On Term Selection Techniques for Patent Prior Art Search

Mona Golestan Far

mona.golestanfar@anu.edu.au

A thesis submitted for the degree of
Master of Philosophy
Research School of Computer Science
The Australian National University

January 2016

Committee in charge:

Tom Gedeon, Chair

Professor of Computer Science
Research School of Computer Science
The Australian National University (ANU)

Scott Sanner, Main

Professor of Computer Science
College of Engineering, Electrical Engineering and Computer Science
Oregon State University
Adjunct Professor at ANU
Principal Researcher, Machine Learning Research Group, NICTA

Hanna Suominen

Adjunct Professor of Computer Science
Research School of Computer Science
The Australian National University
Senior Researcher, Machine Learning Research Group, NICTA

Gabriela Ferraro

Adjunct Professor of Computer Science
Research School of Computer Science
The Australian National University
Researcher, Machine Learning Research Group, NICTA

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

Mona Golestan Far

mona.golestanfar@anu.edu.au

January 26, 2016

Declaration

I hereby declare that this thesis is my original work that 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. Certain aspects of these contributions (mainly Chapter 4 of this thesis) have been published in collaboration with other researchers as follows:

> 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 and Development in Information Retrieval*, SIGIR '15 (Santiago, Chile, August 09 - 13 2015). ACM, New York, NY, USA.

I developed my baseline system based on Reda[1] Bouadjenek's baseline to investigate the reasons that his experiments on different query reformulation techniques ended in only little improvement over the baseline. My key research contributions consisting of experimental design and empirical analysis, which ended in proposing interactive methods as a promising avenue for simple but highly effective term selection in prior art search, have been presented in this thesis.

Mona Golestan Far
January 26, 2016

[1] Reda has been working on patent prior art search for three months as a visitor at NICTA before I start. He also published his results in [Bouadjenek et al., 2015]

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

Acknowledgments

Upon the accomplishment of this thesis, I would like to thank people who influenced and enlightened my academic path. First, I owe a debt of 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. He highly encouraged me (1) to work hard, but with enthusiasm, (2) to feel responsibility for honest contribution to science, (3) to be an independent and efficient thinker, (4) to deal with a research problem by proposing to-the-point research questions, (5) to be brave enough to examine new ideas and to seek for the best solution, and (6) to collaborate with other researchers and to be generous in sharing the success of my work with others. I hope I have learned enough to continue my academic career even without his unselfish support. In my view, the science world needs more people like Scott.

I thank my other committee members, Tom Gedeon (chair of my panel), Gabriela Ferraro, and Hanna Suominen for their supports over the course of my MPhil. I had always Tom's support. Gabriela has been for me when I needed and Hanna kindly provided a thorough review of my thesis though she was on her maternity leave during my research. I am indebted to them all for their time and effort they spent.

I acknowledge the academic and technical support provided by the National ICT of Australia (NICTA) and the Australian National University (ANU) and thank them for their support in my research. Specially, I would like express my gratitude to Bob Williams — the leader of the Machine Learning (ML) group of NICTA. He always encouraged students, including me, to participate in NICTA events (e.g., NICTA ML Retreat and Review); he also supported me to attend NICTA ML Summer School. He created such a positive research atmosphere and culture with a collection of excellent researchers that made my tenure at NICTA an extraordinary fruitful experience. I learned a lot discussing with smart researchers on my neighbourhood like Scott Sanner, Justin Domke, and Aditya Menon and attending their seminars, talks, and reading groups.

During my one-year MPhil program, I met many successful IR researchers like Paul Thomas (and his great IR & friends seminars), Milad Shokuhi, and Leif Azzopardi who left a strong impact on my career. I would also thank David Hawking for his smart comments on my work. Dave and Scott's support encouraged me to have a submission to SIGIR that got accepted ◡̈.

I would like also mention about anonymous reviewers of my SIGIR paper for their useful comments. Their words warm my heart to stay in research and have more contribution to Computer Science and IR community. I always feel confident to go forward when I read their following comment on my paper: "Although the

amount of work on patent query reformulation was large since 2010, but none of this work did that simple but interesting study. The study and results are very interesting and can be very useful for patent examiners in practical search situations". I owe the presentation of my work at Chile (Santiago) to both generous SIGIR travel grant and ANU fund.

I appreciate Reda Bouadjenek, for his contribution to the baseline IR framework. I would like also thank my best fellow postgraduate friend, Ehsan Abbasnejad, for patiently answering my questions and spending time to discuss over new ideas.

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 being always the main support after my parents and for respecting my attempts to follow my goals and values. I owe a debt of love to my parents who have been the main motivation behind every single step in my academic life by their famous recommendation: "Go forward for the highest possible academic degree and never stop learning or experiencing a new path".

Abstract

A patent is a set of exclusive rights granted to an inventor to protect his 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 that may be relevant to a new patent application. Many well-known information retrieval (IR) techniques (e.g., typical query expansion methods), which are proven effective for ad hoc search, 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 (i.e., 0.11 vs. 0.48) and performs twice as well on mean average precision (MAP) as the best participant in CLEF-IP 2010 (i.e., 0.22 vs. 0.48). We find a very clear term selection value threshold for use when choosing terms. We also notice that most of the useful feedback terms are actually present in the original query and hypothesise that the baseline system can be substantially improved by removing negative query terms. We try four simple automated approaches to identify negative terms for query reduction but we are 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

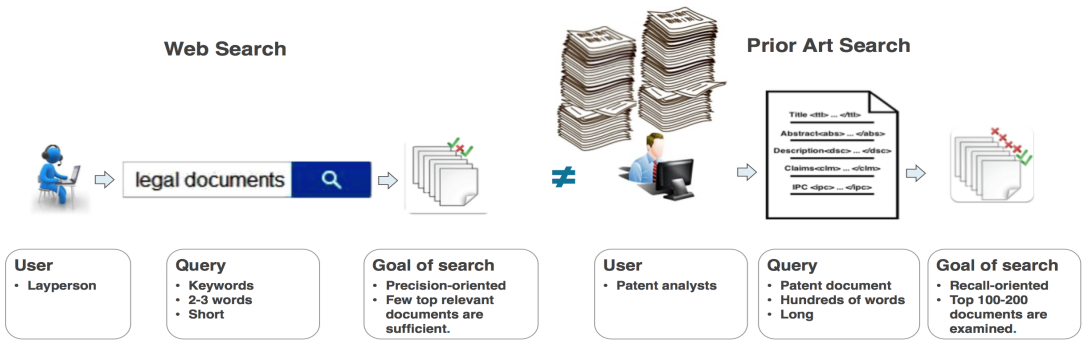


Figure 1.1: The main differences between patent prior art search and a standard web search are: (i) queries are reference patent applications, and (ii) patent prior art search is a recall-oriented task.

1.1 Motivation

Patents are used by legal entities to protect their inventions and they are considered a multi-billion dollar industry of licensing and litigation. Given that a single existing patent may invalidate a new patent application, an efficient patent retrieval system is an important research topic. Patent prior art search involves finding previously granted patents, scientific articles, product descriptions or other published work 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]. Figure 1.1 illustrates the main differences between patent prior art search and a standard web search. 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. 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 typically 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 and high costs involved in creating a patent and infringing patented material [Joho et al., 2010]. 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 complicates the task of formulating queries.

The main users of patent prior art search are patent analysts[1], who are employed to determine the patentability of applications. Patent searchers have to perform an exhaustive and comprehensive search. In general, patent examiners spend about 12 hours to complete an invalidity task by examining approximately 100 patent documents retrieved by 15 different queries on average [Joho et al., 2010]. In addition, the number of patent applications and granted patents is rapidly increasing in recent years. For example, 326,033 patent applications were approved in the US alone[2] in 2014; a number that has doubled in the past 15 years.

Searching based on queries made up of patent applications (patent queries) helps patent examiners to save time and avoid formulating appropriate search queries out of long and difficult patent applications. However, this approach is less effective than a typical web search [Lupu et al., 2013a]. In this thesis, we study query reformulation to transform an initial query (i.e., a 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 [Xue and Croft, 2009b; Piroi, 2010b; Mahdabi and Crestani, 2014]. While prior work has largely focused on specific techniques for query reformulation, we first build an oracular query formed from known relevance judgements for the CLEP-IP 2010 prior art test collection [Piroi, 2010b] (Section 3.1) 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 oracular query evaluation suggest that query reduction methods can outperform state-of-the-art prior art search performance, we proceed to analysing 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 evalu-

[1] Inventors, who want to determine whether their ideas are novel, are not considered as users of patent prior art search, because the prior art search starts with querying by a patent application document; this document is not available when the author is going to ensure the novelty of his idea.

[2] http://www.uspto.gov/web/offices/ac/ido/oeip/taf/us_stat.htm [Accessed on 29/12/2015]

ate 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[3], 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 summarised as follows:

First, we performed a novel analysis in patent IR indicating how the identification of an initial set of relevant documents can improve the retrieval effectiveness of patent prior art search through query formulation and query reduction. We developed an oracular relevance feedback system which extracts terms from the judged relevant documents to formulate oracular queries to determine upper bound performance. Experiments related to oracular system suggest the necessity of precise query reduction and term selection techniques to improve the effectiveness of patent prior art search.

Second, based on our initial analysis, we tried different query reformulation methods to automatically improve the original patent query and approximate the oracular query. However, none of the proposed methods showed significant improvement over the baseline, which aligns with most of the reported literature in patent search. We analysed 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 a noticeable improvement in performance via automated methods.

Third, we proposed labelling the initial set of retrieved results until a relevant document is found that can be used to improve the results of the baseline in a superior way and it outperforms one of the best performing system in patent search (e.g., PATATRAS). We showed this simple interactive relevance feedback approach is also a minimum burden on the users because the first relevant patent can be found 80% of the times in the first 10 results of an initial run.

1.4 Thesis Outline

This chapter introduced our research problem: techniques to reformulate the optimal query for the patent prior art search. The rest of the thesis is organised as the following. Chapter 2 defines all background material required to understand both generic and patent-specific information retrieval (IR). It also reviews previous

[3]Cross-language Evaluation Forum, Intellectual Property Lab

work from a number of related research areas, mainly focusing on the existing query reformulation techniques for both generic and patent-specific IR.

Our main experiments are described in Chapters 3 and 4. In Chapter 3, we first explain the baseline system and related experimental settings; then we briefly describe the test collection used in our experiments. We also discuss errors caused by the data curation and the use of IPC[4] filter within the retrieval process. Chapter 4 covers a thorough analysis on terms in query and top 100 retrieved documents, where we determine the key causes of low effectiveness in prior art search. It also contains our proposed methods to reformulate the query that get improved over the baseline and the best performing patent retrieval system. Finally, Chapter 5 concludes this thesis by summarising the results and discussing interesting directions of future work.

[4]International Patent Classification

Chapter 2

Background and Related Work

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

2.1 Structure of Patents

A patent is a structured document, which consists of four main sections: title, abstract, description, and claims. In addition, it usually contain tables, mathematical and chemical formulas, citations, and technical drawings. The text of a patent document is saved electronically in the patent office as an XML[1] 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].

The structure of patents varies according to the patent office to which the invention is filed. United States patent and trademark office[2] (USPTO), the European patent office[3] (EPO), and the Japan patent office[4] (JPO) are examples of patent offices around the world. There are common sections, which are found in most patent documents. For instance, each patent should contain at least one claim. Other sections such as the "title" of the invention, "classification", "description", "abstract", "summary of invention" may or may not exist according to the patent office. It is very common to find patents filed to the USPTO containing the "abstract" field, but it is not very common in the EPO. Another example of inconsistent use of fields between different patent offices is the presence of explicit fields in USPTO patents called "summary of the invention" and "field of the invention". These two fields are not common in the other patent offices.

In our experiments, we use CLEF-IP 2010 data collection (Section 3.1), which contains patent documents derived from EPO. Hence, we will briefly explain the main sections and some meta-data of EPO patents that are commonly used in a patent retrieval system as follows:

[1] Extensible Markup Language
[2] http://www.uspto.gov
[3] http://www.epo.org
[4] http://www.jpo.go.jp/

```
<?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.

(1) Title
The title of the patent appears in three languages. This is a feature in EPO patents, where the title is stated in English, French, and German.

(2) Abstract
The abstract of the patent document 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.

(3) 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 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.

(4) Claims
The claims section of the patent document lists the 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 generalise 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 the language of the document. Claims translations are only provided for a granted patent.

Nonetheless, patents may contain tables, mathematical and chemical formulas, citations, technical drawing, meta-data (e.g., applicant, inventor, IPC codes, and publication date) or other additional material that can be used to improve the retrieval effectiveness. As it will be described in Section 2.3.5, IPC codes and citations has been widely applied in patent retrieval. Hence, we describe them next.

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

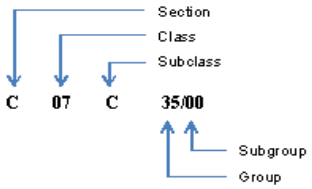


Figure 2.2: An example illustrating the main components of an International Patent Classification code.

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, labelled 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 instance, 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, the 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 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 the Main Code. They may also be assigned further classification and indexing terms to give more details of the contents of the invention, which are called Further Codes.

(6) 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. These citations can be also 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; the final version of the patent gets these parts modified or removed.

2.2 Generic Information Retrieval

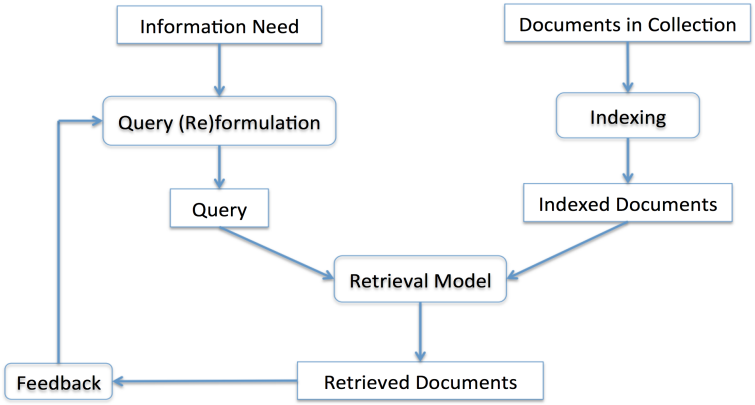


Figure 2.3: Illustration of the process in a generic *IR* system.

An information retrieval (IR) system assists users in finding the information they need. Figure 2.3 illustrates the general IR process. On the collection side, 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 using a retrieval model 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 [Manning et al., 2008, pp. 3–10]. 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.

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 an appropriate ranking algorithm to return relevant documents at top of the ordered list, leading to high effectiveness. Three well-known families of retrieval models are [Croft et al., 2010, p. 233]: (1) vector space models [Salton et al., 1975] (e.g., term frequency and inverse document frequency (TF-IDF)), (2) probabilistic models [Robertson and Zaragoza, 2009] (e.g., BM25[5]), and (3) Language Models (LM) [Ponte and Croft, 1998].

The Vector Space Model

In a vector space model, documents and queries are represented by vectors of term

[5]BM stands for Best Match, and 25 is just a numbering scheme used by Robertson and Walker [1994] to keep track of weighting variants [Croft et al., 2010, p. 249].

weights, and the collection is represented by a matrix of term weights as follows:

$$D_i = [d_{i1}, d_{i2}, d_{i3}, \ldots, d_{im}],$$

$$Q = [q_1, q_2, q_3, \ldots, q_m],$$

$$C = \begin{bmatrix} d_{11} & d_{12} & d_{13} & \cdots & d_{1m} \\ d_{21} & d_{22} & d_{23} & \cdots & d_{2m} \\ d_{31} & d_{32} & d_{33} & \cdots & d_{3m} \\ \vdots & & & & \\ d_{N1} & d_{N2} & d_{N3} & \cdots & d_{Nm} \end{bmatrix},$$

where D_i is a document in the collection C, d_{ik} is a weight for each term t_k in the document D_i, and q_k[6] represents a term in the query Q. The collection is represented by the matrix C_{Nm}, where N is the number of documents in the collection and m is the number of unique terms in the collection. If a term does not appear in a document, the weight for that particular term will be zero.

The TF-IDF weighting function multiplies the occurrence of each term in the document ($c(t_k, Di)$) by the inverse document frequency (idf) measure. idf measures the importance of a term in the collection:

$$idf(t_k) = \log \frac{N+1}{df(t_k)}, \tag{2.1}$$

where $df(t_k)$ is the number of documents in the collection, which contain at least one occurrence of the term t_k, and N is the number of documents in the collection.

Given a query Q, documents are ranked based on the overlap score measure. The TF-IDF score of a document D is the sum, over all query terms, of the TF-IDF weight of each query term q in D. After pivoted normalisation, the TF-IDF score for each document is calculated as follows [Bache and Azzopardi, 2010]:

$$TFIDF(Q, D) = \sum_{q \in Q \cap D} \frac{c(q, D) \times idf(q)}{(1-b) + b.\frac{|D|}{avdl}}, \tag{2.2}$$

where $|D|$ is the size of the document (i.e, the number of words) and $avdl$ is the average document length, $c(q, D)$ is the number of occurrence of each query term in the document D, and $idf(q)$ is the importance of each query term in the collection. TF-IDF 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 as will be described below.

Probabilistic Models

BM25 is a popular and effective ranking algorithm, which extends the scoring function for the binary independence model [Manning et al., 2008, p. 232] to include

[6] We ignore indices to simplify the further equations in this thesis.

document and query term weights. Each document is scored based on the BM25 weighting scheme — often called the Okapi weighting — as follows:

$$BM25(Q,D) = \sum_{q \in Q \cap D} \left(idf(q) \right) \left(\frac{(k_1+1)c(q,D)}{k_1((1-b)+b.\frac{|D|}{avdl}) + c(q,D)} \right). \tag{2.3}$$

The variable k_1 is a positive (i.e., > 0) tuning parameter that calibrates the document term frequency scaling. The setting of $k_1 = 0$ corresponds to a binary model (i.e., no term frequency), and setting a large value for k_1 corresponds to using raw term frequency. The parameter b is also used for tuning ($0 \leq b \leq 1$) 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 normalisation.

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. In this case, BM25 weighting function is calculated as follows:

$$BM25(Q,D) = \sum_{q \in Q \cap D} \left(idf(q) \right) \left(\frac{(k_1+1)c(q,D)}{k_1((1-b)+b\frac{|D|}{avdl}) + c(q,D)} \right) \left(\frac{(k_3+1)c(q,Q)}{k_3 + c(q,Q)} \right), \tag{2.4}$$

where $c(q,Q)$ is 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 normalisation of queries because retrieval is being done with respect to a single fixed query. The tuning parameters of these equations should ideally be set to optimise performance on a development test collection. That is, we can search for values of these parameters that maximise performance on a separate development test collection (either manually or with optimisation methods [Metzler, 2007] such as grid search or something more advanced), and then use these parameters on the actual test collection. In the absence of such optimisation, experiments have shown reasonable values are to set k_1 and k_3 to a value between 1.2 and 2, and b = 0.75 [Manning et al., 2008, p.233][7].

Language Models with Terms Smoothing

The basic idea behind the LM 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 [Zhai and Lafferty, 2004]. Here terms are assumed to occur independently, and the probability is the product of the probability of individual query term q given the document model M_D of document D as follows:

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

[7] These parameters have been derived using small newswire collections and may not be adequate for patent domain (due to the large difference in document length between news and patents).

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

where $c(q,D)$ represents the frequency of term q in document D, and $|D|$ is the size of the document (i.e, the number of words). The overall similarity score for the query and the document could be zero if some 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: The Jelinek-Mercer smoothing language model [Zhai and Lafferty, 2004] combines the relative frequency of a query term 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 $P(q|C)$ as follows:

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

where λ is a tuning parameter controlling the probability assigned to unseen words. The optimal value of the parameter 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. Hence, we have

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

where μ is a tuning parameter. The formula assigns lower score to documents that contain the term, but with fewer occurrence than predicted by the collection language model. As the tuning parameter 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 $2,000$ [Zhai and Lafferty, 2004]. 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 Retrievability

In a complex search like patent (recall oriented) retrieval, it is important that all relevant documents are potentially retrievable by correct query terms. Hence, we briefly explain the study of retrievability in this section. 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 c ranked results via any query terms that they are relevant for [Lupu et al., 2013a]. A document, which is difficult or impossible to be retrieved by a particular retrieval model, would not be retrieved when it is relevant; this leads to a low recall.

Essentially, it is desirable that the retrieval system considers 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 collection 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 score higher.

2. Length normalisation: incorporative term frequency into a model makes 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 TF-IDF and Opaki BM25 favour large terms frequencies [Bashir and Rauber, 2011].

Retrievability Measurement

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

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

where k_{dq} is the rank of the document d in the result set of the query $q \in Q$ and 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. The 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 summarise the amount of bias in the Lorenz Curve and it is computed as follows:

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

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 collection [Bashir and Rauber, 2011].

2.2.3 Query Expansion

One solution to the significant term mismatch between the query and the relevant documents is query expansion (QE) [Efthimiadis, 1996], 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 or from external sources such as Wikipedia, or dictionaries [Cao et al., 2008]. Original queries should be expanded by good terms, unless it can lead to retrieval of irrelevant documents.

Feedback-based Query Expansion

An initial query can be expanded using a feedback from users — explicit relevance feedback — or automatically from top k ranked retrieved documents, assuming they are relevant to the query — pseudo relevance feedback (PRF) [Manning et al., 2008, p.187]. Getting feedback from users needs user studies and interaction while pseudo relevance feedback is an automated process without user interaction.

The Rocchio Algorithm for Relevance Feedback: The Rocchio algorithm [Rocchio, 1971] is a classic algorithm of relevance feedback used mainly for query expansion. In brief, it provides a method of incorporating relevance feedback information into the vector space model representing a query [Manning et al., 2008, p.181]. The Rocchio algorithm is used to modify the query by the partial knowledge of known relevant and irrelevant [8] documents; the goal is to move the query closer to the centroid of the relevant documents but further from irrelevant documents (Figure 2.4). The modified

[8] In this thesis, we use irrelevant and non-relevant interchangeably for documents that are not relevant to the query.

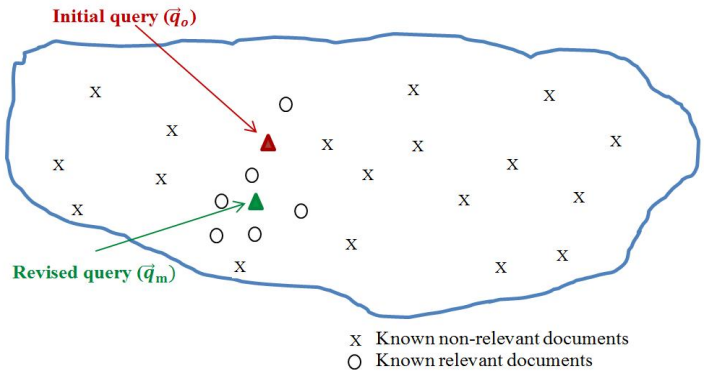


Figure 2.4: Rocchio algorithm for relevance feedback. Some documents have been labelled as relevant and irrelevant and the initial query vector is moved in response to this feedback (redrawn from [Manning et al., 2008, p.182]).

query $\vec{q}_m$ is:

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

where q_0 is the original query vector; D_r and D_{irr} are the set of known relevant and irrelevant 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 higher β and γ [Manning et al., 2008, p.183].

PRF is typically used in Rocchio algorithm. In automatic query expansion, we can include all candidate terms (e.g., all the terms in the known relevant documents) and use the term weighting schema to emphasise on the fact that some terms are better than others or we can never include poor terms [Robertson, 1991]. It is important to distinguish between good expansion terms and bad ones. [Carpineto and Romano, 1999] used term scoring methods based on the differences between the distribution of terms in (pseudo-)relevant documents and the distribution of terms in all documents. Distinguishing between expansion terms only based on their distribution in the feedback documents (i.e., extracting the most frequent terms) and in the whole collection (i.e., extracting the most specific terms) is not sufficient. It can be considered as a term classification problem to separate good expansion terms from others directly according to their potential impact on the retrieval effectiveness; hence, we can apply supervised learning methods for term selection. Classifiers like Support Vector Machines (SVM) [Cao et al., 2008], Naïve Bayes and Logistic Regression [He and Ounis, 2009] can be used to classify terms and feedback documents.

Query Expansion by External Resources

The most common form of query expansion is a global analysis, using dictionaries, WordNet, Wikipedia, or other thesaurus. 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, p.190].

2.2.4 Query Reduction

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 (QR), a technique for dropping unnecessary query terms from long queries, improves the performance.

A common approach to reduce verbose queries is selecting a subset of a long query (or sub-query). A search engine performs more precisely when just the key concepts are used as a query rather than a long query. Hence, the identification of the key query concepts has a 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]. We can choose effective subsets in a query by 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 SVM classification is used [Kumaran and Carvalho, 2009]. In this approach, the quality of query reduction depends on the performance of the predictor and ranking algorithm [Balasubramanian et al., 2010].

As the other approach, we can use query term ranking techniques to select effective terms from a verbose query by ranking them. A vast number of rankings are possible given different settings of individual term weights; for example, we can train a regression model to weight all query words of a verbose query [Lease et al., 2009]. We can also 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. Main IR evaluation metrics are calculated using a contingency table (Table 2.1), where:

True Positive (TP): The number of documents, which are relevant and the system retrieves them.
False Negative (FN): The number of documents, which are relevant but the system does not retrieve them.
False Positive (FP): The number of documents, which are irrelevant but the system retrieves them.

	Relevant	Irrelevant
Retrieved	True Positive (TP)	False Positive (FP)
Not-retrieved	False Negative (FN)	True Negative (TN)

Table 2.1: Contingency table.

True Negative (TN): The number of documents, which are irrelevant and the system does not retrieve them.

Precision and Recall

Precision and recall are the most frequent and basic measures for information retrieval effectiveness [Manning et al., 2008, p.155]. They are calculated with respect to the documents are returned by the search engine in response to a specific query. Precision is the fraction of retrieved documents that are relevant:

$$Precision = \frac{\sharp(relevant\ items\ retrieved)}{\sharp(retrieved\ items)} = \frac{TP}{TP+FP} = P(relevant|retrieved),$$

where the symbol $\sharp$ is read as 'the number of'. Recall is the fraction of relevant documents that are retrieved:

$$Recall = \frac{\sharp(relevant\ items\ retrieved)}{\sharp(relevant\ items)} = \frac{TP}{TP+FN} = P(retrieved|relevant).$$

For many prominent applications, particularly web search, good results on the first page or the first three pages are more important than all relevant documents. Hence, users prefer to look at precision and recall 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" [Manning et al., 2008].

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

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

Average Precision and Mean Average Precision

We can measure mean average precision (MAP) by calculating average precision (AP) on retrieval results. AP is the average of precision at each point where a relevant document is found; it is computed as:

$$AP = \frac{\sum_{r=1}^{N}(Prec(r) \times R(r))}{n}, \tag{2.13}$$

where r is the rank, N the number of documents retrieved, $R(r)$ a binary function of the document relevance at a given rank, $Prec(r)$ is precision at a given cut-off rank r, and n is the total number of relevant documents [Manning et al., 2008, p.159]. Then, for a given set of queries, Q, MAP is calculated by:

$$MAP(Q) = \frac{\sum_{q \in Q} AP(q)}{|Q|}, \tag{2.14}$$

where q is a query in a set of queries Q.

2.3 Patent-specific Information Retrieval

Patent retrieval differs from a generic retrieval due to specific characteristics of patents. For instance, in patent prior art search the query is as long as a patent document, which makes the query (re)formulation more difficult compared to a web search. Patents have some specific features, which can be yielded in retrieval; we will discuss them in this section.

2.3.1 The Study of Retrievability for Patents

Retrievability is specifically critical in recall oriented applications like patent retrieval or legal settings. In these cases, the focus of a system is to retrieve all documents that are relevant rather than to retrieve a subset of documents that best satisfy the query intent. Hence, all documents should at least potentially be retrievable via appropriate query terms. The design of recall oriented retrieval systems has been the centre of attention recently [Fujii et al., 2007; Kontostathis and Kulp, 2008]. 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.

In retrievability, we analyse 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. Experiments show that about 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 are identified using cluster-based [Bashir and Rauber, 2009b] or term-proximity-based methods [Bashir and Rauber, 2010].

Another study [Bache and Azzopardi, 2010] analyses the relationship between retrievability and effectiveness-based measures (Precision, MAP). Results show that the two goals of maximising access and maximising performance are quite compatible. They further conclude that a reasonably good retrieval performance can be

obtained by selecting parameters that maximise 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 a retrievability-based measure without accessing to relevance judgments, making it an attractive alternative for automatic evaluation.

2.3.2 Initial Query Formulation

The patent prior art search begins with a full patent application as a query. A full text as a query is a big challenge because it is not focused on the information that the user needs. In order to achieve good retrieval results, extracting the best representative text with the proper weights is important. In this section, we discuss initial attempts to formulate an effective query out of the patent query as follows:

The Best Field to Extract Query Terms
Patents are structured documents and they consist of several different sections: title, abstract, description, and claims. Different sections use different type of language for describing the invention. The abstract and description use more technical terminology while the claim usually uses legal jargon. Structured indexing keeps the field structure in the index that allows searching in specific fields instead of searching in a full document. In addition, separate fields for meta-data (Section 2.3.5) like IPC code and author can be used to improve the retrieval effectiveness [Magdy et al., 2009].

There are contrasting findings from previous work with respect to which field should be used to extract query terms. Early patent search tasks mainly considered claims to build the query, the same as what examiners start in the novelty process [Takaki et al., 2004; Konishi, 2005; Mase et al., 2005; Fujii, 2007a], whereas recent works show that building queries from the description field is more useful in patent retrieval (considering background summary in US patents equivalent to description field in European patents) [Xue and Croft, 2009a,b; Mahdabi et al., 2011b]. Another research shows that extracting terms according to their TF-IDF scores from every field of the query patent, and giving higher importance to the terms extracted from the abstract, claims, and description fields than to the terms extracted from the title field, is an effective way of constructing a search query [Cetintas and Si, 2012]. Another experiment shows that discarding descriptions from queries improves the MAP up to 30% because descriptions contain more noise than information [Gobeill et al., 2010]. They also show that claims are more informative and the title is poorly informative in retrieval.

Query Formulation Using Phrases
Most query formulation techniques rely on terms; however, formulating queries using phrases has recently obtained encouraging results [Becks et al., 2010]. According to early results, an NLP[9]-based grouping of terms can increase performance compared to the bag-of-words approach [Osborn et al., 1997]. For example, we can

[9]Natural language processing

improve the 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) algorithm 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
Kim and Croft [2014] recently worked on generating diverse queries from the patent query 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 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 set of terms represents one query aspect, then the top k retrieved documents for each sub-query are considered as pseudo relevant documents and those ranked below the top k are considered as irrelevant documents. Finally, the query is generated by a decision tree [KIM, 2014; Kim and Croft, 2014].

2.3.3 Query Expansion for Patents

Although a query is very long in patent prior art search, a significant term mismatch between queries and relevant documents has been reported earlier [Roda et al., 2009; Magdy et al., 2009]. The QE is a suggested solution to cope with the term mismatch problem; however, most QE techniques are ineffective to improve the performance within the patent domain [Kishida, 2002; Konishi, 2005]. We review previous works on the QE techniques for the patent search here.

Query Expansion by Pseudo Relevance Feedback
Previous studies [Magdy and Jones, 2011] showed that PRF is ineffective for patent prior art search. Since the retrieval effectiveness is low at initial retrieval, the assumption that the top k documents are relevant is invalid and leads in adding noise to the query; hence, the improvement using PRF is insignificant. The solutions proposed to cope with this problem are as follows:

- **Selecting documents for PRF based on cluster analysis:** In this approach, a document that clusters lots of high similar documents is considered as relevant and a document that has no nearest neighbour or some neighbours with low similarity is considered irrelevant [Lee et al., 2008]. In the 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 success of this approach highly depends on the appropriate selection of terms out of the patent query; these terms produce the best PRF candidates that can help in improving retrievability during QE [Bashir and Rauber, 2010]. Experiments show a significant improvement for Gini coefficient, which is used to measure retrievability, but there is no report on other main measures (e.g., MAP and recall).

- **Identyfying expansion terms:**
  Term proximity information can be used to identify expansion terms. Given a patent query, first, an initial query is generated by taking, for example, claim terms; then a query-specific lexicon based on definitions of the IPC is created. Among many terms in the lexicon, only expansion terms identified by two adjacency operators used in patent examination[10] (i.e., 'ADJn' and 'NEARn') are used for query expansion [Mahdabi et al., 2013].

- **Predicting the effectiveness of feedback documents:**
  In patent retrieval, the MAP is very low at initial retrieval; hence the top retrieved documents are not always relevant. As a result, there is a high chance that we use irrelevant documents for expansion in PRF. Recently, machine learning methods like regression are used to improve the PRF by predicting the effectiveness of a feedback document [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 techniques 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, show insignificant change to overall retrieval effectiveness, but a degree of improvement for some topics in patent domain. Magdy and Jones [2011] applied the idea of automatically generating the synonyms set (SynSet) using parallel manual translations to create possible synonyms sets (in the CLEF-IP collection, some sections in some patents are translated into three languages: English, French, and German). Although this idea presents better results compared to WordNet, there is still little improvement in retrieval effectiveness. The only QE technique, which achieves the best results, uses a combination of PRF and QE with translation of terms and phrases from German and French [Jochim et al., 2011].

[10]Patent examiners use term proximity heuristics in their searches in Boolean retrieval model in order to reward a document where the matched query terms occur close to each other. Two forms of adjacency operators are used in Boolean retrieval model to address proximity. The 'ADJn' operator which searches for terms within n words proximity in the order specified, and the 'NEARn' operator, which searches for the terms within n words, in either order.

2.3.4 Query Reduction for Patents

In patent prior art search, a query is as long as a patent document; verbose patent queries are not focused on information needed by the user and they may cover more than one topic. Query reduction is a good solution for this problem. We discuss query reduction techniques for patent retrieval as follows:

Query Segmentation:
Decomposing each patent query into coherent sub-topics 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]. In another approach, PRF is adapted for query reduction by decomposing a patent application into constituent text segments and the LM similarities are computed by calculating the probability of generating each segment from the top ranked documents. The least similar segments to the query are removed, hypothesising that the 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].

Query Summarisation:
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. The LM for the query, collection, and each summary is generated after summarising the patent query [Mahdabi et al., 2011a].

Query Term Selection:
Identifying useful query terms and giving them higher weights is important to build an effective query. The simplest proposed approach is 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 this method is that it does 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) [Kullback and Leibler,

1951] between the language model of the query and the whole collection as follows:

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

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 a 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] as follows:

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

where $c(t,Q)$ 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 a smoothing parameter (based on JM LM).

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; however, patents contain additional material, namely, tables, mathematical and chemical formulas, citations, technical drawing, meta-data (e.g., applicant, inventor, IPC codes, and publication date) that can be used to improve the retrieval. We explain the non-text information (e.g., meta-data) in patents that are used to improve the retrieval performance as follows:

The Use of Citation

The most successful use of meta-data to date is the citation lists in order to learn patterns of relevance [Lupu et al., 2013a]. A patent collection is a very dense network of citations that creates a set of interrelations and can be exploited during a prior art search. The large majority of patents rely upon the previous work and patents. The citation relations make this development process visible. Similarly, fundamental patents, which establish a new technology, are exceptional and they tend to be cited very frequently in the whole sub-field. A citation graph of a patent collection is used for identifying patent thickets (i.e., the patent portfolios of several companies that overlap 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, 2009a].

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 the 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 are combined to compute the ranking score for the 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 is used 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 trivial improvements [Lopez and Romary, 2009a]. However, citation information is not always present in the patent application and this method cannot be used in real-life patent search and initial citations by the applicants may not be considered relevant by patent examiners [Magdy and Jones, 2010b; Magdy et al., 2011]. Similar work by Gobeill et al. [2010] and Gurulingappa et al. [2010] also indicate improvement in MAP and recall using citations in patent retrieval.

The Use of IPC Codes

Patents are classified by 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 addition, classifications 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, 2009b]. 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 work, considered IPC code in their search, reported improvement in retrieval effectiveness [Fujita, 2005; Kang et al., 2007; Herbert et al., 2009; Graf et al., 2010; Harris et al., 2009, 2010, 2011; Verma and Varma, 2011a]. It has also reported that using complete IPC code leads to better results compared to the use of 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 on information present in images (flowcharts and diagrams), so, beyond text data, image processing tasks also can contribute to the search [Lupu et al., 2013a]. A 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 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 [Roda et al., 2009; Joho et al., 2010; Piroi et al., 2012; Lupu et al., 2013a]. In recent years machine translation (MT) has become established as the dominant technique for translation in CLIR, which usually achieves 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, 2014].

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, 2009a]. Many patent retrieval tasks re-rank the top retrieved documents obtained from an initial retrieval stage based on additional patent features [Lopez et al., 2010], claim structure [Mase et al., 2005], and considering IPC information of patents and their 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 — as patent prior art search — 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

Patent retrieval evaluation score (PRES) [Magdy and Jones, 2010a] is a novel metric for evaluating recall-oriented IR applications, which is derived from the normalised

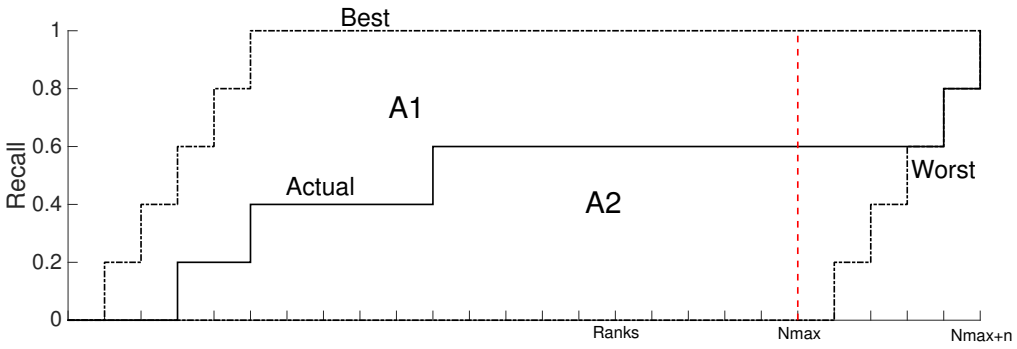


Figure 2.5: PRES curve is bounded between the best case and the new defined worst case (redrawn from [Magdy and Jones, 2010a]).

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 a 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 $N_{\max}$. Any relevant document not retrieved in the top $N_{\max}$ is assumed to be the worst case (Figure 2.5). 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 a higher recall can achieve a lower PRES value when compared to systems with a lower recall but a 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.

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

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.

2.4 Summary

This section covered key background material required to understand both generic and patent-specific IR. It also included literature reviews over previous work from a number of related research areas, mainly focusing on the existing query reformulation techniques for both generic and patent-specific IR.

In nutshell, as we discussed in this section, patent retrieval has studied quite well in previous work because it is an important problem. The number of work and techniques — mostly complicated — on patent retrieval, specially for QE techniques, is large; however, none reported a significant improvement. Hence, we can see the necessity of a precise error analysis showing the reasons that these techniques fail for patent prior art search. Next two chapters report the results of our experimental design and empirical analysis of our baseline patent retrieval system, which make it clear why patent prior art is ineffective and why standard IR techniques by previous studies cannot improve it.

Chapter 3

Baseline IR Framework

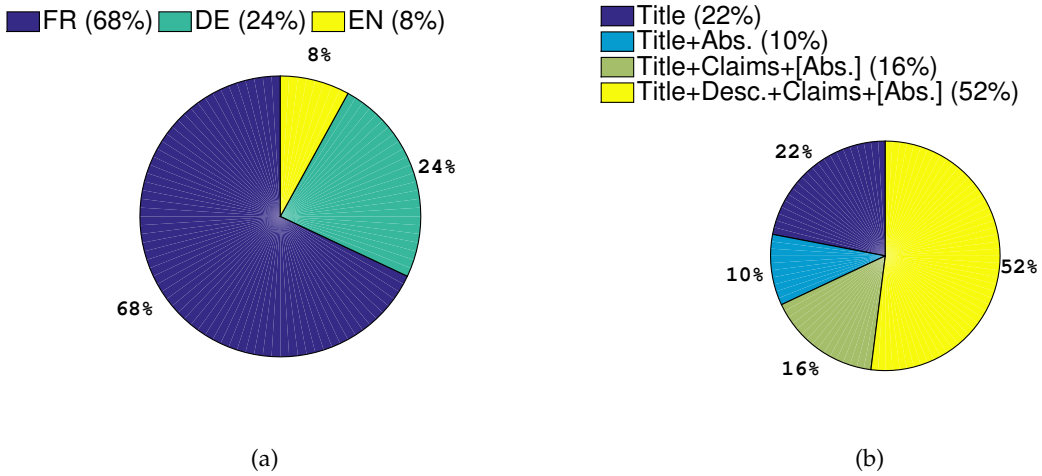
In this chapter, first, we describe the data collection used in our experiments (i.e., CLEF-IP 2010); then we briefly explain our baseline system and the experimental settings. In addition, we analyse two main errors caused by the data curation and our experimental settings.

3.1 Data Collection

The prior art candidate search task (PAC) ran in three subsequent years: 2009, 2010, and 2011 by Cross Language Evaluation Forum for Intellectual Property evaluation track[1] (CLEF-IP). We use CLEF-IP 2010 data collection[2]. The documents in the patent collection are stored as XML files. The documents are derived from data released by the European Patent Office (EPO) and have mixed content in English, German and French. The files contain bibliographic data as well as descriptive text. The XML files are quite comprehensive, containing detailed information on inventors, assignees, priority dates and so on. CLEF-IP 2010 contains 2,6 million patent documents, corresponding to approximately 1,3 million individual patents published until 2001. The prior art candidate search task contain 2,000 topics. The evaluations can be also done on topic subsets where the topic document language is English, German, or French, respectively. These topic subsets are extracted from the large topic set, resulting in 1,348 English language topics, 518 German topics and 134 French topics.

Each patent in the collection consists of multiple version 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. The presence of some patents in the collection with some missing content fields indicated that they are not present in any of versions. The problem of missing data is in some cases so significant that some of these patents consist only of the title.

[1] http://www.ir-facility.org/prior-art-search1
[2] 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 (redrawn from [Magdy, 2012, p.43]).

The patent collection contains 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.

Figure 3.1(a) shows the percentage of the English, German, and French patents in the CLEF-IP 2010 collection. Some patents in the collection do not contain all sections, and some of the non-English patents do not contain translations into English. 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 are 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 disclosed invention. 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 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].

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

IR model	Metric	Patent section					
		Title	Abstract	Description	DescP5[3]	Claims	Claims1
BM25	PRES	0.370	0.488	0.539	0.476	0.504	0.474
	MAP	0.057	0.101	0.131	0.097	0.109	0.094
	A. Recall[4]	0.485	0.594	0.634	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

3.2 Baseline and Experimental Settings

We developed a baseline IR system for patent prior art search on the top of the Lucene[5] search engine[6], which processes queries using both BM25 [Robertson et al., 1994] 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[7] (Section 3.1) that contains 2.6 million patent documents and a subset of 1,281 topics (queries) in the English test set where we determined at least one valid, relevant English document was available. 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 by Magdy [2012]. In our implementation, each section of a patent (title, abstract, description, and claims) 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 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 work [Lopez and Romary, 2009b]. 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 —

[3]DescP5: The first five paragraphs of description.
[4]Average Recall
[5]Apache Lucene 4.10.2.
[6]`http://lucene.apache.org/`
[7]`http://www.ifs.tuwien.ac.at/~clef-ip/`

MAP, average recall, and 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 scoring function. We call the 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 (LMDir) and Jelinek-Mercer (LMJ) smoothing — Section 2.2.1[8]) with Lucene default settings and different sections of the patent query. It can be seen that the results for BM25 and LM with Dirichlet smoothing are very similar, therefore in this thesis, we report our results based on LM. As it can be seen in table 3.1, the figures for LMJ are incompatible with other IR models. The reason is that these numbers were reported based on Lucene's default tuning parameter (i.e., $\lambda = 1$); figures increased by changing this parameter to $\lambda = 0.7$. However, since tuning parameters were not the focus of this thesis, we ignored repeating experiments with new λ.

In addition, we compared 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 used multiple retrieval models and exploited patent metadata and citation structures. All results in this thesis used the 1,348 English topic subset as reported in the PATATRAS evaluation [Piroi, 2010a]. Since the evaluation of our systems used a slightly smaller subset of 1,281 topics, we assumed no relevant results were found by our systems for the 67 remaining topics of the 1,348 topic subset in order to ensure a fair comparison to PATATRAS.

3.3 Errors Caused by Baseline Settings

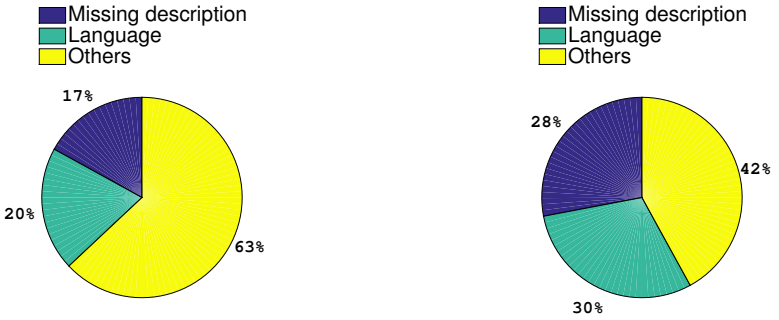
Data curation and IPC filter used in baseline settings are two sources of errors; in this section, we will discuss these two origins of the errors. This analysis only considers false negatives.

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 3.1, some patents in the union collection does not have the contents of some sections. 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.

[8] For IR models, we did not change Lucene default values: $k_1 = 1.2$, $b = 0.75$, $\mu = 2000$, and $\lambda = 1$.



(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.

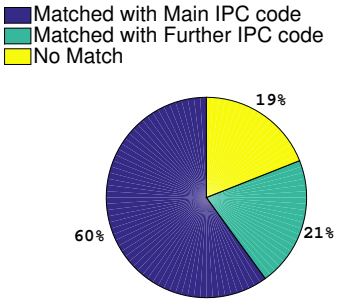
We calculate the percentage of errors caused by data curation in this experiment. As it has been illustrated in Figure 3.2(a), overall, 37% of errors are due to CLEF-IP data curation (missing description and non-English relevant patents[9]) while the majority of relevant patents, which are not retrieved (63%), are full English patent documents (Figure 3.2). These results indicate that the baseline retrieval system is ineffective to retrieve the majority of the relevant patents because of reasons other than missing description and language mismatch. 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

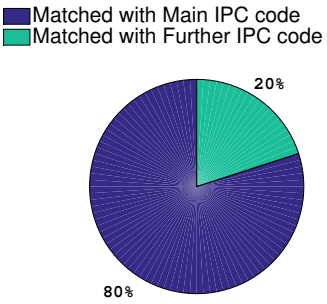
As we mentioned in Section 3.2, IPC codes (Section 2.1) are assigned to patent queries to filter the search results by constraining relevant documents to have common IPC codes with the patent query. In this section, we investigate the errors caused by classification code mismatch between topics (queries) and relevant documents for three different levels of hierarchy.

Filter Type I: Three First Components of IPC Code
First, we examine the effect of filtering out those patents, for 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

[9]Patents, which are with missing description and non-English, are included inside 'Missing discretion' subset.



(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 relevant patents (TPs and FNs).

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 the main IPC code of the query, and about 21% have, at least, one of the further IPC codes 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 the 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. It will very take long time to compare each patent in the whole collection with the query without IPC filter. 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 instead of 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. As the result, 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 we are querying with the description of the patent application, which is consisted of thousands of words.

2. The precision in the top k (e.g., 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.

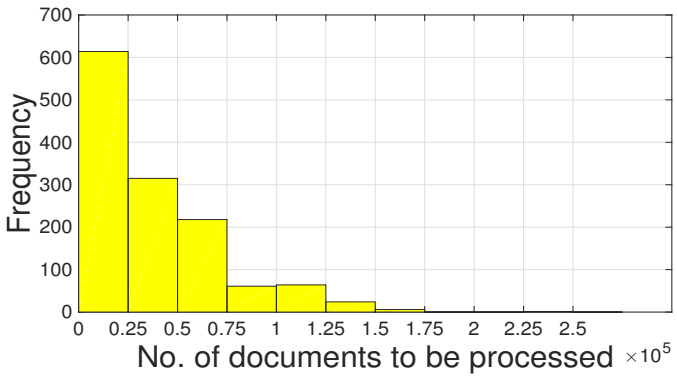


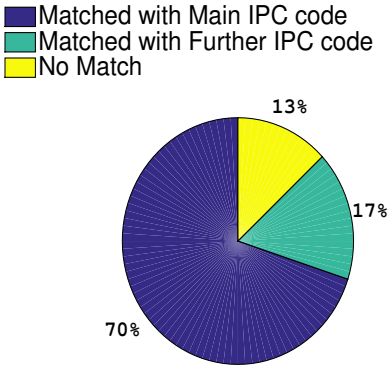
Figure 3.4: The distribution of the number of patents that should be ranked for each query over all test topics (1,303), after applying the IPC filter (filter type I). On average, the matching process for each query is done over 36,254 documents instead of the whole collection (2.6 million documents),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

We justify the efficient computational matching process by the following experiment. 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, this number is 36,254. Therefore, applying the IPC filter computationally saves us considerable amount of time since the system only needs to look at 36,254 documents per query on average instead of the entire collection, which contains 2.6 million patent documents.

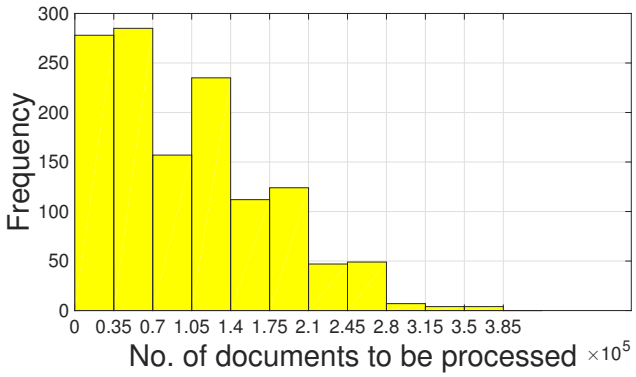
In trade off between losing 19% of relevant patents and making the ranking process faster, we choose faster computation. In addition, we notice that the histogram falls down by increasing the number of documents, which should be processed; this means that for the majority of queries the matching process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

Filter Type II: Two First Components of IPC Code

We hypothesise that the errors will be reduced, if we broaden the filter by selecting the two first components of the query IPC, namely, section, and class (e.g., C07). We repeat the experiments above 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 of the IPC code filter. However, the number of documents, which 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 the matching process does not follow the falling trend as filtering with three first components. Hence, we conclude that this



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



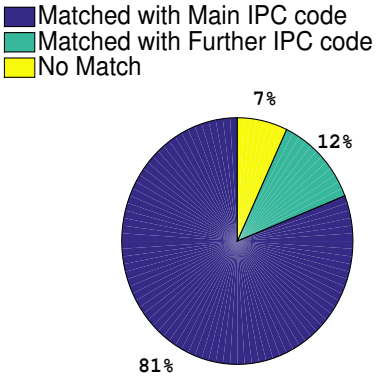
(b) The distribution of the number of patents should be ranked for each query over all test queries (1,303). In average, the matching process for each query is done over 99,754 documents instead of the whole collection (2.6 million documents), which dramatically reduce the computational time.

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

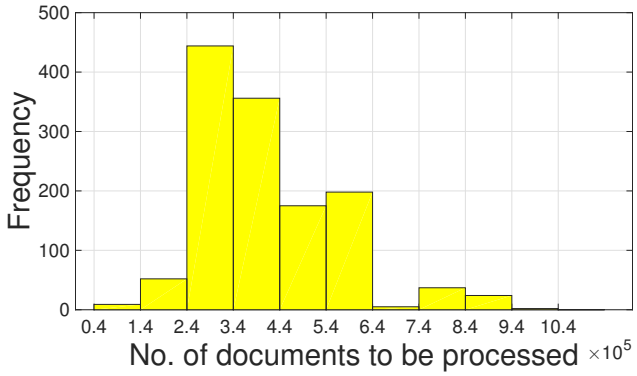
filter is not appropriate since we only reduce the error by 6% whereas the average number of documents, which should be processed, triples.

Filter Type III: First Component of IPC Code

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) 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 that 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 aver-



(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. On 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)

age. This number is much higher than the number for previous filters; this explains that using only the first component of the IPC code is not computationally efficient because it does not reduce the computational time as well it still causes 7% of the errors.

To recap our experiments related to the IPC code filtering, we showed in trade off between the errors related to applying IPC code filter and a computationally efficient matching process, we got the best results when we applied the first three IPC code symbols (e.g., section, class, and subclass) of the patent 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.

3.4 Summary

We explained the settings and development of our baseline system as well data collection we used in our experiments. We found two sources of errors: (1) errors related to data curation, and (2) errors arose from using IPC filter. However, we showed that these errors are not the main causes of low effectiveness in patent prior art search. Hence, in the next chapter, we will focus on term analysis to figure out the main reasons that generic IR techniques fail for patent prior art search (i.e., the reasons, which are not specific to our data collection or experimental set-up).

Chapter 4

Optimal Query Term Selection

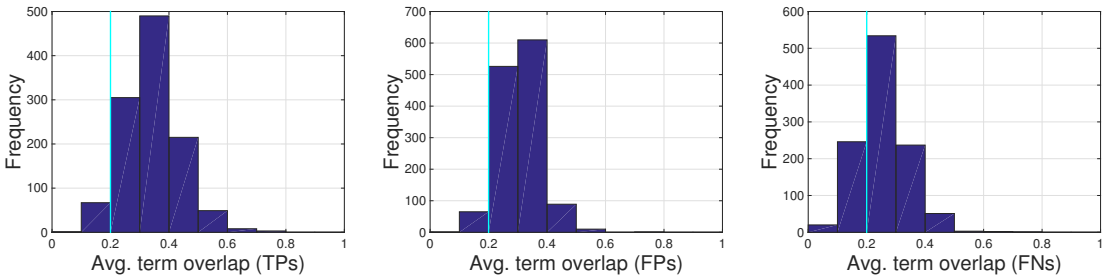
In this chapter, we investigate the problem — ineffective patent prior art search — from a term analysis perspective for both patent query and relevant documents. We are mainly interested in figuring out what is wrong in term matching process between the patent query and relevant patents that the system cannot retrieve relevant patents at top of the search result list. We hypothesise that the patent query contain sufficient terms to retrieve relevant patents and IR techniques like query expansion does not suit for prior art search.

We start with experiments that show that there is 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 queries based on this score; this gives an upper bound performance of standard Okapi BM25 and LM retrieval algorithms. In addition, our experiments demonstrate the sufficiency of terms in the patent query to achieve a high performance. So there is the need for better term selection and term weighting techniques rather than query expansion. We try four simple query reduction approaches to approximate the oracular query; then we discuss why they are not effective. Finally, we show that we can get improvements using an interactive approach, which needs minimum user effort [Golestan Far et al., 2015].

4.1 Term Mismatch

Standard retrieval models rank documents based on term matching between the queries and documents. A significant term mismatch between the patent query and relevant patents has been mentioned the main cause for low effective patent prior art search in previous studies [Roda et al., 2009; Magdy, 2012]. We examine term overlap between a patent query and three important subsets of documents for each query[1] (i) retrieved relevant patents (TPs); (ii) retrieved irrelevant patents (FPs); and (iii) not-retrieved relevant patents (FNs), respectively. We calculate the average term

[1]The cut-off rank has been chosen 100 in all experiments.



(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 in three subsets (TP, FP, FN) over all queries in English test subset[2].

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, Q is a query (we calculate this score for TP, FP, and FN subsets, respectively), D is a collection of TP patents, FP patents, or FN patents for each query, $|D|$ is the number of TP patents, FP patents, or FN patents for each query, $|terms_{Q \cap d}|$ is the number of query terms appearing in each TP, FP, or FN patent, $|Q|$ is the size of the query.

Figure 4.1 compares the distribution of term overlap between the query and documents in three subsets (TP, FP, FN), over all queries in English test subset. We summarise the main conclusions for this experiment as follows:

1. For the majority of queries (around 94% of queries), patent documents with a term overlap higher than 0.2 are retrieved (e.g, TPs and FPs).

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).

This experiment implicitly shows that a low or zero term match is not the main cause of low effectiveness in patent prior art search. Hence, in the next experiments we concentrate on analysing terms in queries and documents.

4.2 Oracular Relevance Feedback System

A query is optimal if it ranks all relevant documents before those that are not relevant, 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

[2]We assume 0.2 is a boundary for the term overlap between patent query and document, where documents with the term overlap over 0.2 are retrieved by our baseline IR system.

terms that appear in all relevant documents, but explicitly discounts all terms that occur in irrelevant documents [Manning et al., 2008, p.182].

Inspired by the idea of an optimal query, we develop an oracular relevance feedback system, which extracts terms from the judged relevant documents to derive an upper bound on performance for the standard Okapi BM25 and LM retrieval algorithms for patent prior art search. We, first, use the oracular relevance feedback to score query terms. Then, we run some experiments related to the existence of the useful terms inside the patent query. Finally, we analyse the system performance for two oracular queries formulated by high-scored terms.

4.2.1 Term Scoring

After the initial retrieval with the original patent query, we build a vocabulary set out of all terms appearing in top 100 retrieved documents; then we use judged relevant documents to score each term. We calculate an oracular relevance feedback score (i.e., $RF(t, Q)$) for each term t in the top 100 retrieved documents given the query Q as follows:

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

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

where $Rel(t)$ is the average term frequency in retrieved relevant patents, and $Irr(t)$ is the average term frequency in retrieved irrelevant patents. According to Equation 4.2, words appearing more frequently in relevant documents achieve higher RF score. We assume that words with a positive score, are Useful Terms because they are more frequent in relevant patents while words with a negative score, are Noisy Terms because they appear more frequently in irrelevant patents. We call terms with positive $RF(t, Q)$ useful terms because we assume that queries, which contain more useful terms, would perform better. We empirically seek a threshold τ for $RF(t, Q)$ to pick up useful terms (as we will show in Section 4.2.5 and Figure 4.5(a)). Hence, useful terms are terms with a positive RF score (i.e., $RF(t, Q) > 0$).

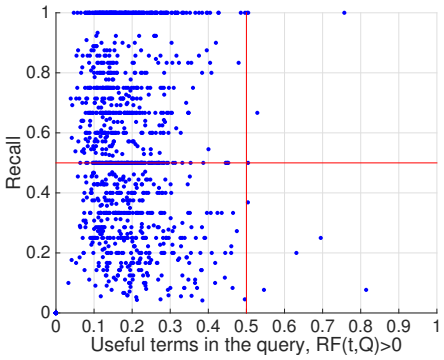
We yield the oracular relevance feedback score: (i) to find a pattern for the system performance versus useful terms; (ii) to show the term overlap with useful terms and noisy terms for TP, FN patents; (iii) to examine the existence of useful terms in different sections of the patent query; and (iv) to formulate oracular queries.

4.2.2 Performance versus Useful Terms

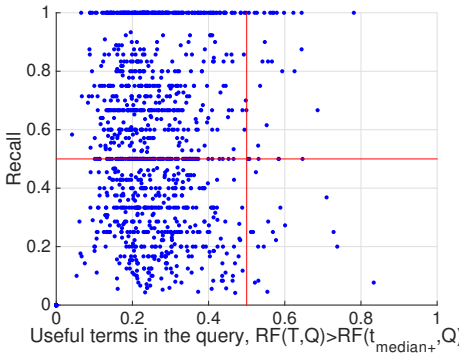
We intuitively expect that queries, which contain more useful terms achieve higher performance. Hence in this experiment, we investigate if more useful terms in the patent query leads to a higher performance. In other words, we seek a pattern that connects system performance and the existence of useful terms in the initial patent query.

We define four different criteria to select useful terms as follows:

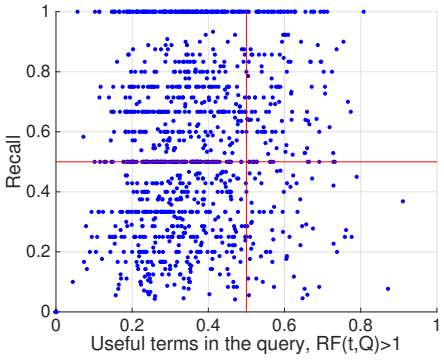
1. terms with positive RF scores (i.e., $RF(t, Q) > 0$));



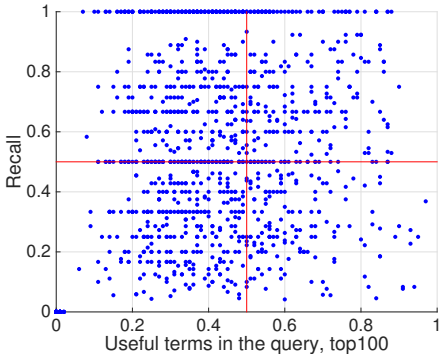
(a) Useful terms: $\{t|RF(t,Q) > 0\}$



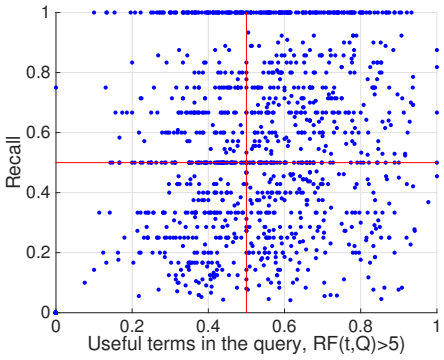
(b) Useful terms: $\{t|RF(t,Q) > RF(t_{+median})\}$



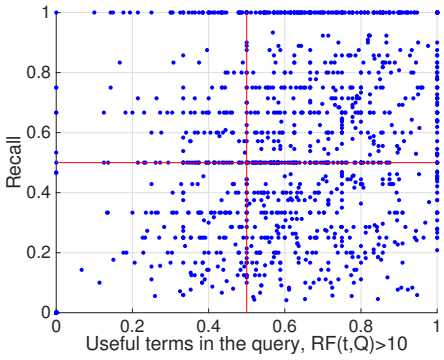
(c) Useful terms: $\{t|RF(t,Q) > 1\}$



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



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

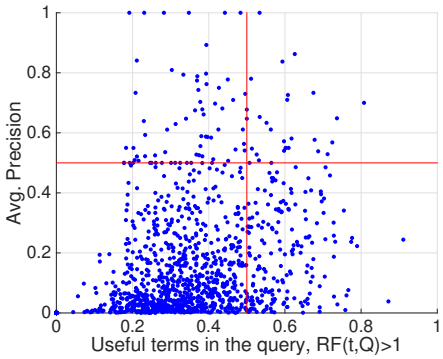


(f) Useful terms: $\{t|RF(t,Q) > 10\}$

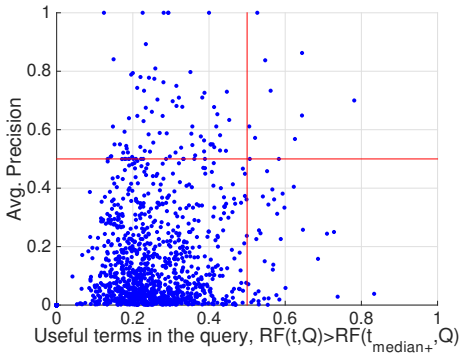
Figure 4.2: Scatter plots of recall versus the existence of useful terms in query for different values of τ.

2. terms with the score higher than the positive median score (i.e., $RF(t,Q) > RF(t_{+median},Q)$);

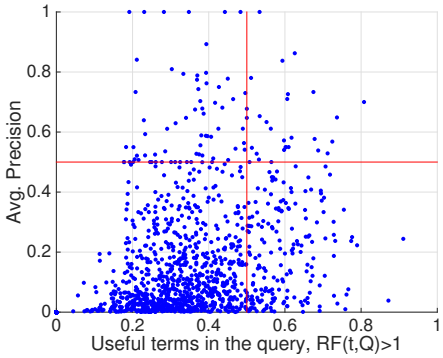
3. terms with the score higher than a constant: 1, 5, and 10 (i.e., $RF(t,Q) > 1,5,10$); and



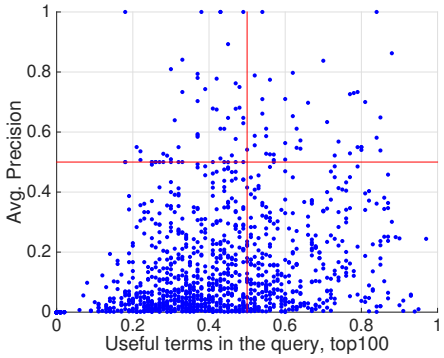
(a) Useful terms: $\{t|RF(t,Q) > 0\}$



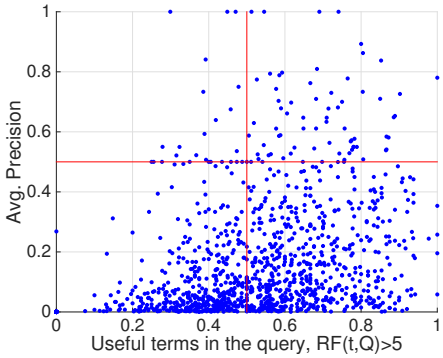
(b) Useful terms: $\{t|RF(t,Q) > RF(t_{+median}, Q)\}$



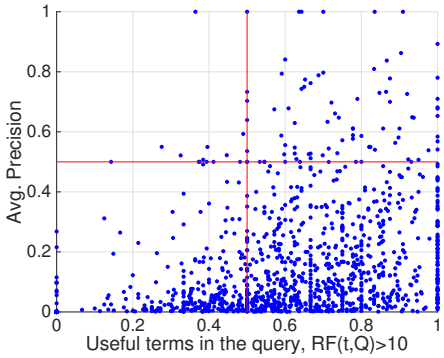
(c) Useful terms: $\{t|RF(t,Q) > 1\}$



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



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

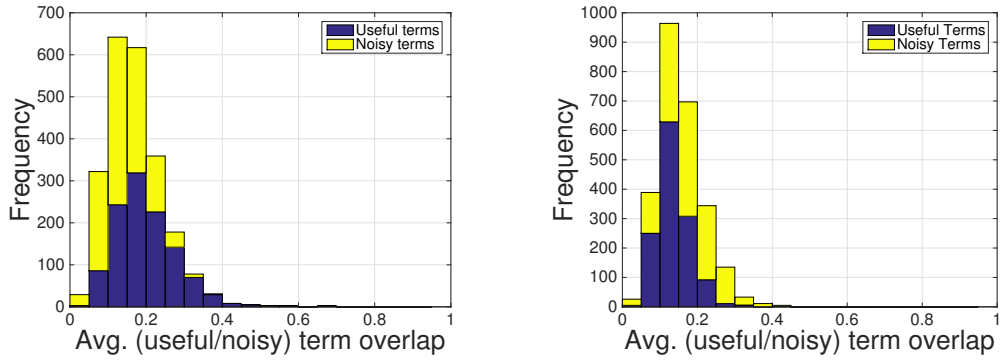


(f) Useful terms: $\{t|RF(t,Q) > 10\}$

Figure 4.3: Scatter plots of average precision (AP) versus the existence of useful terms in query for different values of τ.

4. top 100 high-scored terms.

For each query, we calculate the fraction of useful terms to all query terms. Figure 4.2 shows the scatter plot of recall versus this fraction; each blue dot representing the original patent query. As it can be seen, unlike our first assumption, we do not see



(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.

any correlation between recall and the presence of useful terms in the query and the pattern for the recall is very noisy and irregular. We repeat the experiment for the average precision (AP). Figure 4.3 shows the scatter plot of AP versus the existence of useful terms in the query. Patterns are slightly more meaningful than recall; we can see a very weak correlation between AP and useful terms inside the query for top-scored terms (i.e., $RF(t, Q) > 10$). The rightmost points on Figure 4.3(f) indicate that there are some queries containing the most important terms as well they have high performance. However it does not demonstrate our first assumption. This experiment implicitly indicates that term mismatch between the patent query and relevant documents is not the main reason for the low effectiveness in the patent prior art search. So, it seems that the attention should be focused on re-weighting.

4.2.3 Term Overlap with Useful Terms and Noisy Terms

In Section 4.1, we showed that almost both TP and FP patents have sufficient term overlap with the retrieved documents. In this 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 in the query while irrelevant patents have a higher term overlap with the noisy terms. It means that the proportion of the document, which is useful, is higher in relevant retrieved documents than irrelevant retrieved documents. This experiment shows that noisy terms cause the system to retrieve irrelevant patents at top of the list.

4.2.4 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 the existence of useful terms in different sections of patents. Table 4.1 shows the average number of useful terms

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

in different sections of a patent query. As it can be seen, description has the highest number of useful terms where RF score threshold τ is 0 or 1. The average number of the useful terms in description is higher for $\tau = 0$ than for $\tau = 1$. Compared to other sections, description contains more useful terms, which demonstrates why we achieved higher performance when querying with description (Section 3.2). Table 4.2 shows the average percentage of useful terms in different sections of a patent query. It shows that, overall, useful terms constitute less than 50% of the whole words in each section of patent queries. For example, only 27% of the description of a patent query is made of useful terms, on average, and the rest consists of irrelevant terms. This shows that noisy terms are dominant over the useful terms when querying with the description.

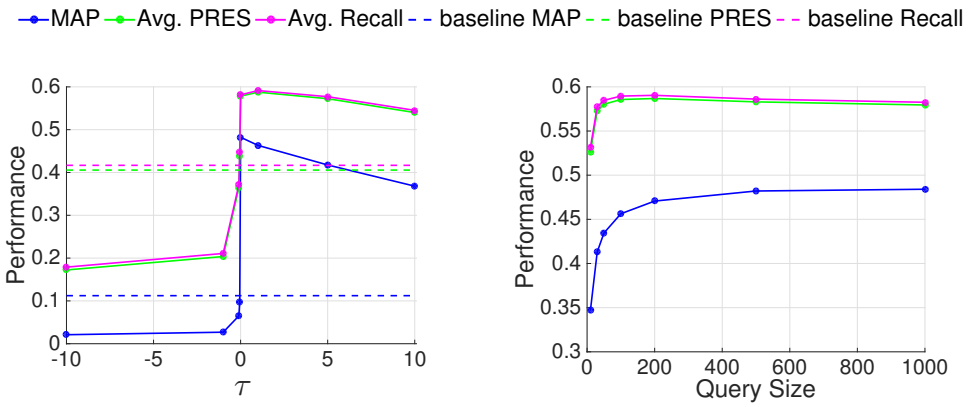
4.2.5 Oracular Query Formulation

As we illustrated in Section 4.2.2, we could not find an informative pattern for the performance and the existence of the useful terms in a patent query. In this section, we examine the system effectiveness for queries formulated using terms selected by an oracular relevance feedback system. We formulate two different oracular queries.

The *first* query is formulated by selecting terms in the top 100 retrieved documents using the oracular relevance feedback score; we call this as oracular query:

$$Oracular\ Query = \{t \in top100 | RF(t, Q) > \tau\}. \tag{4.3}$$

First, we empirically seek to evaluate the threshold τ on $RF(t, Q)$ yielding the best oracular query. Figure 4.5(a) shows that the performance changes by the values of τ and peaks at $\tau = 0$ for the MAP. We also remark that the performance jumps notably over the baseline for the oracular query formulated according to Equation 4.3. Second, we seek to determine what is the best number of terms to formulate the oracular query. Figure 4.5(b) shows that the performance increases notably when we



(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

include up to 200 terms while formulating a query, however it remains quite stable when we include more than 200 terms.

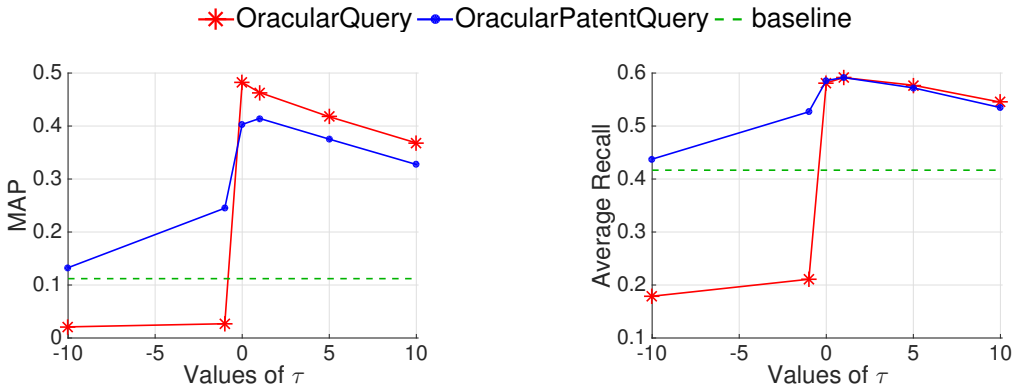
We seek to establish whether terms within a reference patent query are sufficient for obtaining a strong performance, so, we formulate the *second* query by selecting oracular terms that also occur in the reference patent query. We call it oracular patent query:

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

Baseline versus Oracular Query

We compare oracular queries versus the baseline as shown in Figure 4.6. First, we investigate the ideal threshold setting τ for the oracular queries. Notably, there is a rather unexpected steep dropoff in performance for both oracular queries when slightly noisy terms are included (i.e., τ just slightly less than 0). However, this dropoff is less pronounced for the oracular patent query indicating that restriction to query terms in the reference patent may reduce the impact of the noisy terms that are included in the reformulated query. While the oracular query and oracular patent query peak at slightly different thresholds ($\tau = 0$ and $\tau = 1$, respectively), either value of τ yields good performance. However, values of $\tau > 1$ demonstrate a stronger relative decrease in performance due to the exclusion of a large number of useful terms.

In Table 4.3, we compare our best oracular relevance queries with both the baseline patent query and the PATATRAS system. In general we found BM25 and LM to offer very similar performance. Our subsequent results use only LM due to space limitations although results for BM25 are very similar. More importantly, the oracular query using $\tau = 0$ far outperforms the baseline and approximately performs twice as well on MAP as the PATATRAS system, the best competitor in CLEF-IP 2010.



(a) Mean Average Precision τ.

(b) Average Recall

Figure 4.6: Comparing the performance of oracular query, oracular patent query and also baseline for various values of the threshold τ.

Table 4.3: Performance for the Patent Query (baseline), two variants of the Oracular Query[3], and Top CLEF-IP 2010 Competitor (PATATRAS).

		Baseline	PATATRAS	OracularQ $\tau = 0$	OracularPQ $\tau = 1$
LM	MAP	0.112	0.226	**0.482**	**0.414**
	Recall	0.416	0.467	0.582	0.591
BM25	MAP	0.123	0.226	**0.492**	**0.424**
	Recall	0.431	0.467	0.584	0.598

The MAP and the recall for the best oracular patent query (where terms are equally weighted) are respectively lower than the MAP and the recall for the best oracular query. However, the query reduction approach inherent in the oracular patent query is still sufficient to achieve MAP performance appreciably better than PATATRAS (for $\tau \geq 0$) with reduced sensitivity to the inclusion of noisy terms (when $\tau < 0$). This explains why query expansion techniques are not too effective for patent prior art search. We also conclude that the existence of the noisy terms is the main cause of low effectiveness in prior art search. In query expansion, we include noisy terms rather than useful terms that negatively affect the performance.

To summarise, our experiments[4] related to an oracular relevance feedback system suggest two important conclusions:

1. query reduction is sufficient for achieving effective prior art patent retrieval; and

[3]Two variants of the oracular query are: (1) Oracular Query (OracularQ, Equation 4.3), and (2) Oracular Patent Query (OracularPQ, Equation 4.4).

[4]In our experiments, we considered equal weights for all terms to formulate queries, because of the inconsistency of the results when we used term weighting.

2. very precise methods for eliminating poor query terms in the reduction process are needed.

4.3 Query Reduction: Approximating the Oracular 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

We first apply four simple automated query reduction techniques to improve the effectiveness of the patent prior art search. Then we analyse the reasons why these methods fail to considerably achieve a higher performance over the baseline.

4.3.1.1 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, p.27]. Inspired by this fact, we hypothesise that we will improve the performance by pruning out highly frequent terms in the top 100 retrieved documents after an initial run of the patent query. To identify highly frequent terms, we calculate the average term frequency over the top 100 documents for each term — document frequent (*DF*) score — as follows:

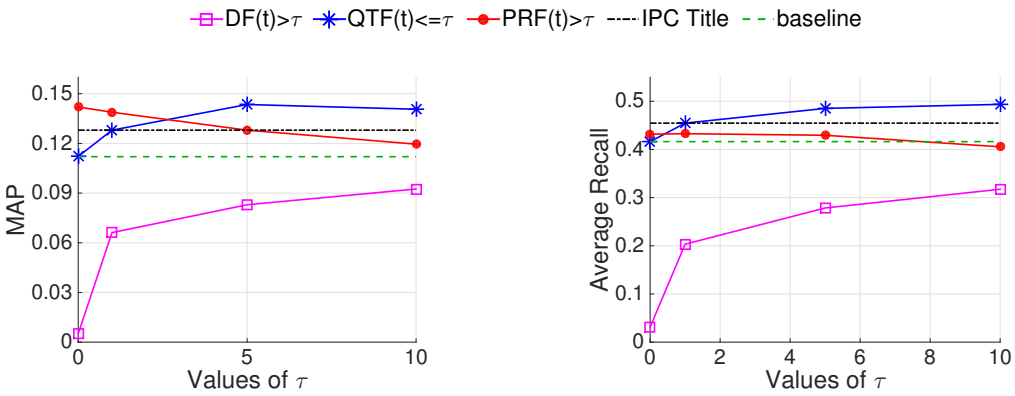
$$DF(t, Q) = \frac{1}{100} \sum_{d \in D} c(t, d), \tag{4.5}$$

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

We remove words with *DF* score higher than τ ($DF(t, Q) > \tau$) from the patent query. As illustrated in Figure 4.7 (magenta line), such pruning hurts the performance. *DF* pruning continues increasing and converges to the baseline as $\tau \to \infty$ (i.e., no pruning). Overall, removing document frequent terms from the patent query is not considered an appropriate approach since it hurts the performance.

4.3.1.2 Removing Infrequent Terms in Patent Query

Frequent terms inside long and verbose queries are considered important [Maxwell and Croft, 2013]. Thus, we hypothesise that removing terms appearing infrequently in the Patent Query may help and hence propose to remove terms with query term frequency (QTF) below a threshold τ ($QTF(t) \leq \tau$). Results in Figure 4.7 (blue line) indicate the performance is slightly better than the baseline when removing low QTF



(a) Mean Average Precision. (b) Average Recall.

Figure 4.7: System performance versus the threshold τ for four different query reduction approaches.

terms. The best MAP is achieved when $\tau = 5$ and it meets the baseline when $\tau = 0$ (i.e., all terms retained).

4.3.1.3 Removing Terms in IPC Titles

The titles of IPC codes indicate the intended content of patents classified under that code by using a single phrase or several related phrases linked together. We used words in IPC code titles associated to each patent query as stopwords to reduce the query, based on the assumption that these terms are common to all patents having the same IPC code label. As it can be seen in Figure 4.7 (black line), this approach slightly helps improving the performance.

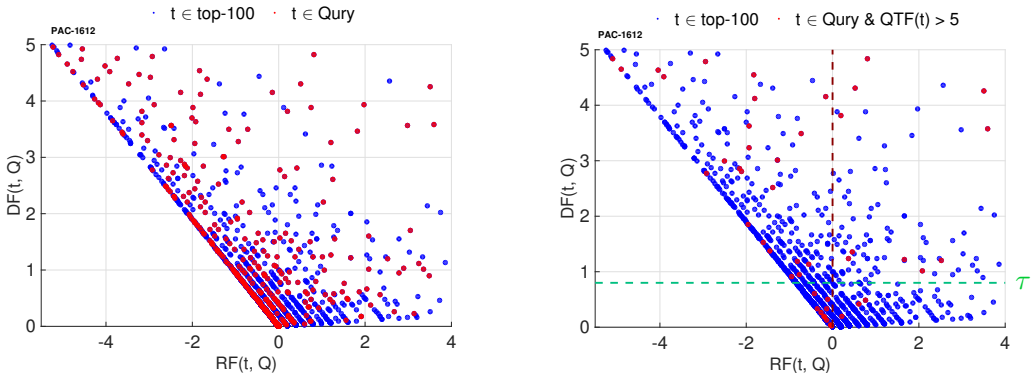
4.3.1.4 Query Reduction Using Pseudo Relevance Feedback

Pseudo relevance feedback (*PRF*) is an automated process without user interaction that 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 the top 5 retrieved documents are relevant and the rest are irrelevant, then we calculate the *PRF* score (i.e., $PRF(t, Q)$) based on this assumption:

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

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

where $Rel(t)$ is the average term frequency in the top 5 retrieved patents, and $Irr(t)$ is the average term frequency in the remained patents in the top 100 retrieved documents. We select the terms in the patent query that have the *PRF* score higher than the threshold τ ($PRF(t, Q) > \tau$) to reformulate a reduced query. Figure 4.7 (red line)



(a) Red points are all terms in the query (e.g, $t \in Query$).

(b) Red points are query terms where QTF is higher than 5 (i.e., $t \in Query \bigwedge QTF(t) > 5$).

Figure 4.8: Scatter plot of DF score versus RF score. This anecdotal example analyses the 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.

shows that this approach is also unsuccessful at achieving a notable improvement over the baseline.

4.3.1.5 Automated Techniques Fail to Approximate Oracular Patent Query

In Section 4.2.5, we showed that terms inside the patent query are sufficient to get a noticeable improvement over the baseline; however, results related to automated query reduction techniques showed only a little improvement. We analyse the causes through the following experiments.

First, we use an anecdotal example — a sample query — to analyse terms selected by proposed query reduction methods. Figure 4.8(a) shows a scatter plot of DF score versus RF score for the sample query — PAC-1612. Each blue point is a vocabulary in top 100 retrieved document vocabulary set. The figure shows a negative correlation between $DF(t, Q)$ and $RF(t, Q)$ score; however it does not indicate that a term with a higher DF score has essentially a lower RF score. Hence in the DF pruning method, the removal of a document frequent term does not mean that the term is not important to represent the document. As it is illustrated in Figure 4.8(b), by removing document frequent terms (i.e., $DF(t, Q) > \tau$), we also remove many useful terms (e.g., terms with $RF(t, Q) > 0$). In Figure 4.8(a), the red points represent all query terms and in Figure 4.8(b), they represent query terms with term frequency higher than 5 (i.e., $QTF(t) > 5$) — where we got the best performance. The comparison of Figures 4.8(a) and 4.8(b) shows that by pruning out terms with $QTF(t) < 5$, on the one hand, we are removing considerable amount of noisy terms; on the other hand, we are removing useful terms too. In addition, the remaining terms are not purely useful terms because they are still contaminated by noisy terms (e.g., terms with $RF(t, Q) < 0$).

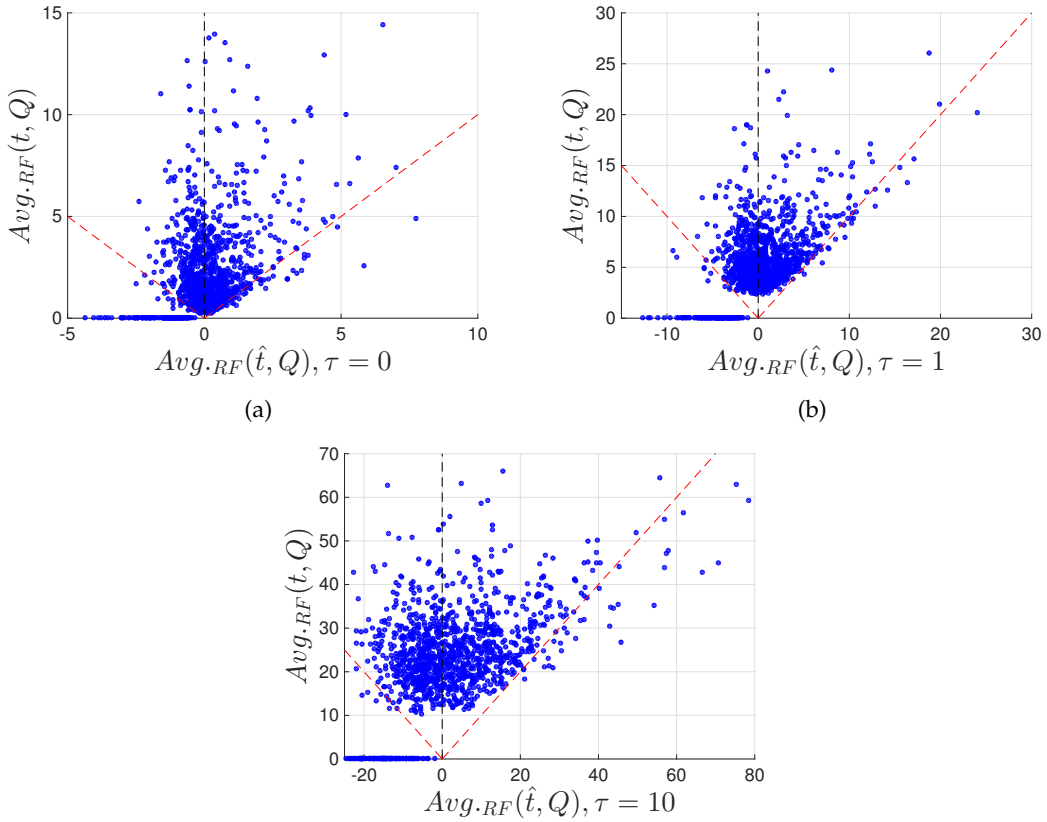


Figure 4.9: Comparing *RF* score of top relevance feedback terms and top pseudo relevance feedback terms for three different values of the threshold τ (i.e., $0, 1, 10$).

In the second experiment, we analyse why the term selection technique using pseudo relevance feedback fails to approximate the oracular patent query. 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* terms (i.e., $RF(t,Q) > \tau$), $Avg_{RF}(\hat{t},Q)$ is the average *RF* score for top *PRF* terms (i.e., $PRF(\hat{t},Q) > \tau$), τ is a threshold for the score to select terms, t is a symbol for terms with high *RF* score and $\hat{t}$ is a symbol for terms with high *PRF* score.

```
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 4.10: The top terms scored by each of four methods on a sample query (except for IPC title terms which are not scored); whether the term is pruned or retained depends on each approach. Numerical oracular scores $RF(t, Q)$ are provided indicating whether the term was actually useful (blue/positive) or noisy (red/negative).

Figure 4.9 shows a scatter plot of the average *RF* score for top relevance feedback terms versus the average *RF* score for top pseudo relevance feedback terms. First, we observe that the *RF* score of top relevance feedback terms is higher than the *RF* score of top pseudo relevance feedback terms for queries (i.e., $Avg_{RF}(t, Q) > Avg_{RF}(\hat{t}, Q)$). We can also see that for about half of the queries, $Avg_{RF}(\hat{t}, Q)$ is negative that indicates we are selecting noisy terms by their pseudo relevance feedback score rather than useful terms. Second, we can find a weak positive correlation toward selecting positive terms by pseudo relevance feedback; this is the reason why we could obtain a small improvement when applying query reduction using *PRF*.

Finally, we analyse the reasons that four proposed query reduction approaches fail to approximate the oracular patent query, using an anecdotal example of a sample query about an invention related to "emulsifier". Figure 4.10 shows the raw abstract of the invention, and the top high-scoring terms (except for IPC title terms which are not scored, but simply displayed) and their associated *RF* scores for each approach. Terms are chosen based on their scores for each approach as follows:

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

It can be seen that the four methods fail to clearly discriminate between useful and noisy terms. Important stemmed terms like "enzym" and "starch" would be pruned according to *DF*; in contrast, *QTF* and *PRF* both score "starch" highly and retain it, but also retain other noisy terms (e.g., highly noisy terms like "amylos" with $RF(t, Q) = -20.15$ or "amylopectin" with $RF(t, Q) = -7.14$). Over half of the IPC title terms are noisy and appropriate to remove, but critical useful stemmed terms like "emulsifi" are also removed. Critically, all methods retain noisy terms (red/negative) and results from Section 4.2.5 showed that the inclusion of even slightly noisy terms can significantly hurt performance. Overall, all methods fail to retain only the oracular query terms (blue/positive) and do worse than PATATRAS.

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 evaluate the impact of minimal user interaction, we next analyse the performance of an oracular patent query (Equation 4.4) derived from *only* the top k ranked relevant documents identified in the search results (for small k) — we assume that the remaining documents in the top 100 are irrelevant. Using this approach, Table 4.4 shows that we can *double* the MAP in comparison to our baseline and also outperform the PATATRAS system by identifying only the *first* relevant document ($k = 1$). The MAP *triples* by the identification of three first relevant documents ($k = 3$).

Table 4.4: System performance using minimal relevance feedback in comparison with baseline and PATATRAS. τ is RF score threshold, and k indicates the number of top relevant patents.

	Baseline	PATATRAS	$k=1$ $\tau=0$	$k=1$ $\tau=1$	$k=3$ $\tau=0$	$k=3$ $\tau=1$
MAP	0.112	0.226	0.288	**0.289**	**0.369**	0.368
Recall	0.416	0.467	0.479	0.484	0.547	0.550

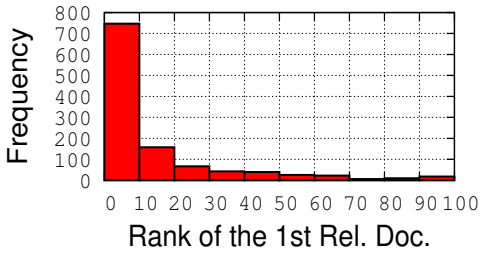


Figure 4.11: The distribution of the first relevant document rank over test queries.

Furthermore, to establish the minimal interaction required by this approach, Figure 4.11 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.

4.4 Summary

Our experiments demonstrated the sufficiency of terms inside the patent query for an efficient prior art search. However, we showed that the inclusion of even slight noisy terms can significantly hurt the performance.

We achieved an upper bound on performance through building an oracular query from known relevance judgements. Our attempts to automatically approximate the oracular query using different query term selection techniques could not retain only oracular query terms and the reformulated query was contaminated with sufficient noisy terms to lead to ineffective performance. Finally, we found that by identifying only the first relevant document from the search results we could outperform the baseline as well the PATATRAS system.

Chapter 5

Conclusions

In this thesis, we investigated the reasons that make patent prior art search less effective than other search applications. We started with recognising errors due to data curation and baseline settings; these contribute a small portion of the total errors. We hypothesized that the main portion of the errors is 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 studies proposed different solutions to improve retrieval effectiveness including e.g., query expansion, we focused on analysing terms in the patent query and the top 100 retrieved patents. After defining an oracular query based on relevance judgements, we established (1) the sufficiency of the standard retrieval scoring models and (2) query reduction methods to achieve state-of-the-art patent prior art search performance. After finding that automated methods for query reduction 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 when a single relevant document is provided to the system. Given that such a simple interactive method for query reduction with a standard retrieval model outperforms 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[1].

5.1 Contributions

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

1. **Analysis of optimal queries:** 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 the oracular system suggested the necessity of precise query reduction and term selection techniques to improve the effectiveness in patent prior art search.

[1] We used only the CLEF-IP collection to empirically investigate the proposed solution due to the limited availability of additional test collections. The use of only one test collection for this investigation may limit the generalizability of the findings reported here.

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 to prune the negative terms out. We showed that these approaches were inefficient because they could not discriminate between useful terms and noisy terms. Since generic IR models demonstrated to be over-sensitive to the existence of noisy terms, the performance improved a little by our proposed automated techniques.

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

5.2 Future Work

In this research, we analysed the key reasons making generic IR methods ineffective for patent prior art through various experiments that may open further research enquiries on the topic of prior art search. We describe the limitations and discuss further improvements as follows:

5.2.1 Exploring Other Term Scoring Methods

Our term scoring method was inspired by Rocchio optimal query [Manning et al., 2008, p.181]. We used this score to select query terms that 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

We demonstrated that useful terms in the patent query are sufficient for an effective retrieval. We showed that a query, formulated using a precise selection of useful terms, considerably outperforms the baseline and PATARAS. We could not approximate the oracular query using automated techniques because the retrieval models are over-sensitive to noisy terms and our proposed reduction approaches were incapable of discriminating between useful terms and noisy terms. Hence, we need more sophisticated query term selection techniques, which differentiate useful terms from noisy terms. For example, query term selection technique, proposed by Maxwell and Croft [2013] using affinity graph and random walk, can be applied for patent prior art search. Other example is the work by Kumaran and Carvalho [2009]; they used learning to rank all sub-sets of the original query (sub-queries) based on their predicted quality, and select the top sub-query.

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 and network analysis is a proper alternative to a traditional IR based on term matching process when the retrieval problem based on term matching is difficult. Patents are rich in meta data and highly structured in terms of entities and relations; for example the bibliographic meta data in the patent XML file contains details about its inventor, organisation, and other information that can build a multidimensional graph. Regarding this network structure of patents, we can find possible prior works in the scientific contributions (e.g., research articles) of inventors. Also competitive organisations may have developed the same or very close idea prior to the idea, which is claimed in the 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*, CIKM '08 (Napa Valley, California, USA, 2008), 561–570. Association for Computing Machinery (ACM), New York, NY, USA. doi:10.1145/1458082.1458157. (cited on page 13)

Bache, R. and Azzopardi, L., 2010. Improving access to large patent corpora. *T. Large-Scale Data- and Knowledge-Centered Systems*, 2 (2010), 103–121. doi:10.1007/978-3-642-16175-9_4. (cited on pages 10, 13, and 18)

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, 49, and 56)

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*, SIGIR '10 (Geneva, Switzerland, 2010), 571–578. ACM, New York, NY, USA. doi:10.1145/1835449.1835545. (cited on page 16)

Bashir, S. and Rauber, A., 2009a. Analyzing document retrievability in patent retrieval settings. In *Database and Expert Systems Applications, 20th International Conference, DEXA 2009, Linz, Austria, August 31 - September 4, 2009. Proceedings*, 753–760. Springer Berlin Heidelberg. doi:10.1007/978-3-642-03573-9_63. (cited on page 18)

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*, CIKM '09 (Hong Kong, China, 2009), 1863–1866. ACM, New York, NY, USA. doi:10.1145/1645953.1646250. (cited on pages 13, 18, and 20)

Bashir, S. and Rauber, A., 2010. Improving retrievability of patents in prior-art search. In *Advances in Information Retrieval, 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31, 2010. Proceedings*, 457–470. Springer Berlin Heidelberg. doi:10.1007/978-3-642-12275-0_40. (cited on pages 18 and 21)

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 (JASIST)*, 62, 8 (2011), 1515–1532. doi:10.1002/asi.21549. (cited on pages 13 and 14)

Becks, D.; Mandl, T.; and Womser-Hacker, C., 2010. Phrases or terms? the impact of different query types. In *CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy*. http://ceur-ws.org/Vol-1176/CLEF2010wn-CLEF-IP-BecksEt2010.pdf. (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*, SIGIR '08 (Singapore, Singapore, 2008), 491–498. ACM, New York, NY, USA. doi:10.1145/1390334.1390419. (cited on page 16)

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*, WSDM '10 (New York, New York, USA, 2010), 31–40. ACM, New York, NY, USA. doi:10.1145/1718487.1718492. (cited on page 16)

Bouadjenek, M. R.; Sanner, S.; and Ferraro, G., 2015. A study of query reformulation for patent prior art search with partial patent applications. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, ICAIL '15 (San Diego, California, 2015), 23–32. ACM, New York, NY, USA. doi:10.1145/2746090.2746092. (cited on page vii)

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*, SIGIR '08 (Singapore, Singapore, 2008), 243–250. ACM, New York, NY, USA. doi:10.1145/1390334.1390377. (cited on pages 14 and 15)

Carpineto, C. and Romano, G., 1999. Towards more effective techniques for automatic query expansion. In *Research and Advanced Technology for Digital Libraries*, vol. 1696 of *Lecture Notes in Computer Science*, 126–141. Springer Berlin Heidelberg. doi:10.1007/3-540-48155-9_10. (cited on page 15)

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 (JASIST)*, 63, 3 (2012), 512–527. doi:10.1002/asi.21708. (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 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy*. http://ceur-ws.org/Vol-1176/CLEF2010wn-CLEF-IP-DhondtEt2010.pdf. (cited on page 24)

D'HONDT, E.; VERBERNE, S.; ALINK, W.; AND CORNACCHIA, R., 2011. Combining document representations for prior-art retrieval. In *CLEF 2011 Labs and Workshop, Notebook Papers, 19-22 September 2011, Amsterdam, The Netherlands*. http://ceur-ws.org/Vol-1177/CLEF2011wn-CLEF-IP-DHondtEt2011.pdf. (cited on page 20)

EFTHIMIADIS, E. N., 1996. Query expansion. *Annual Review of Information Systems and Technology (ARIST)*, (1996). (cited on page 14)

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*, SIGIR '07 (Amsterdam, The Netherlands, 2007), 793–794. ACM, New York, NY, USA. doi:10.1145/1277741.1277912. (cited on pages 19 and 24)

FUJII, A., 2007b. Integrating content and citation information for the NTCIR-6 patent retrieval task. In *Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, NTCIR-6, National Center of Sciences, Tokyo, Japan, May 15-18, 2007*, 377–380. National Institute of Informatics (NII). doi:10.1016/j.ipm.2006.11.004. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings6/NTCIR/76.pdf. (cited on page 24)

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 18)

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 (Jun. 2005), 207–235. doi:10.1145/1105696.1105853. (cited on page 24)

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*, PaIR '11 (Glasgow, Scotland, UK, 2011), 13–18. ACM, New York, NY, USA. doi:10.1145/2064975.2064981. (cited on page 22)

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*, CIKM '11 (Glasgow, Scotland, UK, 2011), 1953–1956. ACM, New York, NY, USA. doi:10.1145/2063576.2063863. (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, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 -*

October 2, 2009, Revised Selected Papers, vol. 6241 of *Lecture Notes in Computer Science*, 444–451. Springer Berlin Heidelberg. doi:10.1007/978-3-642-15754-7_53. (cited on pages 19 and 24)

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 and Development in Information Retrieval*, SIGIR '15 (Santiago, Chile, August 09 - 13 2015). ACM, New York, NY, USA. (cited on page 39)

Graf, E.; Frommholz, I.; Lalmas, M.; and Van Rijsbergen, K., 2010. Knowledge modeling in prior art search. In *Advances in Multidisciplinary Retrieval, First Information Retrieval Facility Conference, IRFC 2010, Vienna, Austria, May 31, 2010. Proceedings*, vol. 6107 of *Lecture Notes in Computer Science*, 31–46. Springer Berlin Heidelberg. doi:10.1007/978-3-642-13084-7_4. (cited on page 24)

Gurulingappa, H.; Müller, B.; Klinger, R.; Mevissen, H.; 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 *Proceedings of The Nineteenth Text REtrieval Conference, TREC 2010, Gaithersburg, Maryland, USA, November 16-19, 2010*. http://trec.nist.gov/pubs/trec19/papers/fraunhofer-scai.chem.rev.pdf. (cited on page 24)

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*, PaIR '10 (Toronto, ON, Canada, 2010), 27–32. ACM, New York, NY, USA. doi:10.1145/1871888.1871894. (cited on pages 7 and 24)

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 Berlin Heidelberg. (cited on page 24)

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*, PaIR '09 (Hong Kong, China, 2009), 29–32. ACM, New York, NY, USA. doi:10.1145/1651343.1651350. (cited on page 24)

He, B. and Ounis, I., 2009. Finding good feedback documents. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09 (Hong Kong, China, 2009), 2011–2014. ACM, New York, NY, USA. doi:10.1145/1645953.1646289. (cited on page 15)

Hearst, M. A., 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23, 1 (1997), 33–64. (cited on page 22)

Herbert, B.; Szarvas, G.; and Gurevych, I., 2009. Prior art search using international patent classification codes and all-claims-queries. In *Multilingual Information*

Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers, vol. 6241 of *Lecture Notes in Computer Science*, 452–459. Springer Berlin Heidelberg. doi:10.1007/978-3-642-15754-7_54. (cited on page 24)

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 22)

Jochim, C.; Lioma, C.; and Schütze, H., 2011. Expanding queries with term and phrase translations in patent retrieval. In *Multidisciplinary Information Retrieval - Second Information Retrieval Facility Conference, IRFC 2011, Vienna, Austria, June 6, 2011. Proceedings*, vol. 6653 of *Lecture Notes in Computer Science*, 16–29. Springer Berlin Heidelberg. doi:10.1007/978-3-642-21353-3_3. (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*, IIiX '10 (New Brunswick, New Jersey, USA, 2010), 13–24. ACM, New York, NY, USA. doi:10.1145/1840784.1840789. (cited on pages 2, 25, and 32)

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. doi:10.1016/j.ipm.2006.11.006. (cited on page 24)

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*, SIGIR '14 (Gold Coast, Queensland, Australia, 2014), 891–894. ACM, New York, NY, USA. doi:10.1145/2600428.2609467. (cited on page 20)

Kishida, K., 2002. 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, NTCIR-3, Tokyo, Japan, October 8-10, 2002*. National Institute of Informatics (NII). http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/NTCIR3-PATENT-KishidaK.pdf. (cited on page 20)

Konishi, K., 2005. Query terms extraction from patent document for invalidity search. In *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, NTCIR-5, National Center of Sciences, Tokyo, Japan, December 6-9, 2005*, vol. 5. National Institute of Informatics (NII). http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings5/data/PATENT/NTCIR5-PATENT-KonishiK.pdf. (cited on pages 19 and 20)

Kontostathis, A. and Kulp, S., 2008. The effect of normalization when recall really matters. In *Proceedings of the 2008 International Conference on Information & Knowledge Engineering, IKE 2008, July 14-17, 2008, Las Vegas, Nevada, USA*, 96–101. CSREA Press. (cited on page 18)

Kullback, S. and Leibler, R. A., 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, (1951), 79–86. (cited on page 22)

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*, SIGIR '09 (Boston, MA, USA, 2009), 564–571. ACM, New York, NY, USA. doi:10.1145/1571941.1572038. (cited on pages 16 and 56)

Lease, M.; Allan, J.; and Croft, W. B., 2009. Regression rank: Learning to meet the opportunity of descriptive queries. In *Advances in Information Retrieval, 31th European Conference on IR Research, ECIR 2009, Toulouse, France, April 6-9, 2009. Proceedings*, vol. 5478 of *Lecture Notes in Computer Science*, 90–101. Springer Berlin Heidelberg. ISBN 978-3-642-00957-0. doi:10.1007/978-3-642-00958-7_11. (cited on page 16)

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*, SIGIR '08 (Singapore, Singapore, 2008), 235–242. ACM, New York, NY, USA. doi:10.1145/1390334.1390376. (cited on page 20)

Lopez, P. and Romary, L., 2009a. Multiple retrieval models and regression models for prior art search. In *Working Notes for CLEF 2009 Workshop co-located with the 13th European Conference on Digital Libraries (ECDL 2009) , Corfù, Greece, September 30 - October 2, 2009.* http://ceur-ws.org/Vol-1175/CLEF2009wn-CLEFIP-LopezEt2009.pdf. (cited on pages 23, 24, and 25)

Lopez, P. and Romary, L., 2009b. PATATRAS: retrieval model combination and regression models for prior art search. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers*, vol. 6241 of *Lecture Notes in Computer Science*, 430–437. Springer Berlin Heidelberg. ISBN 978-3-642-15753-0. doi:10.1007/978-3-642-15754-7. (cited on pages 24 and 31)

Lopez, P.; Romary, L.; et al., 2010. Experiments with citation mining and key-term extraction for prior art search. In *CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy*, vol. 1176 of *CEUR Workshop Proceedings*. CEUR-WS.org. http://ceur-ws.org/Vol-1176/CLEF2010wn-CLEF-IP-LopezEt2010.pdf. (cited on pages 25 and 32)

Lupu, M.; Hanbury, A.; et al., 2013a. Patent retrieval. *Foundations and Trends in Information Retrieval*, 7, 1 (2013), 1–97. doi:10.1561/1500000027. (cited on pages 2, 13, 23, 24, and 25)

Lupu, M.; Piroi, F.; and Hanbury, A., 2013b. Evaluating flowchart recognition for patent retrieval. In *Proceedings of the 5th International Workshop on Evaluating Information Access, EVIA 2013, National Center of Sciences, Tokyo, Japan, June 18*, 37–44. National Institute of Informatics (NII). http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings10/pdf/EVIA/08-EVIA2013-LupuM.pdf. (cited on page 25)

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, 29, 30, 31, and 39)

Magdy, W. and Jones, G. J., 2010a. 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*, SIGIR '10 (Geneva, Switzerland, 2010), 611–618. ACM, New York, NY, USA. doi:10.1145/1835449.1835551. (cited on pages xvii, 25, and 26)

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*, PaIR '11 (Glasgow, Scotland, UK, 2011), 19–24. ACM, New York, NY, USA. doi:10.1145/2064975.2064982. (cited on pages 20 and 21)

Magdy, W. and Jones, G. J., 2014. Studying machine translation technologies for large-data CLIR tasks: a patent prior-art search case study. *Information Retrieval*, 17 (2014), 492–519. doi:10.1007/s10791-013-9231-6. (cited on page 25)

Magdy, W. and Jones, G. J. F., 2010b. Applying the KISS principle for the CLEF-IP 2010 prior art candidate patent search task. In *CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy*. http://ceur-ws.org/Vol-1176/CLEF2010wn-CLEF-IP-MagdyEt2010.pdf. (cited on page 24)

Magdy, W.; Leveling, J.; and Jones, G. J., 2009. Exploring structured documents and query formulation techniques for patent retrieval. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers*, vol. 6241 of *Lecture Notes in Computer Science*, 410–417. Springer Berlin Heidelberg. doi:10.1007/978-3-642-15754-7_48. (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 - 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings*, vol. 6611 of *Lecture Notes in Computer Science*, 725–728. Springer Berlin Heidelberg. doi:10.1007/978-3-642-20161-5_80. (cited on page 24)

Mahdabi, P.; Andersson, L.; Hanbury, A.; and Crestani, F., 2011a. Report on the CLEF-IP 2011 experiments: Exploring patent summarization. In *CLEF 2011 Labs and Workshop, Notebook Papers, 19-22 September 2011, Amsterdam, The Netherlands*. http://ceur-ws.org/Vol-1177/CLEF2011wn-CLEF-IP-MahdabiEt2011.pdf. (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*, SIGIR '12 (Portland, Oregon, USA, 2012), 505–514. ACM, New York, NY, USA. doi:10.1145/2348283.2348353. (cited on page 21)

Mahdabi, P. and Crestani, F., 2012. Learning-based pseudo-relevance feedback for patent retrieval. In *Multidisciplinary Information Retrieval - 5th International Retrieval Facility Conference, IRFC 2012, Vienna, Austria, July 2-3, 2012 Proceedings*, vol. 7356 of *Lecture Notes in Computer Science*, 1–11. Springer Berlin Heidelberg. doi:10.1007/978-3-642-31274-8_1. (cited on page 21)

Mahdabi, P. and Crestani, F., 2014. Patent query formulation by synthesizing multiple sources of relevance evidence. *ACM Trans. Inf. Syst.*, 32, 4 (Oct. 2014), 16:1–16:30. doi:10.1145/2651363. (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*, SIGIR '13 (Dublin, Ireland, 2013), 113–122. ACM, New York, NY, USA. doi:10.1145/2484028.2484056. (cited on page 21)

Mahdabi, P.; Keikha, M.; Gerani, S.; Landoni, M.; and Crestani, F., 2011b. Building queries for prior-art search. In *Multidisciplinary Information Retrieval - Second Information Retrieval Facility Conference, IRFC 2011, Vienna, Austria, June 6, 2011. Proceedings*, vol. 6653 of *Lecture Notes in Computer Science*, 3–15. Springer Berlin Heidelberg. doi:10.1007/978-3-642-21353-3_2. (cited on pages 19 and 23)

Manning, C. D.; Raghavan, P.; and Schütze, H., 2008. *Introduction to Information Retrieval*, vol. 1. Cambridge University Press. (cited on pages xvii, 9, 10, 11, 14, 15, 16, 17, 18, 41, 48, and 56)

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 (Jun. 2005), 190–206. doi:10.1145/1105696.1105702. (cited on pages 19 and 25)

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*, SIGIR '13 (Dublin, Ireland, 2013), 583–592. ACM, New York, NY, USA. doi:10.1145/2484028.2484096. (cited on pages 48, 49, and 56)

Metzler, D., 2007. Using gradient descent to optimize language modeling smoothing parameters. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07 (Amsterdam, The Netherlands, 2007), 687–688. ACM, New York, NY, USA. doi:10.1145/1277741.1277859. (cited on page 11)

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*, CIKM '97 (Las Vegas, Nevada, USA, 1997), 216–221. ACM, New York, NY, USA. doi:10.1145/266714.266899. (cited on page 19)

Pérez-Iglesias, J.; Rodrigo, Á.; and Fresno, V., 2010. Using BM25F and KLD for pattern retrieval. In *CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy*. http://ceur-ws.org/Vol-1176/CLEF2010wn-CLEF-IP-PerezEt2010.pdf. (cited on page 23)

Piroi, F., 2010a. CLEF-IP 2010: Prior art candidates search evaluation summary. Technical report, Technical Report IRF TR 2010 00003, Information Retrieval Facility, Vienna. (cited on page 32)

Piroi, F., 2010b. 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 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*. http://ceur-ws.org/Vol-1178/CLEF2012wn-CLEFIP-PiroiEt2012.pdf. (cited on page 25)

Ponte, J. M. and Croft, W. B., 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98 (Melbourne, Australia, 1998), 275–281. ACM, New York, NY, USA. doi:10.1145/290941.291008. (cited on page 9)

Porter, M. F., 1980. An algorithm for suffix stripping. In *Program*, vol. 14, 130–137. (cited on page 31)

Robertson, S. and Zaragoza, H., 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3, 4 (2009), 333–389. doi:10.1561/1500000019. (cited on page 9)

Robertson, S. E., 1991. On term selection for query expansion. *Journal of Documentation*, 46, 4 (1991), 359–364. doi:10.1108/eb026866. (cited on page 15)

Robertson, S. E. and Walker, S., 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94 (Dublin, Ireland, 1994), 232–241. Springer-Verlag New York, Inc., New York, NY, USA. http://dl.acm.org/citation.cfm?id=188490.188561. (cited on page 9)

Robertson, S. E.; Walker, S.; Jones, S.; Hancock-Beaulieu, M. M.; and Gatford, M., 1994. Okapi at trec-2. In *The Second Text REtrieval Conference (TREC-2)*, 21–34. Gaithersburg, MD: NIST. http://research.microsoft.com/apps/pubs/default.aspx?id=67648. (cited on page 31)

Rocchio, J. J., 1971. Relevance feedback in information retrieval. In *Salton [1971]*, 313–323. (cited on page 14)

Roda, G.; Tait, J.; Piroi, F.; and Zenz, V., 2009. CLEF-IP 2009: Retrieval experiments in the intellectual property domain. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers*, 385–409. doi:10.1007/978-3-642-15754-7_47. (cited on pages 20, 25, and 39)

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)

Salton, G., 1971. *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA. (cited on page 68)

Salton, G.; Wong, A.; and Yang, C. S., 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18, 11 (Nov. 1975), 613–620. doi:10.1145/361219.361220. (cited on page 9)

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*, CIKM '04 (Washington, D.C., USA, 2004), 399–405. ACM, New York, NY, USA. doi:10.1145/1031171.1031251. (cited on pages 19 and 22)

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*, ICAIL '11 (Pittsburgh, Pennsylvania, 2011), 249–255. ACM, New York, NY, USA. doi:10.1145/2018358.2018393. (cited on pages 20 and 24)

Verma, M. and Varma, V., 2011b. Exploring keyphrase extraction and IPC classification vectors for prior art search. In *CLEF 2011 Labs and Workshop, Notebook Papers,*

19-22 September 2011, Amsterdam, The Netherlands. http://ceur-ws.org/Vol-1177/CLEF2011wn-CLEF-IP-VermaEt2011.pdf. (cited on page 25)

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*, CIKM '09 (Hong Kong, China, 2009), 2037–2040. ACM, New York, NY, USA. doi:10.1145/1645953.1646295. (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*, SIGIR '09 (Boston, MA, USA, 2009), 808–809. ACM, New York, NY, USA. doi:10.1145/1571941.1572139. (cited on pages 2, 19, and 32)

Zhai, C. and Lafferty, J., 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22, 2 (Apr. 2004), 179–214. doi:10.1145/984321.984322. (cited on pages 11, 12, and 31)

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*, IIiX '10 (New Brunswick, New Jersey, USA, 2010), 245–254. ACM. doi:10.1145/1840784.1840820. (cited on page 2)