Patent Prior-art Search

Mona Glestan Far

mona.golestanfar@anu.edu.au

A thesis submitted for the degree of
YOUR DEGREE NAME
The Australian National University

January 2015

mona.golestanfar@anu.edu.au 2015

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

Mona Glestan Far

mona.golestanfar@anu.edu.au

19 January 2015

to my xxx, yyy (yyy is the people you want to dedicated this thesis to.)

Acknowledgments

Who do you want to thank?

Abstract

Put your abstract here.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The patent system is designed to encourage disclosure of new technologies and novel ideas by granting exclusive rights on the use of inventions to their inventors, for a limited period of time. An important requirement for a patent to be granted is that the invention, it describes, is novel which means there is no earlier patent, publication or public communication of a similar idea. To ensure the novelty of an invention, patent offices as well as other Intellectual Property (IP) service providers perform searches called âĂŸprior art searchesâĂŹ or âĂŸvalidity searchesâĂŹ. Since the number of patents in a company's patent portfolio affects the company market value, well-performed prior art searches are of high importance [Piroi et al., 2013a].

Patent Retrieval
Evaluation of patent retrieval was proposed in NTCIR-2 in 2001 [Leong, 2001]. Since then patent retrieval has featured as a track in all NTCIR campaigns. The *Clef-Ip Lab* and its tasks have evolved considerably over the last five years, from a rough approximation of a prior art search task in 2009, to, in 2013, a good simulation of the passage-level search carried out by patent searchers [Piroi et al., 2013b]. Patent retrieval is of interest in IR research since it is of commercial interest and is a challenging IR task with different characteristics to popular IR tasks such as precision-orientated ad hoc search on news archives or web document collections. Various patent search tasks have been created in mentioned campaigns including:
Ad-hoc search: A number of topics are used to search a patent collection with the objective of retrieving a ranked list of patents that are relevant to this topic [Iwayama et al., 2003].

Invalidity search. The claims of a patent are considered as the topics, and the objective is to search for all relevant documents (patents and others) to find whether the claim is novel or not [Joho et al., 2010; Fujii et al., 2004]. All relevant documents are needed, since missing only one document can lead to later invalidation of the claim or the patent itself.

Passage Search: The same as invalidity search, but because patents are usually long, the task focuses on indicating the important fragments in the relevant documents [Fujii et al., 2007].

Prior-art Search: This is the main search task carried out in patent offices; it is concerned with finding all relevant patents that are potential to invalidate the novelty of a patent application or at least that have common parts to that patent [Roda et al., 2010]. The full patent application submitted to the patent office is considered as the topic, and patent citations that are identified by the patent office are taken as the relevant documents, therefore the objective is to find these citations of patents automatically. Prior-art search in patent retrieval focuses on finding any kind of patents relevant to the patent application in hand; this is different from invalidity search which focuses on finding any type of document that proves that a given claim in a patent application is not novel.

Main challenges of prior-art search

Query length: The query is a full patent application instead of just keywords which are short. So it is not focused on information need. The problem with a full document as a query is that, it might refer to multiple topics. Even in the case of a single invention, different components of the new device or process which may be described in verbose patent application.

Recall-oriented retrieval task: Prior-art search is a recall-oriented retrieval task, where not missing a relevant document is more important than retrieving a small number of the most relevant documents at the top rank. Usually, in prior-art search, the patent examiners will carefully examine the first 100 or 200 documents retrieved by the search engine instead of browsing just the top few results. Missing one patent could result in a multimillion dollar lawsuit due to a patent infringement [Arampatzis et al., 2007; Magdy and Jones, 2010].

Term Mismatch: The biggest challenge in patent retrieval is the significant term mismatch between the query and relevant document (Due to: Usage of new inventive words, Rewording to avoid repetition, Non-standardized acronyms: invented by authors, Synonyms: signal and wave). Magdy [Magdy et al., 2010] reported from their analysis presented for CLEF-IP 2009 prior-art search that 12% of the relevant patents do not share any terms in common with patent topics after filtering out stop words [Magdy and Jones, 2011].

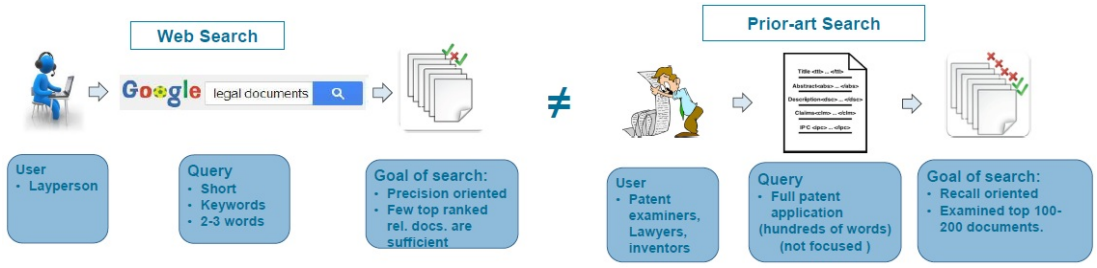

Figure 1.1: The main differences between patent prior-art search and an standard web search are: (1) the user is an expert (professional patent examiner), (2) the query is a full application not just keywords, and (3) it is a recall oriented task.

Problem Statement : Applying standard information retrieval (IR) techniques to patent search is not effective and needs applying supplementary methods to improve its effectiveness. Although lots of methods has been proposed in recent years but still reported results for different tasks of patent search show lower retrieval effectiveness compared to other IR applications [Lupu et al., 2013]. For example, it is generally expected to achieve a mean average precision (MAP) less than 0.1, which is still regarded as an acceptable level of effectiveness. The results of various evaluation campaigns [Lupu et al., 2013; Joho et al., 2010; Roda et al., 2010; Piroi et al., 2012] concluded that patent search task is certainly not a solved problem and many challenges in applying IR solutions in the intellectual property domain remain to be overcome. In this research, we focus on enhancing possible general and patent-specific IR techniques to improve the effectiveness of baseline patent prior-art search.

1.1 Thesis Statement

I believe A is better than B.

1.2 Introduction

Put your introduction here. You could use \fix{ABCDEFG.} to leave your comments, see the box at the left side.

You have to rewrite your thesis!!!

1.3 Thesis Outline

How many chapters you have? You may have Chapter 2, Chapter **??**, Chapter **??**, Chapter **??**, and Chapter 6.

Chapter 2

Background and Related Work

2.1 General Information Retrieval (IR)

In general, an information retrieval system assists users in finding the information they need, Fig. 2.1 illustrates the general IR process. First, a repository of indexed documents is created from a collection of documents to be searched for. Users formulate the information they need as a query. In the matching process, the query and documents representations are compared and the result would be a ranked list of documents.

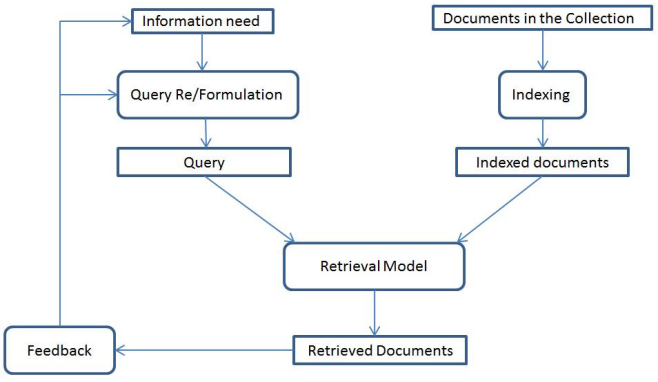

Figure 2.1: Simple illustration of the process in a general IR system.

2.1.1 Retrieval Models

2.1.2 The Study of Retrievablity

2.1.3 Query Expansion (QE)

2.1.4 Query Reduction (QR)

2.1.5 IR Evaluation Metrics

2.2 Patent-specific IR

2.2.1 The Study of Retrievablity for patents

2.2.2 Query Formulation

2.2.3 Query Expansion for Patents

2.2.4 Query Reduction for Patents

2.2.5 The Use of Metadata

2.2.6 Multilinguality

2.2.7 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 [**?**]. Many patent retrieval tasks rerank the top retrieved documents from initial retrieval stage based on additional patent feature [**?**], claim structure [**?**], and considering IPC information of patent and its neighbours to retrieve similar patents [**?**].

2.2.8 Evaluation Metrics for Patent Retrieval

At the begging of each chapter, please introduce the motivation and high-level picture of the chapter. You also have to introduce sections in the chapter.

Section 2.3 xxxx.

Section 2.4 yyyy.

2.3 Motivation

2.4 Related work

You may reference other papers. For example: Generational garbage collection [Lieberman and Hewitt, 1983; Moon, 1984; Ungar, 1984] is perhaps the single most important advance in garbage collection since the first collectors were developed in the early

1960s. (doi: "doi" should just be the doi part, not the full URL, and it will be made to link to dx.doi.org and resolve. shortname: gives an optional short name for a conference like PLDI '08.)

2.5 Summary

Summary what you discussed in this chapter, and mention the story in next chapter. Readers should roughly understand what your thesis takes about by only reading words at the beginning and the end (Summary) of each chapter.

Chapter 3

Baseline Framework

Chapter 4

Term Analysis

4.1 Term Mismatch

4.2 Relevance Feedback

4.2.1 Discriminative Words

4.2.2 RF Optimal Query Formulation

4.2.3 Query Reduction Using RF

4.3 Pseudo Relevance Feedback

4.3.1 PRF Query

4.3.2 Query Reduction Using RF

4.4 Section-based Analysis

4.5 Noisy words

4.5.1 Document Frequent Words

4.5.2 IPC Code Definition

4.6 Summary

Same as the last chapter, summary what you discussed in this chapter and be the bridge to next chapter.

Chapter 5

Improve the Performance by Minimum Users Efforts

Chapter 6

Conclusions

Summary your thesis and discuss what you are going to do in the future in Section 6.4.

6.1 Overview

6.2 Summary

6.3 Contributions

6.4 Future Work

Good luck.

Bibliography

Arampatzis, A.; Kamps, J.; Koolen, M.; and Nussbaum, N., 2007. Access to legal documents: Exact match, best match, and combinations. (2007). (cited on page 2)

Fujii, A.; Iwayama, M.; and Kando, N., 2004. Overview of patent retrieval task at ntcir-4. In *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization*, 232–241. (cited on page 1)

Fujii, A.; Iwayama, M.; and Kando, N., 2007. Overview of the patent retrieval task at the ntcir-6 workshop. In *Proceedings of the Sixth NTCIR Workshop Meeting*, 359–365. (cited on page 1)

Iwayama, M.; Fujii, A.; Kando, N.; and Takano, A., 2003. Overview of patent retrieval task at ntcir-3. In *Proceedings of the ACL-2003 workshop on Patent corpus processing-Volume 20*, 24–32. Association for Computational Linguistics. (cited on page 1)

Joho, H.; Azzopardi, L. A.; and Vanderbauwhede, W., 2010. A survey of patent users: an analysis of tasks, behavior, search functionality and system requirements. In *Proceedings of the third symposium on Information interaction in context*, 13–24. ACM. (cited on pages 1 and 3)

Leong, M., 2001. Patent data for ir research and evaluation. In *Proceedings of the 2nd NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, Tokyo, Japan*, 359–365. (cited on page 1)

Lieberman, H. and Hewitt, C., 1983. A real-time garbage collector based on the lifetimes of objects. *Communications of the ACM*, 26, 6 (Jun. 1983), 419–429. doi: 10.1145/358141.358147. (cited on page 6)

Lupu, M.; Hanbury, A.; et al., 2013. Patent retrieval. *Foundations and Trends in Information Retrieval*, 7, 1 (2013), 1–97. (cited on page 3)

Magdy, W. and Jones, G. J., 2010. Pres: a score metric for evaluating recall-oriented information retrieval applications. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 611–618. ACM. (cited on page 2)

Magdy, W. and Jones, G. J., 2011. A study on query expansion methods for patent retrieval. In *Proceedings of the 4th workshop on Patent information retrieval*, 19–24. ACM. (cited on page 2)

Magdy, W.; Leveling, J.; and Jones, G. J., 2010. Exploring structured documents and query formulation techniques for patent retrieval. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, 410–417. Springer. (cited on page 2)

Moon, D. A., 1984. Garbage collection in a large LISP system. In LFP '84: *Proceedings of the 1984 ACM Symposium on LISP and Functional Programming* (Austin, Texas, USA, Aug. 1984), 235–246. ACM, New York, New York, USA. doi:10.1145/800055.802040. (cited on page 6)

Piroi, F.; Lupu, M.; and Hanbury, A., 2013a. Overview of clef-ip 2013 lab. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, 232–249. Springer. (cited on page 1)

Piroi, F.; Lupu, M.; and Hanbury, A., 2013b. Passage retrieval starting from patent claims. a clef-ip 2013 task overview. In *CLEF (Online Working Notes/Labs/Workshop)*. (cited on page 1)

Piroi, F.; Lupu, M.; Hanbury, A.; Sexton, A. P.; Magdy, W.; and Filippov, I. V., 2012. Clef-ip 2012: Retrieval experiments in the intellectual property domain. In *CLEF (Online Working Notes/Labs/Workshop)*. (cited on page 3)

Roda, G.; Tait, J.; Piroi, F.; and Zenz, V., 2010. Clef-ip 2009: retrieval experiments in the intellectual property domain. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, 385–409. Springer. (cited on pages 2 and 3)

Ungar, D., 1984. Generation scavenging: A non-disruptive high performance storage reclamation algorithm. In SDE 1: *Proceedings of the 1st ACM SIGSOFT/SIGPLAN Software Engineering Symposium on Practical Software Development Environments* (Pittsburgh, Pennsylvania, USA, Apr. 1984), 157–167. ACM, New York, New York, USA. doi:10.1145/800020.808261. (cited on page 6)