Patent Information Retrieval

Mihai Lupu, Allan Hanbury
Vienna University of Technology, Vienna, Austria
{lupu,hanbury}@ifs.tuwien.ac.at

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.

Prerequisite knowledge and/or skills: none.

2 Tutorial Description

Patents are legal documents issued by a government, which grants a set of rights of exclusivity and protection to the owner of an invention. As such, it is one of the main instruments of Intellectual Property protection, and a multi-billion dollar industry world-wide. For the IR researcher, it is an opportunity to investigate existing techniques and develop new methods for a domain which has IR at its very core. Any new invention or innovation, as well as the work of all academic or industry researchers, and of all patent searchers in every patent office across the world, revolves around the need to know exactly what has already been done and published before.

The total number of patents in force worldwide at the end of 2008 was approximately 6.7 million, with more that half a million new patents per year[1]. The tutorial will introduce the participants to the specificities of the patent domain, its rules and customs, and how patents are formed as documents.

For example, a patent requires the invention to be publicly disclosed, but because public disclosure may be against the economic interests of the owner of the invention, the accessibility of the patent's content depends on the outcome of a struggle between the applicants and the examiners at the patent offices. The typical outcome is a text that indeed describes the invention, but still remains inscrutable to most people—and hard to handle for off-the-shelf indexing methods—due to a large amount of neologisms and intentionally over-generalized expressions. Experiments have shown as many as 12% of the documents relevant to a topic did not have any word in common with the topic, and that the cosine measure was higher for non-relevant documents than for relevant ones [14]. Legal requirements in various countries lead to a multitude of other idiosyncrasies, such as very long sentences caused by a requirement to express something in one sentence only [1].

The patenting process requires the documents to adhere to a certain structure, as well as the creation of a substantial amount of metadata, such as, for example, several classification hierarchies for the subject domains. The tutorial will explain in which sense these elements are important for patent professionals and how they may be of use to IR researchers. Among these, of particular interest are the search reports created by experts in the patent offices during the review process, as they contain manually produced relevance judgements.

Such relevance judgements are an important source of information, but they must also be considered carefully, as different forms of patent search have potentially different sets of relevant documents. State of the Art Search, Evidence of Use Search, Pre-Filing Patentability Search,

[1] http://www.wipo.int/ipstats/en/statistics/patents/

Patentability or Novelty Search, Clearance or Freedom to Operate Search, and Validity or Invalidity Search are all different activities that a professional searcher engages in, and a search system may target one or several of these forms of search.

The tutorial will also introduce the participants to the evaluation of IR engines for the patent domain. The requirements of the professional users in this area are quite different from those of the general public, and we will discuss the use of Cranfield-based evaluation methods as well as the experience of more interactive evaluation efforts, such as the PatOlympics[2].

As a key form of Intellectual Property protection, patents represent a valueable source of scientific and technology information, extremely important for the industry, but under-analyzed in academia. The objectives of this tutorial, as outlined next, focus on changing this.

2.1 Objectives and Relevance

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. The tutorial will cover the full spectrum of IR research and its applications / implications for patent IR, as demonstrated in the following:

IR sub-field	**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? [17] • the global patent collection has manually created relevance judgments across languages? [20] • and that it also has an international classification scheme covering all patents? [5] • a patent has been overthrown in 1997 at the USPTO for prior art disclosed in ancient Sanskrit texts? [8]
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? [21] • 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? [2]
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 [7] • the USPTO publishes the examiner's search strategy and results for each application[3] • an examination division at the EPO always consists of three technical examiners? [4]

[2] http://www.ir-facility.org/events/irf-symposium/irf-symposium-2011/patolympics

[3] http://portal.uspto.gov/external/portal/pair

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? [9] • published search reports can be used to learn to rank and provide significant retrieval improvements? [10] • common search strategies involve different features (inventors, owners, classes, references), whose weights need to be balanced? [11]
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 [15] • 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 [7]
Evaluation (e.g., test collections, effectiveness measures, experimental design)	• there already exist over 5 test collections dedicated to patent search [16][19][12] • patent search includes at least 3 different types of search use-cases, for which different effectiveness measures are needed [2]
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 [24] • 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? [22] • by their definition, patents’ core entities have not previously been seen? • entities are the fundamental way to searching chemical patents? [6]
Multimedia IR (e.g., Image search, video search, speech/audio search, music IR)	• there are 9 types of images in patents? [18] • patent images are black-and-white, not even grayscale? [23] • currently, engineering patent searchers have no option but to manually review thousands of images? [3]

3 Agenda

1. **Session 1 : 09:30-11:00**
   (a) **Introduction** The patent domain - how it works, international standardization efforts, organizations (30 minutes)

   (b) **Metadata** Social networks of inventors and assignees, page rank information (15 minutes)

   (c) **Full Text** Content, genres, multilinguality (45 minutes)

2. **Session 2 : 11:30-13:00**
   (a) **Non-text data in patents** (30 minutes)

   (b) **Types of searches** Use-cases, examples of queries from the USTPO (30 minutes)

(c) **Evaluation** Cranfied and non-Cranfield in the patent domain (30 minutes)

4 Contact and Biography

Mihai Lupu obtained his PhD degree under the Singapore-MIT Alliance at the National University of Singapore, where he researched data retrieval in peer-to-peer networks. Now, his research interests continue in the area of information retrieval, with emphasis on patent retrieval and evaluation of domain-specific IR methods. Between 2009 and 2011, he has been the co-organizer of the TREC-CHEM evaluation campaign and the PaIR workshop series. He is co-editor of the recent book on *Current Challenges in Patent Information Retrieval* (Springer, Information Retrieval Series, 2011 [13])

Allan Hanbury is Senior Researcher at the Vienna University of Technology, Austria. He is scien- tific coordinator of the EU-funded KHRESMOI Integrated Project on medical information search and analysis. He is also co-organiser of the CLEF-IP evaluation track. He is a member of the PROMISE Network of Excellence on IR evaluation, the MUMIA COST Action on multilingual and multifaceted interactive information access, and the CEEPUS (Central European Exchange Program for University Studies) Network on image processing, information engineering & interdisciplinary knowledge exchange. For the latter network, he is co-organising a summer school in July 2012 focussing on medical information analysis. He was leader of the Evaluation, Integration and Standards work package of the MUSCLE EU Network of Excellence, and has led a number of Austrian national projects. His research interests include information retrieval, multi-modal information retrieval, and the evaluation of information retrieval methods. He is author or co-author of over 70 publications in refereed journals and international conferences.

They can be contacted at:

```
Vienna University of Technology
FavoritenStrasse 9-11/188
Vienna, 1040, Austria
+43-1-58801 188314 (Mihai Lupu)
+43-1-58801 188310 (Allan Hanbury)
{lupu,hanbury}@ifs.tuwien.ac.at
```

References

[1] Manual of Patent Examination Procedure, Section 608.01(m), Revision July 2010. http://www.uspto.gov/web/offices/pac/mpep/index.htm.

[2] D. Alberts, C. B. Yang, D. Fobare-DePonio, K. Koubek, S. Robins, M. Rodgers, E. Simmons, and D. DeMarco. *Current Challenges in Patent Information Retrieval*, chapter 1 : Introduction to Patent Searching - Practical Experience and Requirements for Searching the Patent Space. Springer Verlag, 2011.

[3] D. DeMarco and A. Davis. Mechanical patent searching: A moving target. In *Proc. of PIUG Annual Conference*, 2010.

[4] European Patent Office. *Guidelines for Substantive Examination*, 2010.

[5] C. G. Harris, R. Arens, and P. Srinivasan. Using classification code hierarchies for patent prior art searches. In M. Lupu, K. Mayer, J. Tait, A. J. Trippe, and W. B. Croft, editors, *Current Challenges in Patent Information Retrieval*. Springer, 2011.

[6] J. D. Holliday and P. Willet. *Current Challenges in Patent Information Retrieval*, chapter 17: Representation and Searching of Chemical-Structure Information in Patents. Springer Verlag, 2011.

[7] H. Homan. Making the case for patent searchers? *Searcher*, 2004.

[8] Indian Ministry of Environment and Forests. Know instances of patenting on the ues of medicinal plants in india. http://pib.nic.in/newsite/erelease.aspx?relid=61511, 2010.

[9] Y. Kim, J. Seo, and W. B. Croft. Automatic boolean query suggestion for professional search. In *Proc. of SIGIR*, 2011.

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

[11] P. Lopez and L. Romary. PATATRAS: Retrieval Model Combination and Regression Models for Prior Art Search. In C. Peters, G. Di Nunzio, M. Kurimo, T. Mandl, D. Mostefa, A. Peñas, and G. Roda, editors, *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, volume 6241 of *LNCS*. Springer, 2010.

[12] M. Lupu, J. Huang, J. Zhu, and J. Tait. TREC Chemical Information Retrieval - An Evaluation Effort for Chemical IR Systems. *WPI Journal*, 2011.

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

[14] W. Magdy, J. Leveling, and G. J. F. Jones. DCU @ CLEF-IP 2009: Exploring standard IR techniques on patent retrieval. In *Workshop of the CLEF 2009, Revised Selected Papers*, LNCS 6241, 2010.

[15] H. Moohan. Report of the superworkshop for expert users. In *Proc. of the Patent Information Conf.*, 2011.

[16] H. Nanba, A. Fujii, M. Iwayama, and T. Hashimoto. Overview of the Patent Retrieval Task at the NTCIR-8 Workshop. In *Proc. of NTCIR-8*, 2010.

[17] N. Oostdijk, E. D'hondt, H. van Halteren, and S. Verberne. Genre and domain in patent texts. In *Proc. of PaIR*, 2010.

[18] F. Piroi, M. Lupu, A. Hanbury, and V. Zenz. Clef-ip 2011: Retrieval in the intellectual property domain. In *CLEF (Notebook Papers/Labs/Workshop)*, 2011.

[19] F. Piroi and J. Tait. Clef-ip 2010: Retrieval experiments in the intellectual property domain. In *Proc. of CLEF*, 2010.

[20] G. Roda, V. Zenz, M. Lupu, K. Järvelin, M. Sanderson, and C. Womser-Hacker. So Many Topics, So Litlle Time. *SIGIR Forum*, 43(1), 2009.

[21] J. Tait, M. Lupu, H. Berger, G. Roda, M. Dittenbach, A. Pesenhofer, E. Graf, and K. van Rijsbergen. Patent Search: An Important New Test Bed for IR. In *Proc. of DIR*, 2009.

[22] F. Versloot. The data and their coverage in the epo's collections. In *Proc. of Patent Information Conference*, 2011.

[23] S. Vrochidis. Towards patent image retrieval. In *Proc. of ICIC*, 2009.

[24] X. Wang, X. Zhang, and S. Xu. Patent co-citation networks of fortune 500 companies. *Scientometrics*, 88:761–770, 2011. 10.1007/s11192-011-0414-x.