Inf Retrieval (2014) 17:407–411
DOI 10.1007/s10791-014-9245-8

Guest editorial: Special issue on information retrieval in the intellectual property domain

Allan Hanbury · Mihai Lupu · Noriko Kando · Barrou Diallo · Stephen Adams

Received: 18 August 2014 / Accepted: 1 September 2014 / Published online: 10 September 2014

1 Introduction

Intellectual property (IP) management consists of all the processes linked to the recognition, the publication and the exploitation of creations of the human mind. Patents on inventions are granted by 74 patent offices worldwide under the condition that the patent claims are new, inventive and show industrial applications. Information retrieval (IR) has been recognized as the computing field appropriate to aid in assessing the novelty aspect of the IP processes by proposing methods to search, compare and analyse patent documents.

Patents are complex search objects. They attempt to describe that most elusive concept, the "invention," in a manner that distinguishes it from what has gone before. Unfortunately for the inventor, many of the modern methods for communicating new technical information, such as video, animations, CAD-CAM displays or even, most recently, 3D printing, are little accepted within the patent system, with the result that many new patent

A. Hanbury (✉) · M. Lupu
Institute of Software Technology and Interactive Systems, Vienna University of Technology, Vienna, Austria
e-mail: hanbury@ifs.tuwien.ac.at

M. Lupu
e-mail: lupu@ifs.tuwien.ac.at

N. Kando
Information and Society Research Division, National Institute of Informatics (NII), Tokyo, Japan
e-mail: noriko.kando@nii.ac.jp

B. Diallo
European Patent Office, Rijswijk, The Netherlands
e-mail: bdiallo@epo.org

S. Adams
Magister Ltd., Lostwithiel, UK
e-mail: stephen.adams@magister.co.uk

applications struggle to render the details of a complex innovative product or process through the medium of the written word. The last 10 years has seen substantial improvements in the proportion of patents that has been digitised for electronic searching purposes, and this has made it possible, albeit a challenge, to introduce more information technology support into existing workflows. Indeed, professional patent searchers are still using—and obtaining high-quality, reproducible results from—search tools which have changed little over the last 40 years. With the exception of specialist techniques for handling 2D chemical and biochemical structures, current tools pay little attention to the complexity of patent text.

Patents play a major role in society as they are meant to encourage innovation and invention globally. As a consequence, companies are filing increasing numbers of patents in order to protect their intangible assets and the value of their research—after a drop of 3.9 % in the number of patent application filings worldwide in 2009 compared to 2008, they have again been increasing by around 7.6 % in 2010, 8.1 % in 2011 and reached an increase of 9.2 % in 2012. This corresponds to an estimated 2.35 million patent filings worldwide in 2012 (WIP 2013).

The impact of IR on the economic world, through its influence on the validation and the enforcement of patents is potentially immense. Indeed, the whole purpose of defining the novelty of an invention is to be able to get a clear overview of the technological landscape at the date of the patent filing. IR offers the means to find and compare digital artefacts, in particular text documents such as patents. Not limited to text chunks, IR gives also the possibility to compare images such as technical figures or flowcharts, where inventions are explained. This gives an increasing role for multimedia IR in retrieving prior art documents. Additionally, novelty is not limited to a country or a region. Access to collections of documents (scientific articles, standards, reports, and academic theses) through the internet offers today the possibility of getting a more comprehensive overview of the state-of-the-art. Again, IR has played a fundamental role in society by breaking the language barriers in both the querying aspects and the understanding of the search results in different languages. The state-of-the-art of IR for patents is described in a recent overview of patent retrieval (Lupu and Hanbury 2013).

In summary, patent retrieval and intellectual property specialists in the twenty-first century face many challenges. They must search very large numbers of documents expressing complex technological concepts through sophisticated legal clauses in multiple languages. Despite a great deal of theoretical development in IR techniques and machine translation approaches, advanced search tools for patent professionals are still in their infancy. IR still has a large potential to speed up and make more effective the whole chain of processes involved in IP.

2 The special issue

Patent information retrieval is a cross cutting research area as it contains domains such as text retrieval, multilingual information retrieval, image retrieval, natural language processing, and text categorisation. This special issue presents research results on topics related to IR in the intellectual property domain in order to advance the current state-of-the-art of patent search tools. To the reader who may be unfamiliar with the challenges facing the patent world, this collection of papers is a good introduction to the range of issues which face both patent office and industrial searchers.

The previous special issue on patent processing was published in Information Processing and Management in 2007 (Fujii et al. 2007). A significant amount of work has been done in the area of IR in the intellectual property domain since then, spurred on by the evaluation campaigns and workshops held in the last 6 years, most notably:

- The CLEF-IP[1] Evaluation Campaign (2009–2013). The main focus of this evaluation campaign was prior art search, beginning with full patent document retrieval and shifting over the years toward retrieving relevant passages from patents. Additional tasks on patent image analysis and patent classification were also organised.
- The NTCIR[2] patent tasks (2001–2013). The tasks focussed on patent retrieval in the beginning, shifting toward a focus on machine translation in the end. Patent classification, passage retrieval and technical trend map creation tasks were also organised.
- The TREC-CHEM[3] Track (2009–2011). This track focussed on the creation of technology surveys in the chemical domain, with additional tasks in the later years on chemical structure recognition.
- The series of Patent Information Retrieval (PaIR) Workshops[4] (2008–2011), collocated with the Conference on Information and Knowledge Management (CIKM).

Six of the papers in this special issue use resources made available by the CLEF-IP Evaluation Campaign.

3 Paper overview

Seven papers were accepted for publication in this special issue after a thorough review by at least three reviewers each. These papers recognize at least two major factors which impact the work of the patent searcher. The first factor is the multimodal and multilingual nature of the data to be searched, while the second is the process of developing an appropriate strategy for searching the data, which includes query expansion and optimisation steps. Three of the papers consider the use of query expansion in improving patent retrieval results, while one paper considers the use of multiple query representations for the same purpose. The remaining three papers cover the areas of text categorization, image retrieval and multilingual retrieval for patents. We now briefly summarize the papers in this special issue, highlighting their potential contributions to advancing patent search in practice.

We begin by considering the papers related to search strategy. One of the main problems in patent retrieval is vocabulary mismatch—in different patents, different words or phrases are often used to describe similar concepts. This is to a small extent due to a conscious effort on the part of the patent authors to make patents less findable, but mostly because patents represent technological advances for which the vocabulary is not yet fixed. For this reason, the use of carefully crafted lists of synonyms in queries to patent search systems is an important aspect of professional patent search. Three of the papers deal with the use of automated *query expansion* to solve the vocabulary mismatch problem. Note that query expansion seems to have the connotation that "more is better" when developing a

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

[2] http://research.nii.ac.jp/ntcir/index-en.html.

[3] http://www.ir-facility.org/trec-chem.

[4] http://www.ir-facility.org/pair-workshops.

search string for retrieval, whereas professional searchers are aware that the optimum strategy applied to current search tools may often be quite concise: the most important step is choosing the *most appropriate* search terms, not simply identifying additional ones. The three papers in this category each use a different source of information for the query expansion: patent citation graphs, Wikipedia and query logs of patent examiners. The paper by Parvaz Mahdabi and Fabio Crestani uses the citation links in a pseudo relevance result set to build a topic dependent citation graph. The term distribution of the documents in the citation graph is used to build a query model, which is used to expand the original query. The paper by Bashar Al-Shboul and Sung-Hyon Myaeng uses primary and secondary categories as well as titles of Wikipedia pages for query expansion, with a particular focus on the use of phrases instead of words. Finally, Wolfgang Tannebaum and Andreas Rauber use query logs from patent searchers at the United States Patent and Trademark Office (USPTO) to create term networks of synonyms for several USPTO patent classes, which are then used for automated query expansion.

The paper by Dong Zhou et al. also examines manipulating the query to improve patent retrieval, but uses *multiple query representations* instead of query expansion. Three query representations are automatically generated from a query patent document and submitted to the same search engine. The list of search results produced by each query is treated as a collection of 'ratings,' which are analyzed using collaborative filtering algorithms and finally merged into a single document ranking.

Cross-language information retrieval (CLIR) is an essential tool for patent retrieval, as patent documents are written in multiple languages, and are eligible as prior art irrespective of the language used. The paper of Walid Magdy and Gareth Jones presents an adapted machine translation (MT) specifically designed for CLIR. As the translations of query text used for CLIR do not have to be morphologically and syntactically correct, an MT system is trained on text which has been pre-processed with stop word removal and stemming. This leads to a significant decrease in the MT computational and training resource requirements. This is a promising area, although the most pressing challenge for the professional searcher is in the Asian languages rather than French and German.

Patent categorisation is generally used to automatically place a patent into categories of a hierarchical patent classification system, such as the international patent classification (IPC) or the recently adopted cooperative patent classification (CPC). This categorisation has an application in patent offices to efficiently direct new patent applications to the most competent examiners. The paper by Eva D'hondt et al. examines the impact of concept drift over time on patent categorization accuracy, where a temporal mismatch between training data and newer patent documents to be categorized could lead to a deterioration in accuracy. It proposes methods for training data selection and combination of text representations to improve patent categorization. Eventually, such an approach could lead to better ways of retrieving from the older literature, which is still legally part of the state-of-the-art for every new patent search.

To achieve the necessary discrimination between a new invention and the prior state of the art, modern search must be able to take account of everything known about the content of the patent databases. One way forward is to develop better ways of exploiting the relatively small, but highly significant, volume of non-text information as part of the search process. There are situations where the key information is not in the text of the patent, but in the patent drawings, where it cannot be located by text retrieval approaches. The paper of Marçal Rusiñol et al. presents a method for processing a specific type of patent image. It converts flowchart images into a structured textual representation that can be queried using

text retrieval approaches or compared to each other using various metrics for structural similarity.

4 Conclusion

This special issue presents seven papers proposing solutions to challenges in patent retrieval. The majority of the papers focus on the manipulation of queries to improve search results, with further papers covering the issues of cross-language retrieval, categorisation and image search.

It is interesting to note that none of the evaluation campaigns in the intellectual property domain are still active. This is however not due to all challenges in this domain being solved, as is clear from the papers in this special issue and from the less than optimum results of the evaluation campaigns. The collections and tasks that are still available as a result of the evaluation campaigns continue to be a fertile ground for innovative developments in Information Retrieval.

There are also some challenges in retrieval in the intellectual property domain that are not tackled by any of the papers in this special issue. This includes (Lupu and Hanbury 2013): *information fusion* to take into account in a unified way the patent text, drawings and metadata; *federated search* to allow searching of disparate sources of information that could constitute prior art; and *collaborative search* to take into account that patent searching is often done by a group of people, such as an information specialist collaborating with a technological domain specialist and a patent lawyer. It is also necessary to tackle the human factor in patent searching—that point where searcher and machine combine.

From the searcher's point of view, patent information exists primarily as a resource to inform commercial decisions. Research such as that described in this special issue could help ultimately to improve the accessibility of patent information to all interested parties.

Acknowledgments The guest editors would like to thank the reviewers for their efforts in reviewing papers submitted to the special issue.

References

WIPO Economics and Statistics Series. (2013) *World intellectual property indicators*. WIPO publication no. 941E/2013.

Fujii, A., Iwayama, M., & Kando, N. (2007). Introduction to the special issue on patent processing. *Information Processing and Management*, *43*(5), 1149–1153. doi:10.1016/j.ipm.2006.11.004.

Lupu, M., & Hanbury, A. (2013). Patent retrieval. *Foundations and Trends in Information Retrieval*, *7*, 1–97. doi:10.1561/1500000027.