep drop-off in performance.

Table 4.3: Performance for the Patent Query, Oracular Query, and Top CLEF-IP 2010 (PATATRAS).

	PATATRAS	Pat.Query W:TF	Pat.Query W:1	Oracular W:RF(t,Q)	Oracular W:1	Oracular(PQ) W:1
PRES	N/A	0.535	0.427	0.609	0.609	0.617
MAP	0.264	0.162	0.118	**0.462**	**0.507**	**0.436**
Recall	N/A	0.549	0.438	0.613	0.612	0.622



(a) Oracular Query performance versus the threshold τ.



(b) Oracular Query performance versus the query size.

Figure 4.5: Oracular Query performance versus various values of the threshold τ and query size

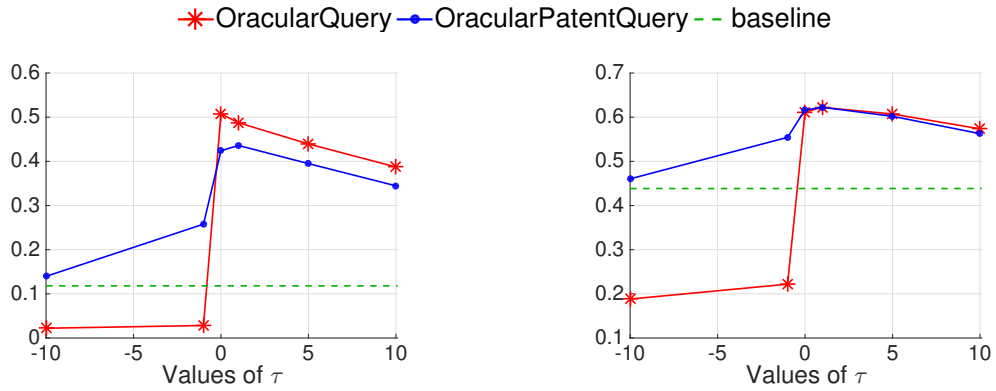
2. We achieve the highest MAP for the Oracular Query formulated by selecting the terms with $RF(t, Q) > 0$.

Figure 4.5(b) shows that the performance increases notably when we include terms up to 200 while formulating a query, but it remains quit unchanged when we include more than 200 terms.

We seek to establish that the terms within a reference Patent Query are sufficient for a strong performance, so, we formulate the second query by selecting oracular terms that also occur in the reference patent query We call it Oracular Patent Query:

$$\textit{Oracular Patent Query} = \{t \in Q | RF(t, Q) > \tau\} \tag{4.4}$$

As it has been shown in Table 4.3 and Figure 4.6, the system performance for the Oracular Patent Query is also considerably improved compared to the baseline and PATATRAS. The results indicate that the Patent Query has sufficient terms for an improved performance. We compare MAP and Recall for both Oracular Query and Oracular Patent Query in Figure 4.6. On one hand, We notice that MAP for Oracular Patent Query is slightly less than MAP for Oracular Query. We justify that it is due to some extra terms in top-100 vocabulary set that they are absent within the Patent



(a) Mean Average Precision τ.

(b) Average Recall

Figure 4.6: Comparing the performance of Oracular Query and Oracular Patent Query for various values of the threshold τ

Query. On the other hand, Oracular Patent Query performance drops slower for negative values of τ, which shows that Oracular Patent Query contains less Noisy Terms than Oracular Query. This explains why query expansion techniques is not too effective for patent prior art search. We also conclude that the the existence of the Noisy Terms is the main cause of low effectiveness in prior art search.

To recap, our experiments related to Oracular Relevance Feedback system suggest two important conclusions:

1. Query reduction should suffice for effective prior art patent retrieval; and

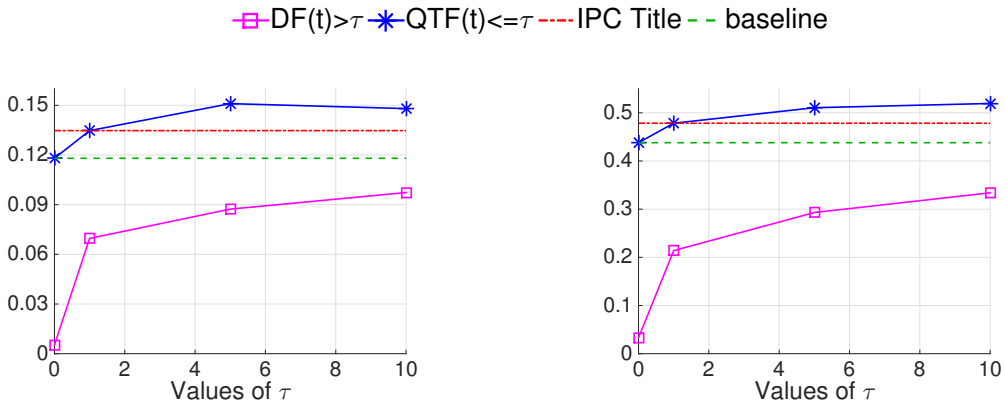
2. Very precise methods for eliminating poor query terms in the reduction process are required.

4.3 Query Reduction: Approximating the Oracular Patent Query

The gain achieved using the Oracular Patent Query method motivates us to explore various methods to approximate the terms selected by this query without "peeking at the answers" provided by the actual relevance judgements. We first attempt this via fully automated methods and then proceed to evaluate semi-automated methods based on interactive relevance feedback methods.

4.3.1 Automated Reduction

For automated reduction we first examine three simple approaches, then we apply Pseudo Relevance Feedback for query term selection.



(a) Mean Average Precision.

(b) Average Recall.

Figure 4.7: Comparing system performance for three different query reduction approaches and their changes with a threshold τ.

4.3.1.1 Simple Query Reduction Approaches

We, first, apply three simple approaches to reduce the initial patent queries aiming at approximating the Oracular Patent Query: (i) removing document frequent terms; (ii) removing less frequent terms in Patent Query; (iii) Removing Terms in IPC Titles.

Removing Document Frequent Terms

In standard IR, removing terms appearing highly frequently across documents in the collection can improve retrieval effectiveness [Manning et al., 2008]. Inspired by this fact, we hypothesize that we will improve the performance by pruning out highly frequent words in top-100 retrieved documents after an initial run of Patent Query. To identify highly frequent terms, we calculate the average term frequency over top-100 documents for each word and we call it Document Frequent (DF) score, as follows:

$$DF(t, Q) = \frac{1}{100} \sum_{d_i \in D} TF(t, d_i) \tag{4.5}$$

where $D = \{d \in \text{Top-100 retrieved documents}\}$, and $TF(t, d_i)$ is the term frequency of each term in document d_i.

We remove words with DF score higher than τ ($DF(t, Q) > \tau$) from Patent Query. Figure 4.7 illustrates how the performance change by different values of the threshold τ illustrates that removing document frequent words for different threshold τ hurts the performance (magenta line). As it can be seen, the performance converges with the baseline performance when τ goes higher (e.g., 500). This means that there is no term with such a high DF score, and we do not empirically remove any term from the Patent Query. Overall, removing document frequent terms from Patent Query does not consider an appropriate approach since it ruins the performance.

Removing Less Frequent Terms in Patent Query
Frequent terms inside long and verbose queries are considered important [Maxwell and Croft, 2013]. However, we hypothesise that we may improve the effectiveness by removing terms appearing less frequently in the Patent Query. Therefore, we remove terms with the frequency less than the threshold τ ($QTF(t) <= \tau$). The blue line in Figure 4.7 indicates that the performance gets slightly better than the baseline when we remove less frequent terms in Patent Query. As it can be seen the best MAP achieved when $\tau = 5$.

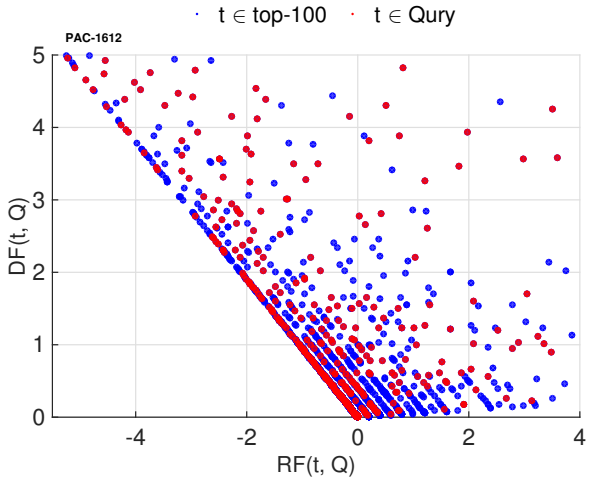
Removing Terms in IPC Titles
The titles of classification indicate their intended content by using a single phrase or several related phrases linked together. We used words in IPC code titles for each patent query to reduce the query, based on the assumption that they are common to all patents belonging to the same category and may be considered as stop-words. As it can be seen in Figure 4.7 (red line), this approach slightly helps the performance.

We showed in above experiments that removing document frequent terms did not help the effectiveness and two other approaches had a trivial influence on improving the system effectiveness. In the following experiment, we use an anecdotal example of a sample Patent Query to analyse why these approaches did not help. Figure 4.8 shows a scatter plot of DF score and RF score for a sample query — PAC-1612. Each blue point is a vocabulary in top-100 retrieved document vocabulary set. First, we remark a negative correlation between $DF(t, Q)$ and $RF(t, Q)$, however, it does not help because as it has been illustrated in Figure 4.8(b), by removing document frequent terms ($DF(t, Q) > \tau$), we will remove many Useful Terms ($RF(t, Q) > 0$) . Red points in Figure 4.8(a) are all query terms and red points in Figure 4.8(b) are query terms with term frequency higher than 5 ($QTF(t) > 5$) — where we got the best performance. Comparing Figures 4.8(a) and 4.8(b) show that we remove considerable amount of Noisy Terms by removing terms with $QTF(t) < 5$. On the other hand, it can be seen that many Useful Terms are also removed. Remained terms are not purely Useful Terms since they are contaminated with the Noisy Terms. We conclude that we achieve a trivial improvement over the baseline because proposed reduction techniques cannot precisely filter the Noisy Terms out.

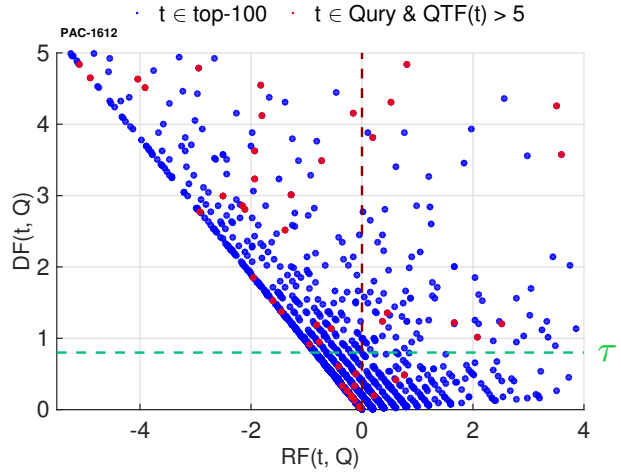
4.3.1.2 Query Reduction Using Pseudo Relevance Feedback

Pseudo Relevance Feedback (*PRF*) is an automated process without user interaction which assumes the top k ranked documents are relevant and the others are irrelevant [Baeza-Yates and Ribeiro-Neto, 2011]. We use *PRF* to select query terms Maxwell and Croft [2013] the same as what we did for Oracular Relevance Feedback system (Section 4.2). We assume that top 5 retrieved documents are relevant and the rest are irrelevant, then we calculate *PRF* score based on this assumption:

$$PRF(t, Q) = Rel(t) - Irr(t) \tag{4.6}$$



(a) Query terms ($t \in Query$) versus Document Frequent terms.



(b) Query terms ($t \in Query \bigwedge QTF(t) > 5$) versus Document Frequent terms

Figure 4.8: Anecdotal example for simple query reduction approaches. Blue points are all terms in a vocabulary set made of top-100 retrieved documents and red points are terms in the Patent Query.

$$t \in \{\text{terms in top-100 retrieved documents}\}$$

In the next step, we select the terms in the Patent Query that have the *PRF* score higher than the threshold τ ($PRF(t) > \tau$) to reformulate a reduced query. Figure 4.9 shows slight improvement over the baseline. Compared to the Oracular Term Selection system, this approach did not also help to get any notable improvement over the baseline.

We analyse why the term selection technique using Pseudo Relevance Feedback

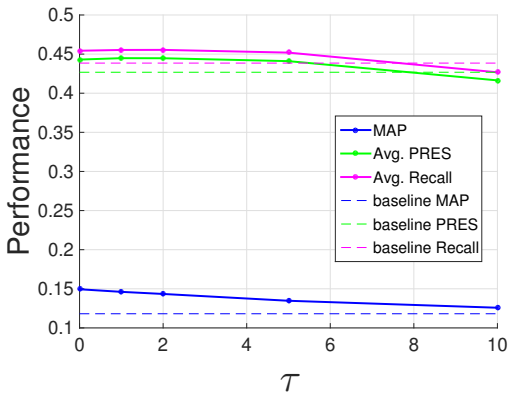
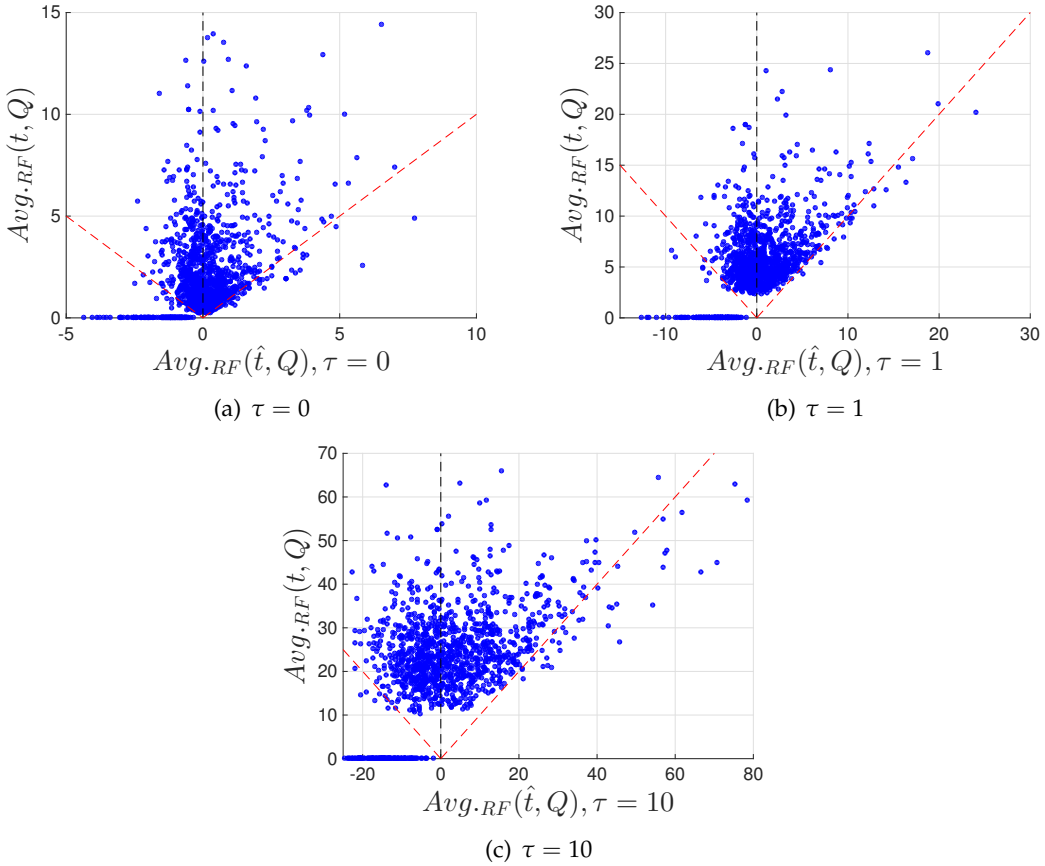


Figure 4.9: Query reduction using PRF for various value of the threshold τ.



(a) $\tau = 0$ (b) $\tau = 1$

(c) $\tau = 10$

Figure 4.10: Comparing *RF* score of top Relevance Feedback terms and Pseudo Relevance Feedback terms for different values of the threshold τ.

fails to approximate the Oracular Patent Query in the following experiment. We seek for a pattern between top Relevance Feedback terms and top Pseudo Relevance

Feedback terms. For this purpose, we calculate the average *RF* score of both terms with top *RF* score and terms with top *PRF* score for each query as follows:

$$Avg_{RF}(t, Q) = \frac{1}{|t|} \sum RF(t, Q), \quad t \in \{t | RF(t, Q) > \tau\} \tag{4.7}$$

$$Avg_{RF}(\hat{t}, Q) = \frac{1}{|\hat{t}|} \sum RF(\hat{t}, Q), \quad \hat{t} \in \{\hat{t} | PRF(\hat{t}, Q) > \tau\} \tag{4.8}$$

where $Avg_{RF}(t, Q)$ is the average *RF* score for top *RF*-scored terms ($RF(t, Q) > \tau$), $Avg_{RF}(\hat{t}, Q)$ is the average *RF* score for top *PRF*-scored terms ($PRF(\hat{t}, Q) > \tau$), τ is a threshold for the score to select terms, t is a symbol for terms with high RF score, $\hat{t}$ is a symbol for terms with high *PRF* score.

Figure 4.10 shows a scatter plot of average *RF* score for top Relevance Feedback terms and top Pseudo Relevance Feedback terms. First, we observe that almost the RF score of top Relevance Feedback terms is lower than the RF score of top Pseudo Relevance Feedback terms for almost all queries ($Avg_{RF}(t, Q) > Avg_{RF}(\hat{t}, Q)$). We can also see that for about half of the queries, $Avg_{RF}(\hat{t}, Q)$ is negative, which indicates we are selecting Noisy Terms by their Pseudo Relevance Feedback score rather than Useful Terms. Second, we can find a very slight positive correlation toward selecting positive terms by Pseudo Relevance Feedback, which is the reason that we could get a slight improvement.

As the last experiment for the automated query reduction, we illustrate why four proposed query reduction approaches failed to approximate the Oracular Patent Query using an anecdotal example of a sample query about an invention related to "emulsifier". Figure 4.11 shows the raw abstract of the invention, and terms and their associated *RF* scores for each approach. Terms are chosen based on the scores for each approach as follows:

$$\{t | DF(t)/QTF(t)/PRF(t) > 10\}$$

For the IPC title terms, all terms appearing in IPC title are displayed since they do not have any score. It can be seen that the four methods fail clearly to discriminate between Useful Terms and Noisy Terms. As one example, important stemmed terms like "enzym" and "starch" have been removed by Document Frequent pruning approach, which hurts query quality. As another example, retaining IPC code title terms yields more Noisy Terms than Useful Terms (19 out of 32, and few of them with a very negative score like "amylos" or "saccharid"). This can justify slight improvement in performance when we prune terms in IPC title. Overall, all methods may retain highly negative terms and results from Section 4.2.2 showed that the inclusion of even slightly negative terms can significantly hurt the performance.

```
PAC-1293

Abstract: The invention relates to an emulsifier, a method for
preparing said emulsifier, and to its use in various applications
, primarily food and cosmetic applications. The invention also
relates to the use of said emulsifier for the creation of an
elastic, gelled foam. An emulsifier according to the invention is
based on a starch which is enzymatically converted, using a
specific type of enzyme, and modified in a specific
esterification reaction.

(1) DF Terms: starch:14.64, enzym:29.49, amylos:-20.15,
oil:8.63, dispers:-8.66, ph:-4.55, dry:-6.21, heat:-2.26,
product:-5.48, slurri:-11.48, viscos:7.77, composit:-4.49,
reaction:-1.97, food:-11.94, agent:5.19, debranch:-10.58,
reduc:-6.37, fat:-12.83, prepar:-0.82, hour:-5.42,
waxi:19.41, deriv:11.97, content:-3.38, aqueou:0.38,
saccharid:-11.95, ml:-0.79, cook:-10.04, modifi:5.65,
solid:5.50, sampl:6.27, mix:2.48, minut:-1.68, dri:-0.91,
gel:-9.85, activ:5.98, corn:-5.27, alpha:12, sprai:-2.74

(2) QTF Terms: starch:14.64, emulsifi:6.72, succin:-3.46,
enzym:29.49, emuls:12.66, hydrophob:5.45, anhydrid:-5.47,
reaction:-1.97, octenyl:-0.66, stabil:3.64, alkenyl:0.06,
reagent:1.17, carbon:0.12, potato:3.74, alkyl:-0.33,
wt:-4.57, ether:1.96, enzymat:-3.45, convers:10.44,
chain:-5.53, atom:0.03, ph:-4.55, treat:-0.89,
ammonium:-1.96, food:-11.94, amylos:-20.15,
glucanotransferas:-0.86, glycidyl:-0.40, glycosyl:-0.02,
dry:-6.21, deriv:11.97, transferas:0.89, foam:-0.49,

(3) IPC title Terms:cosmet:3.77, toilet:0.18, prepar:-0.82,
case:0.47, accessori:-0.01, store:-0.37, handl:0.07,
pasti:-0.17, substanc:-1.21, fibrou:-0.01, pulp:-1.28,
constitut:-0.06, paper:1.26, impregn:-0.11, emulsifi:6.72,
wet:-0.28, dispers:-8.66, foam:-0.49, produc:-0.57,
agent:5.19, relev:0.18, class:0.053, lubric:-0.38,
emuls:12.66, fuel:-0.011, deriv:11.97, starch:14.64,
amylos:-20.15, compound:-0.63, saccharid:-11.95,
radic:1.03, acid:-3.19

(4) PRF Terms: starch:14.64, encapsul:17.50, chees:-4.22,
oil:8.63, hydrophob:5.45, agent:5.19, casein:-2.19,
degrad:17.13, deriv:11.97, tablet:5.30, debranch:-10.58,
imit:-1.13, viscos:7.77, oxid:5.97, activ:5.98, osa:9.32,
funnel:2.68, amylas:26.06, amylopectin:-7.14, maiz:20.61,
blend:-3.17, waxi:19.41, convert:31.81,
```

Figure 4.11: Four query reduction approaches on a sample query. Top terms retained by each method are shown. Numerical oracular scores $RF(t, Q)$ are provided indicating whether the term was useful (blue/positive) or noisy (red/negative).

Table 4.4: System performance using minimal relevance feedback. τ is RF score threshold, and k indicates the number of first relevant retrieved patents.

	$k=1$ $\tau=0$	$k=1$ $\tau=1$	$k=3$ $\tau=0$	$k=3$ $\tau=1$
MAP	0.303	**0.304**	0.388	0.387
Recall	0.504	0.509	0.576	0.579

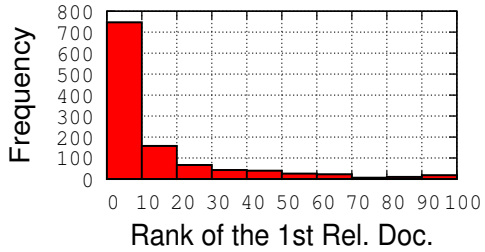


Figure 4.12: The distribution of the first relevant document rank over test queries.

4.3.2 Semi-automated Interactive Reduction

Our sample analysis of specific queries and terms selected via our oracular approach suggests that automated methods fall far short of optimal term selection. This leads us to explore another approach of approximating the oracular query derived from relevance judgements by using a subset of relevance judgements through interactive methods. Specifically, to minimize the need for user interaction, in this section we analyse the performance of an oracular query derived from only the first relevant document identified in the search results. Using this approach, table 4.4 shows that we can double the MAP in comparison to our baseline and also outperform the PATATRAS system.

Furthermore, to establish the minimal interaction required by this approach, Figure 4.12 indicates that the baseline methods return a relevant patent approximately 80% of the time in the first 10 results and 90% of the time in the first 20 results. Hence, such an interactive approach requires relatively low user effort while achieving state-of-the-art performance.

Chapter 5

Conclusions

In this thesis, we investigated the reasons that patent prior art searches are less effective than the other web search applications. We started with recognising errors due to data curation and baseline settings that make small portions of the whole retrieval errors. However, the main portion of the errors are due to term matching process of retrieval ranking functions. Hence, we looked at the patent prior art search from a term selection perspective. While previous works proposed different solutions to improve retrieval effectiveness, we focused on term analysis of the patent query and top-100 retrieved patents. After defining an Oracular Query based on relevance judgements, we established both the sufficiency of the standard LM retrieval scoring models and query reduction methods to achieve state-of-the-art patent prior art search performance. After finding that automated methods for query reduction approaches fail to offer significant performance improvements, we showed that we can double the MAP with minimum user interaction by approximating the Oracular Query through a relevance feedback approach with a single relevant document. Given that such simple interactive methods for query reduction with a standard LM retrieval model outperform highly engineered patent-specific search systems from CLEF-IP 2010, we concluded that interactive methods offer a promising avenue for simple but highly effective term selection in patent prior art search.

5.1 Contributions

We briefly summarise the major contributions of our work as follows:

1. **Development of an Oracular term selection system:** We built an oracular term selection system from known relevance judgements to formulate an oracular query that far outperformed the baseline and the best-performed competitor on CLEF-IP 2010. Experiments related to oracular system suggests the necessity of precise query reduction and term selection techniques to improve the effectiveness of patent prior art search.

2. **Analysis of automated query reduction techniques for patent prior art search:** We examined four simple query reduction methods to select the positive terms and prune out negative terms. We illustrated that these approaches are ineffi-

cient because they can not discriminate between useful terms and noisy terms. Since our system is over-sensitive to the existence of noisy terms, we could not achieve high performance via these simple methods.

3. **Proposal of a semi-interactive method for query term selection:** We showed that a simple minimal interactive relevance feedback approach, where terms are selected by only the first retrieved relevant document performs as well as a highly engineered patent-specific system on CLEF-IP 2010.

5.2 Future Work

In this research, we analysed the key reasons that generic IR methods are not effective for patent prior art through various experiments, which may open further research on the topic of prior art search. Thus, we describe the limitations and discuss further improvements.

5.2.1 Exploring other Term Scoring Methods

Our term scoring method inspired by Rocchio optimal query [Manning et al., 2008]. We used this score to select query terms, which resulted in a remarkable improvement in the performance. However, exploring other existing term scoring techniques like Kullback-Leibler divergence [Baeza-Yates and Ribeiro-Neto, 2011] may improve the results.

5.2.2 Exploring more Sophisticated Query Reduction Methods

One of the most important findings of our research was the existence of useful terms sufficiently inside the reference Patent Query. We showed that a query formulated by selecting these terms considerably outperforms the baseline. Thus, in this thesis, we tried four simple query reduction techniques. However, we only got slight improvement over the baseline because the retrieval models are over-sensitive to noisy terms and our proposed reduction approaches were incapable of discriminating useful terms and noisy terms. Given the necessity of a precise query term selection technique that can differentiate useful terms from noisy terms, applying more sophisticated query reduction approaches — e.g, query term selection technique proposed in [Maxwell and Croft, 2013] using affinity graph and random walk — is an important open area of research for query term selection in patent prior art search.

5.2.3 Considering Phrasal Concepts for Query Reformulation

Our research was limited to only single terms in patent documents. However, one important characteristic of patents is that inventors use longer technical terms to describe their research ideas. Hence, phrasal concepts and terminology are frequently used as keywords in target patent documents. Hence, an obvious extension of this work is extracting phrasal concepts while reformulating the query.

5.2.4 Patent Retrieval Using Meta-data Social Information

A retrieval based on meta-data social information and social network analysis is a proper alternate to a traditional IR based on term matching process, when the retrieval problem based on term matching is difficult — like patent prior art search. Bibliographic meta-data in the XML file of a patent document contains details about its inventor, organisation, and other information, which can give rise in more effective retrieval. Recent studies aim at improving the IR process with information coming from social networks. This is commonly known as social IR. Regarding this social network structure of patents, it is possible to find possible prior works in the social profile of other inventors with the same research interest. Also competitive organisations may have developed the same or very close idea prior to the novel idea claimed in a patent application.

Bibliography

Azzopardi, L. and Vinay, V., 2008. Retrievability: an evaluation measure for higher order information access tasks. In *Proceedings of the 17th ACM conference on Information and knowledge management*, 561–570. ACM. (cited on page 12)

Bache, R. and Azzopardi, L., 2010. Improving access to large patent corpora. In *Transactions on large-scale data-and knowledge-centered systems II*, 103–121. Springer. (cited on pages 9, 12, and 17)

Baeza-Yates, R. A. and Ribeiro-Neto, B. A., 2011. *Modern Information Retrieval - the concepts and technology behind search, Second edition*. Pearson Education Ltd., Harlow, England. (cited on pages 2, 28, 46, and 54)

Balasubramanian, N.; Kumaran, G.; and Carvalho, V. R., 2010. Exploring reductions for long web queries. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 571–578. ACM. (cited on page 15)

Bashir, S. and Rauber, A., 2009a. Analyzing document retrievability in patent retrieval settings. In *Database and Expert Systems Applications*, 753–760. Springer. (cited on page 17)

Bashir, S. and Rauber, A., 2009b. Improving retrievability of patents with cluster-based pseudo-relevance feedback documents selection. In *Proceedings of the 18th ACM conference on Information and knowledge management*, 1863–1866. ACM. (cited on pages 12, 17, and 20)

Bashir, S. and Rauber, A., 2010. Improving retrievability of patents in prior-art search. In *Advances in Information Retrieval*, 457–470. Springer. (cited on pages 17 and 20)

Bashir, S. and Rauber, A., 2011. On the relationship between query characteristics and ir functions retrieval bias. *Journal of the American Society for Information Science and Technology*, 62, 8 (2011), 1515–1532. (cited on page 12)

Becks, D.; Mandl, T.; and Womser-Hacker, C., 2010. Phrases or terms? the impact of different query types. In *CLEF (Notebook Papers/LABs/Workshops)*. (cited on page 19)

Bendersky, M. and Croft, W. B., 2008. Discovering key concepts in verbose queries. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 491–498. ACM. (cited on page 15)

Bendersky, M.; Metzler, D.; and Croft, W. B., 2010. Learning concept importance using a weighted dependence model. In *Proceedings of the third ACM international conference on Web search and data mining*, 31–40. ACM. (cited on page 15)

Cao, G.; Nie, J.-Y.; Gao, J.; and Robertson, S., 2008. Selecting good expansion terms for pseudo-relevance feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 243–250. ACM. (cited on pages 13 and 14)

Cetintas, S. and Si, L., 2012. Effective query generation and postprocessing strategies for prior art patent search. *Journal of the American Society for Information Science and Technology*, 63, 3 (2012), 512–527. (cited on page 19)

Croft, W. B.; Metzler, D.; and Strohman, T., 2010. *Search engines: Information retrieval in practice*. Addison-Wesley Reading. (cited on page 9)

D'hondt, E. and Verberne, S., 2010. Clef-ip 2010: Prior art retrieval using the different sections in patent documents. In *CLEF (Notebook Papers/LABs/Workshops)*. (cited on page 23)

D'hondt, E.; Verberne, S.; Alink, W.; and Cornacchia, R., 2011. Combining document representations for prior-art retrieval. In *CLEF (Notebook Papers/Labs/Workshop)*. (cited on page 19)

Fujii, A., 2007a. Enhancing patent retrieval by citation analysis. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 793–794. ACM. (cited on pages 19 and 22)

Fujii, A., 2007b. Integrating content and citation information for the ntcir-6 patent retrieval task. In *Proceedings of NTCIR-6 Workshop Meeting*, 377–380. (cited on page 22)

Fujii, A.; Iwayama, M.; and Kando, N., 2007. Introduction to the special issue on patent processing. *Information Processing & Management*, 43, 5 (2007), 1149–1153. (cited on page 17)

Fujita, S., 2005. Revisiting document length hypotheses: A comparative study of japanese newspaper and patent retrieval. *ACM Transactions on Asian Language Information Processing (TALIP)*, 4, 2 (2005), 207–235. (cited on page 23)

Ganguly, D.; Leveling, J.; and Jones, G. J., 2011a. United we fall, divided we stand: A study of query segmentation and prf for patent prior art search. In *Proceedings of the 4th workshop on Patent information retrieval*, 13–18. ACM. (cited on page 21)

Ganguly, D.; Leveling, J.; Magdy, W.; and Jones, G. J., 2011b. Patent query reduction using pseudo relevance feedback. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, 1953–1956. ACM. (cited on page 22)

Gobeill, J.; Pasche, E.; Teodoro, D.; and Ruch, P., 2010. Simple pre and post processing strategies for patent searching in clef intellectual property track 2009. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, 444–451. Springer. (cited on pages 19 and 23)

Golestan Far, M.; Sanner, S.; Bouadjenek, R.; Ferraro, G.; and Hawking, D., 2015. On term selection techniques for patent prior art search. In *Proceedings of the 38th international ACM SIGIR conference on Research & development in information retrieval*. ACM.

Graf, E.; Frommholz, I.; Lalmas, M.; and Van Rijsbergen, K., 2010. Knowledge modeling in prior art search. In *Advances in Multidisciplinary Retrieval*, 31–46. Springer. (cited on page 23)

Gurulingappa, H.; Müller, B.; Klinger, R.; Mevissen, H.-T.; Hofmann-Apitius, M.; Friedrich, C. M.; and Fluck, J., 2010. Prior art search in chemistry patents based on semantic concepts and co-citation analysis. In *TREC*. (cited on page 23)

Harris, C. G.; Arens, R.; and Srinivasan, P., 2010. Comparison of ipc and uspc classification systems in patent prior art searches. In *Proceedings of the 3rd international workshop on Patent Information Retrieval*, 27–32. ACM. (cited on pages 7 and 23)

Harris, C. G.; Arens, R.; and Srinivasan, P., 2011. Using classification code hierarchies for patent prior art searches. In *Current challenges in patent information retrieval*, 287–304. Springer. (cited on page 23)

Harris, C. G.; Foster, S.; Arens, R.; and Srinivasan, P., 2009. On the role of classification in patent invalidity searches. In *Proceedings of the 2nd international workshop on Patent information retrieval*, 29–32. ACM. (cited on page 23)

He, B. and Ounis, I., 2009. Finding good feedback documents. In *Proceedings of the 18th ACM conference on Information and knowledge management*, 2011–2014. ACM. (cited on page 14)

Hearst, M. A., 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23, 1 (1997), 33–64. (cited on page 21)

Herbert, B.; Szarvas, G.; and Gurevych, I., 2010. Prior art search using international patent classification codes and all-claims-queries. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, 452–459. Springer. (cited on page 23)

Itoh, H.; Mano, H.; and Ogawa, Y., 2003. Term distillation in patent retrieval. In *Proceedings of the ACL-2003 workshop on Patent corpus processing-Volume 20*, 41–45. Association for Computational Linguistics. (cited on page 18)

Jochim, C.; Lioma, C.; and Schütze, H., 2011. Expanding queries with term and phrase translations in patent retrieval. In *Multidisciplinary Information Retrieval*, 16–29. Springer. (cited on page 21)

Joho, H.; Azzopardi, L. A.; and Vanderbauwhede, W., 2010. A survey of patent users: an analysis of tasks, behavior, search functionality and system requirements. In *Proceedings of the third symposium on Information interaction in context*, 13–24. ACM. (cited on pages 1, 2, 23, and 27)

Kang, I.-S.; Na, S.-H.; Kim, J.; and Lee, J.-H., 2007. Cluster-based patent retrieval. *Information processing & management*, 43, 5 (2007), 1173–1182. (cited on page 23)

KIM, Y., 2014. *SEARCHING BASED ON QUERY DOCUMENTS*. Ph.D. thesis, University of Massachusetts Amherst. (cited on page 20)

Kim, Y. and Croft, W. B., 2014. Diversifying query suggestions based on query documents. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 891–894. ACM. (cited on page 20)

Kishida, K., 2003. Experiment on pseudo relevance feedback method using taylor formula at ntcir-3 patent retrieval task. In *Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering*. Citeseer. (cited on page 20)

Konishi, K., 2005. Query terms extraction from patent document for invalidity search. In *Proc. of NTCIR*, vol. 5. (cited on pages 19 and 20)

Kontostathis, A. and Kulp, S., 2008. The effect of normalization when recall really matters. In *IKE*, 96–101. (cited on page 17)

Kumaran, G. and Carvalho, V. R., 2009. Reducing long queries using query quality predictors. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 564–571. ACM. (cited on page 15)

Lease, M.; Allan, J.; and Croft, W. B., 2009. Regression rank: Learning to meet the opportunity of descriptive queries. In *Advances in Information Retrieval*, 90–101. Springer. (cited on page 15)

Lee, K. S.; Croft, W. B.; and Allan, J., 2008. A cluster-based resampling method for pseudo-relevance feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 235–242. ACM. (cited on page 20)

Lopez, P. and Romary, L., 2009. Multiple retrieval models and regression models for prior art search. *arXiv preprint arXiv:0908.4413*, (2009). (cited on pages 22 and 24)

Lopez, P. and Romary, L., 2010. Patatras: Retrieval model combination and regression models for prior art search. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, 430–437. Springer. (cited on pages 23 and 27)

Lopez, P.; Romary, L.; et al., 2010. Experiments with citation mining and key-term extraction for prior art search. In *CLEF 2010-Conference on Multilingual and Multimodal Information Access Evaluation*. (cited on pages 24 and 28)

Lupu, M.; Hanbury, A.; et al., 2013a. Patent retrieval. *Foundations and Trends in Information Retrieval*, 7, 1 (2013), 1–97. (cited on pages 2, 12, 22, and 23)

Lupu, M.; Piroi, F.; and Hanbury, A., 2013b. Evaluating flowchart recognition for patent retrieval. In *The Fifth International Workshop on Evaluating Information Access (EVIA)*, 37–44. (cited on page 23)

Magdy, W., 2012. *Toward higher effectiveness for recall-oriented information retrieval: A patent retrieval case study*. Ph.D. thesis, Dublin City University. (cited on pages xvii, 1, 5, 27, 29, 30, and 37)

Magdy, W. and Jones, G. J., 2010a. Applying the kiss principle for the clef-ip 2010 prior art candidate patent search task. (2010). (cited on page 23)

Magdy, W. and Jones, G. J., 2010b. Pres: a score metric for evaluating recall-oriented information retrieval applications. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 611–618. ACM. (cited on pages xvii and 25)

Magdy, W. and Jones, G. J., 2011. A study on query expansion methods for patent retrieval. In *Proceedings of the 4th workshop on Patent information retrieval*, 19–24. ACM. (cited on page 21)

Magdy, W. and Jones, G. J., 2013. Studying machine translation technologies for large-data clir tasks: a patent prior-art search case study. *Information Retrieval*, (2013), 1–28. (cited on page 24)

Magdy, W.; Leveling, J.; and Jones, G. J., 2010. Exploring structured documents and query formulation techniques for patent retrieval. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, 410–417. Springer. (cited on pages 19 and 20)

Magdy, W.; Lopez, P.; and Jones, G. J., 2011. Simple vs. sophisticated approaches for patent prior-art search. In *Advances in Information Retrieval*, 725–728. Springer. (cited on page 23)

Mahdabi, P.; Andersson, L.; Hanbury, A.; and Crestani, F., 2011a. Report on the clef-ip 2011 experiments: Exploring patent summarization. In *CLEF (Notebook Papers/Labs/Workshop)*. (cited on page 22)

Mahdabi, P.; Andersson, L.; Keikha, M.; and Crestani, F., 2012. Automatic refinement of patent queries using concept importance predictors. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, 505–514. ACM. (cited on page 21)

Mahdabi, P. and Crestani, F., 2012. Learning-based pseudo-relevance feedback for patent retrieval. In *Multidisciplinary Information Retrieval*, 1–11. Springer. (cited on page 21)

Mahdabi, P. and Crestani, F., 2014. Patent query formulation by synthesizing multiple sources of relevance evidence. *ACM Transactions on Information Systems*, (2014). (cited on page 2)

Mahdabi, P.; Gerani, S.; Huang, J. X.; and Crestani, F., 2013. Leveraging conceptual lexicon: query disambiguation using proximity information for patent retrieval. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, 113–122. ACM. (cited on page 20)

Mahdabi, P.; Keikha, M.; Gerani, S.; Landoni, M.; and Crestani, F., 2011b. Building queries for prior-art search. In *Multidisciplinary Information Retrieval*, 3–15. Springer. (cited on pages 18 and 19)

Manning, C. D.; Raghavan, P.; and Schütze, H., 2008. *Introduction to information retrieval*, vol. 1. Cambridge university press Cambridge. (cited on pages xvii, 10, 14, 15, 38, 45, and 54)

Mase, H.; Matsubayashi, T.; Ogawa, Y.; Iwayama, M.; and Oshio, T., 2005. Proposal of two-stage patent retrieval method considering the claim structure. *ACM Transactions on Asian Language Information Processing (TALIP)*, 4, 2 (2005), 190–206. (cited on pages 19 and 24)

Maxwell, K. T. and Croft, W. B., 2013. Compact query term selection using topically related text. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, 583–592. ACM. (cited on pages 46 and 54)

Miller, G. A.; Beckwith, R.; Fellbaum, C.; Gross, D.; and Miller, K. J., 1990. Introduction to wordnet: An on-line lexical database*. *International journal of lexicography*, 3, 4 (1990), 235–244. (cited on page 21)

Osborn, M.; Strzalkowski, T.; and Marinescu, M., 1997. Evaluating document retrieval in patent database: a preliminary report. In *Proceedings of the sixth international conference on Information and knowledge management*, 216–221. ACM. (cited on page 19)

Pérez-Iglesias, J.; Rodrigo, A.; and Fresno, V., 2010. Using bm25f and kld for patent retrieval. In *CLEF-2010 (Notebook Papers/LABs/Workshops)*. (cited on page 18)

Piroi, F., 2010. CLEF-IP 2010: Retrieval experiments in the intellectual property domain. In *CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy*. (cited on page 2)

Piroi, F.; Lupu, M.; Hanbury, A.; Sexton, A. P.; Magdy, W.; and Filippov, I. V., 2012. Clef-ip 2012: Retrieval experiments in the intellectual property domain. In *CLEF (Online Working Notes/Labs/Workshop)*, vol. 1178 of *CEUR Workshop Proceedings*. CEUR-WS.org. http://ceur-ws.org/Vol-1178. (cited on page 23)

Porter, M. F., 1980. An algorithm for suffix stripping. In *Program*, vol. 14, 130–137. (cited on page 27)

Robertson, S. E.; Walker, S.; Jones, S.; Hancock-Beaulieu, M.; and Gatford, M., 1993. Okapi at TREC-2. In *TREC*, 21–34. (cited on page 27)

Roda, G.; Tait, J.; Piroi, F.; and Zenz, V., 2010. Clef-ip 2009: retrieval experiments in the intellectual property domain. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, 385–409. Springer. (cited on pages 20, 23, and 37)

Sahlgren, M.; Hansen, P.; and Karlgren, J., 2002. English-japanese cross-lingual query expansion using random indexing of aligned bilingual text data. In *The Philosophical Writings of Gottlob Frege*. Citeseer. (cited on page 21)

Takaki, T.; Fujii, A.; and Ishikawa, T., 2004. Associative document retrieval by query subtopic analysis and its application to invalidity patent search. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, 399–405. ACM. (cited on pages 19 and 21)

Verma, M. and Varma, V., 2011a. Applying key phrase extraction to aid invalidity search. In *Proceedings of the 13th International Conference on Artificial Intelligence and Law*, 249–255. ACM. (cited on pages 19 and 23)

Verma, M. and Varma, V., 2011b. Exploring keyphrase extraction and ipc classification vectors for prior art search. In *CLEF (Notebook Papers/Labs/Workshop)*. (cited on page 24)

Xue, X. and Croft, W. B., 2009a. Automatic query generation for patent search. In *Proceedings of the 18th ACM conference on Information and knowledge management*, 2037–2040. ACM. (cited on page 19)

Xue, X. and Croft, W. B., 2009b. Transforming patents into prior-art queries. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 808–809. ACM. (cited on pages 2, 19, and 27)

Zhai, C. and Lafferty, J., 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, 22, 2 (2004), 179–214. (cited on pages 11 and 27)

Zhang, J. and Kamps, J., 2010. Search log analysis of user stereotypes, information seeking behavior, and contextual evaluation. In *Proceedings of the third symposium on Information interaction in context*, 245–254. ACM. (cited on page 2)