Patent Information Retrieval

An Instance of Domain-specific Search

Mihai Lupu
Vienna University of Technology
Favoriten Strasse 9-11/188
Vienna 1040, Austria
lupu@ifs.tuwien.ac.at

ABSTRACT

The tutorial aims to provide the IR researchers with an understanding of how the patent system works, the challenges that patent searchers face in using the existing tools and in adopting new methods developed in academia.

At the same time, the tutorial will inform the IR researcher about the unique opportunities that the patent domain provides: a large amount of multi-lingual and multi-modal documents, the widest possible span of covered domains, a highly annotated corpus and, very importantly, relevance judgements created by experts in the fields and recorded electronically in the documents.

The combination of these two objectives leads to the main purpose of the tutorial: to create awareness and to encourage more emphasis on the patent domain in the IR community. Table 1 provides details on how the tutorial covers the topics of the SIGIR conference.

Categories and Subject Descriptors

J.1 [**Computer Applications**]: Administrative Data Processing—*Law,Business,Government*; A.1 [**General Literature**]: Introduction and Survey

General Terms

Documentation, Human Factors, Languages, Algorithms

Keywords

Patents

1. TUTORIAL DESCRIPTION

1.1 Intended Audience

PhD Students, junior and senior researchers looking for a new challenge for their algorithms. In terms of general IR, the audience can have any level: beginner, intermediate or advanced. In terms of patent IR, the audience is expected to consist of novices in the domain.

1.2 Agenda Details

Introduction (∼30 minutes)

The tutorial starts by introducing the attendees to the patent system. We discuss how it works, international standardization efforts, key players. In doing so, we define some of the terms that we will be using in the rest of this session.

Types of Searches (∼30 minutes)

To understand what patent IR is we must understand the tasks that a professional searcher has to solve. While at first sight this may be simply "search for patents", an in-depth look identifies different search tasks, with different types of requests for information and different expectations in terms of the results.

Evaluation (∼30 minutes)

The different tasks the professional searcher has to accomplish are crystalized in benchmarking tasks. The domain is particularly suited for large-scale IR evaluation campaigns, since search reports are public information and can be used as relevance judgements. We will talk about the pros and cons of this use of search reports, and about how to run your own tests and create your own test collections.

Contents (∼75 minutes)

The largest part of this tutorial will be on the actual content of the patent documents and what we can do with it. First, I will describe the standard metadata available for each patent document. It covers a set of different physical and legal persons, hyper-references between patents, dates, and several types of classifications. From this, we can already extract interesting observation pertaining to the evolution of technologies and their key players.

We can learn even more from the actual text, and I will present here examples of content and describe the current state-of-the-art in using this unstructured information for retrieval. Issues such as the use of different genres and languages, as well as hidden semantic information will be presented.

Finally, I will briefly talk about image search in this domain. The particularity here is that most of the image retrieval methdos, based on colors and textures, are powerless in this domain, where 2-bit color depth images are the rule.

Data Sources (∼15 minutes)

In conclusion, I will summarize the further resources available to the researcher interested in the domain, as well as the sources of actual patent data publicly available.

2. REFERENCES

Table 1, with references for each of its claims, is available at `http://mihailupu.net/patentTutorial`.

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

Table 1: Patent IR touches on all aspects of IR research

What was *your* paper about?	Did you know that... (and would your method still work in this case)?
Document Representation and Content Analysis (e.g., text representation, document structure, linguistic analysis, non-English IR, cross-lingual IR, information extraction, sentiment analysis, clustering, classification, topic models, facets)	• patent documents are highly structured and cover different genres within the same document? • the global patent collection has manually created relevance judgments across languages? • and that it also has an international classification scheme covering all patents? • a patent has been overthrown in 1997 at the USPTO for prior art disclosed in ancient Sanskrit texts?
Queries and Query Analysis (e.g., query representation, query intent, query log analysis, question answering, query suggestion, query reformulation)	• a patent search process always starts with a multi-page document describing the invention? • patent search professionals are experts in creating large queries with both keywords and metadata? • "a system having a storage for storing data, an output device to display information, a terminal for entering information, and a component that modifies the input data and controls flow of data between different parts" is a computer?
Users and Interactive IR (e.g., user models, user studies, user feedback, search interface, summarization, task models, query logs, personalized search)	• a patent search process can take up to five months • the USPTO publishes the examiner's search strategy and results for each application • an examination division at the EPO always consists of three technical examiners?
Retrieval Models and Ranking (e.g., IR theory, language models, probabilistic retrieval models, feature-based models, learning to rank, combining searches, diversity)	• the most popular model among patent searchers is boolean, because it provides clear evidence as to why a document was in the retrieved list or not? • published search reports can be used to learn to rank and provide significant retrieval improvements? • common search strategies involve different features (inventors, owners, classes, references), whose weights need to be balanced?
Search Engine Architectures and Scalability (e.g., indexing, compression, MapReduce, distributed IR, P2P IR, mobile devices)	• the only source for absolutely reliable patent legal status data are the national patent offices • more and more National Patent Offices publish their data online, creating a de-facto distributed repository of patent data?
Filtering and Recommending (e.g., content-based filtering, collaborative filtering, recommender systems, profiles)	• a patent searcher has to cull through thousands or tens of thousands of patents for a validity search
Evaluation (e.g., test collections, effectiveness measures, experimental design)	• there already exist over 5 test collections dedicated to patent search • patent search includes at least 3 different types of search use-cases, for which different effectiveness measures are needed
Web IR and Social Media Search (e.g., link analysis, social tagging, social network analysis, advertising and search, blog search, forum search, CQA, adversarial IR, vertical and local search)	• patents form an extensive 'social' network • the objective of a patent claim is to provide as wide as coverage as possible, while disclosing as little as possible
IR and Structured Data (e.g., XML search, ranking in databases, desktop search, entity search)	• patents are distributed as XML files? • by their definition, patents' core entities have not previously been seen? • entities are the fundamental way to searching chemical patents?
Multimedia IR (e.g., Image search, video search, speech/audio search, music IR)	• there are 9 types of images in patents? • patent images are black-and-white, not even grayscale? • currently, engineering patent searchers have no option but to manually review thousands of images?