Published in Studies in health technology and informatics, 2012, vol.180, p. 204-209 which should be cited to refer to this work

An advanced search engine for patent analytics in medicinal chemistry

Emilie PASCHE[a,1,2], Julien GOBEILL[b,2], Douglas TEODORO[a], Arnaud GAUDINAT[b], Dina VISHNYAKOVA[a], Christian LOVIS[a] and Patrick RUCH[b]
[a] *SIMED, University Hospitals of Geneva and University of Geneva, Switzerland*
[b] *BiTeM, Information Sciences Department, University of Applied Sciences, Geneva, Switzerland*

Abstract. Patent collections contain an important amount of medical-related knowledge, but existing tools were reported to lack of useful functionalities. We present here the development of TWINC, an advanced search engine dedicated to patent retrieval in the domain of health and life sciences. Our tool embeds two search modes: an ad hoc search to retrieve relevant patents given a short query and a related patent search to retrieve similar patents given a patent. Both search modes rely on tuning experiments performed during several patent retrieval competitions. Moreover, TWINC is enhanced with interactive modules, such as chemical query expansion, which is of prior importance to cope with various ways of naming biomedical entities. While the related patent search showed promising performances, the ad-hoc search resulted in fairly contrasted results. Nonetheless, TWINC performed well during the Chemathlon task of the PatOlympics competition and experts appreciated its usability.

Keywords. Information retrieval, Patent, Ad-hoc search, Related article search

Introduction

In recent years, patent collections have greatly increased, forming in 2009 a set of more than 50 millions of patents [1]. These collections can be considered as an important and high quality source of knowledge as they contain not only detailed and validated information, but also because by definition their content is rarely found elsewhere. Indeed, it has been shown that a significant number of patents contain unique information not available in any other source [2]. Moreover, such collections are of great interest for biomedical research. Indeed, among the chemistry-related patent applications filed between 2005 and 2009, over 150'000 patent applications concerned the biotechnology field and over 310'000 patent applications were related to pharmaceuticals [3]. Therefore the use of such corpus is essential to information retrieval in biomedical domain.

[1] Corresponding Author. Emilie Pasche, University Hospitals of Geneva, Division of Medical Information Sciences, Rue Gabrielle-Perret-Gentil 4, 1211 Geneva 14, Switzerland; E-mail: emilie.pasche@hcuge.ch.
[2] These two authors contributed equally to the development of the TWINC application.

Nevertheless, studies focused on features requirements for patent search [4] showed the frequent lack of useful functionalities in existing tools. Moreover, nomenclatures, such as genes' and drugs' names, contain many naming ambiguities and synonyms. For example, the drug *acetaminophen* contains 17 additional names in the Medical Subject Headings (MeSH) terminology, such as *paracetamol*. Therefore, to be effective, a tool should at least be able to deal with synonyms.

Thus, several information retrieval (IR) campaigns have emerged promoting the development of IR systems based on patent collections. The Text Retrieval Conference (TREC) has set up the track TREC-Chem [5] proposing two search tasks: a Technical Survey (TS) task and a Prior Art (PA) search task. While during 2009 and 2010, the focus was put on the general chemistry domain, the TREC-Chem 2011 track focused on biomedical and pharmaceutics requirements. This competition provides both a quantitative assessment, but also a comparison of the different search engines developed. At the same time, the PatOlympics competition [6] has explored the development of interface-based tools and the ChemAthlon task has provided a qualitative assessment of chemistry-related patent search.

We have developed TWINC (To WIN ChemAthlon), a web-based interactive and user-friendly application dedicated to chemistry-based patent search to assist life and health specialists. Our tool provides two search modes: an ad-hoc search, to retrieve a set of documents that best fulfill an information need given a short query composed of a few keywords; and a related patent search, to retrieve a set of similar patents given a query composed of several paragraphs extracted from a patent. It also includes three main interactive features: an International Patent Classification (IPC) classifier to automatically attribute IPC codes to a query [7]; a chemical query expansion to retrieve additional way of naming the chemical terms present in the query; and a relevance feedback feature to refine the query based on relevant results. This paper describes the main features of TWINC.

1. Data and Methods

1.1. Data

A collection of 1.3 millions of patents related to chemistry is provided by the TREC campaign [8]. This collection comes from various patent organizations such as EPO (European Patent Office), USPTO (United States Patent and Trademark Office) and WIPO (World Intellectual Property Organization) patent offices.

Ad hoc search is evaluated using a set of six topics (Figure 1) defined by means of natural language by TREC evaluators for the TS task of TREC-Chem 2011. Relevance judgments for each topic are obtained by a stratified sampling approach.

Acetylcholinesterase inhibitors is a potential target for Alzheimer's disease so identifying potent inhibitors of this human enzyme may lead to new treatments of this devastating disease

Figure 1. Example of an ad-hoc search topic related to biomedical domain

Related patent search is evaluated using a set composed of 1000 topics extracted from patent applications and created for the PA task of TREC-Chem 2011. For that

particular task, relevance judgments are constructed based on the original citations of the patents used as topics.

TWINC, our patent search tool, is qualitatively assessed based on two sets of three queries authored by professional patent officers for the ChemAthlon 2010 and 2011 experiments. The experts define the relevance judgments during the live session [6].

1.2. Methods

In this section, we present the pipeline of the patent search approach. We also describe the methods of the chemical query expansion module. Finally, methods to assess the TWINC application are exposed.

Both ad-hoc search and related patent search follow a similar pipeline of three steps. First, the collection is pre-processed, consisting in both the selection of relevant sections and on the normalization of the patent content. Based on the results obtained in the CLEF-IP 2009 [9], only title, abstract and claims are selected for indexing. Patent content normalization is performed using the MeSH terminology. Second, a set of relevant documents is retrieved using the Terrier search engine with the weighting schema BM25. Third, the results are post-processed relying on the re-ranking of the results based on the co-citations networks, where a patent highly cited by other patents will be re-ranked higher. More detailed information on the tuning can be found in [10].

The chemical query expansion feature takes place in a three-stages pipeline. First, the boundaries of chemical terms are detected using an open source chemical Named Entity Recognition tool, called Oscar-3 [11]. Second, a MeSH categorizer [12] normalizes these terms and maps them to unique identifiers. Third, we perform the query expansion by adding synonyms found in different thesauri, including descriptors available in terminological resources such as MeSH or PubChem.

The TWINC's Graphical User Interface is developed with Flex technology and includes the features described above. Usability is assessed during the ChemAthlon task of PatOlympics and consists of three sessions of 20 minutes each. During each session, a different intellectual property expert, assisted by a member of our team, queries the system with one of the topics. He evaluates the quality of the results returned and can submit up to 200 documents. The tool performance is evaluated with two criteria: the number of relevant documents among the up-to-200 submitted documents (so-called recall) and the global appreciation of the system by the experts (so-called user-happiness).

2. Results

The ad hoc search shows a baseline mean average precision (MAP) of 6.3%. When the MeSH normalization is performed prior to indexing, the MAP raises up to 8.9%. Compared to other systems, our approach was ranked as last out of four participants for the ad hoc search task. However, the query sample, consisting of 6 topics, was regarded too small to derive statistical significance difference in the results.

The related patent search with baseline settings obtains a MAP of 5.9%, while the MeSH normalization results in a decrease of the MAP down to 3%. Thus it was decided to disregard MeSH normalization for this task. On the opposite, the co-citations networks strategy reports a positive impact with a MAP of 8.2%. Our approach obtained top-performing results, as we were ranked first out of two

participants in TREC-Chem 2011, confirming results obtained in TREC-Chem 2009, where we were ranked first out of eight participants (Figure 2).

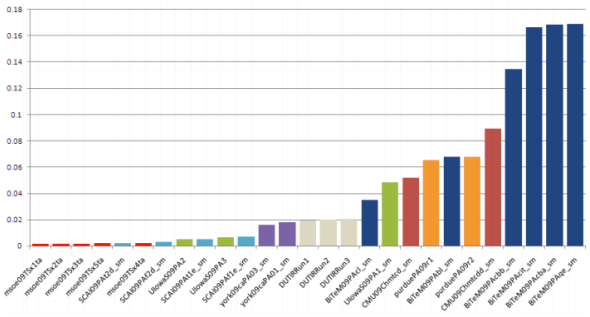

Figure 2. MAP of the runs of the participants of TREC-Chem 2009 (courtesy of M. Lupu). TWINCs' runs are depicted by dark blue color.

The chemical expansion feature was not evaluated during the PA task of TREC-Chem 2011, but during TREC-Chem 2009. It has showed a modest improvement of MAP from 17.9% to 18.2% (+1.7%).

The TWINC application is freely available for non-commercial use at http://casimir.hesge.ch/TWINC (Figure 3). Evaluated within the PatOlympics competition (Table 1), we obtained two successive years the jury's choice out of five participants for its usability. Concerning the performances, we were ranked first out of two teams in 2010, and second in 2011, again with two participating teams.

Figure 3. Prototype of TWINC

	2010		2011	
Teams	**Relevant documents**	**User happiness**	**Relevant documents**	**User happiness**
BiTeM	55	4	55	4.66
Spinque	12	2.33	-	-
CMU	-	-	75	3.33

Table 1. Results of ChemAthlon 2010 and 2011

3. Discussion

We obtained in TREC-Chem 2011 competitive results for the related patent search, but less encouraging results for the ad-hoc search. The re-ranking strategy based on citations seems to be effective. Nevertheless, these results should be balanced by the very limited set of concurrent teams.

Regarding the chemical query expansion, the improvement obtained by our approach is quite humble (+1.7%). But this feature was evaluated on the PA task, during which the MeSH normalization strategy, which has a similar aim, also resulted in contrasted results. Thus, we can expect that this feature will have more impact for the TS task. Indeed, while PA topics are constituted of a large portion of text, query expansion can result in an overload of synonyms bringing a significant noise to the system. Moreover, a large text will probably already contain a subset of the synonyms, while a short query will basically contain only one form of a chemical entity. Thus short topics from the TS task will probably benefits more from this approach.

We have thus developed a user-friendly tool matching some of the requirements of experts for biomedical IR in patent collections. Indeed, it is of major importance to provide tools to deal with various ways of naming chemical entities, as it will help different communities of users to communicate. Indeed, while a drug can be named with its substance (e.g. *fosfomycin*) or with its brand name (e.g. *Monuril*), a biochemist could potentially failed to retrieve patents of interest if he only searches with the substance.

Acknowledgements. The DebugIT project (http://www.debugit.eu) is receiving funding from the European Community's Seventh Framework Programme under grant agreement n°FP7-217139, which is gratefully acknowledged. The EAGLi question-answering framework has been initiated thanks to the SNF Grant # 325230-120758.

References

[1] Bonino D, Ciaramella A, Corno F. Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics. World Patent Information. 2010 March; 32(1):30-38.
[2] Hunt D, Nguyen L, Rodgers M. Patent searching: tools and techniques. New Jersey: John Wiley and Sons; 2007.
[3] Economics and Statistics Division, WIPO. World Intellectual Property Indicators, 2011 edition. 2011.
[4] Azzopardi L, Joho H, Vanderbauwhede W. A Survey on Patent Users Search Behavior, Search functionality and System requirements. IRF Report. 2010.
[5] Lupu M, Piroi F, Huang XJ, Zhu J, Tait J. Overview of the TREC 2009 Chemical IR Track. In Proceedings of the Text REtrieval Conference. 2009.
[6] Lupu M. PatOlympics : an infrastructure for interactive evaluation of patent retrieval tools. In Proceedings of the DESIRE '11 Conference. 2011.
[7] Teodoro D, Gobeill J, Pasche E, Ruch P, Vishnyakova D, Lovis C. Automatic IPC encoding and novelty tracking for effective patent mining. In Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies. 2010:309-317.
[8] Lupu M, Gurulingappa H, Filippov I, Jiashu Z, Fluck J, Zimmermann M, Huang J, Tait J. Overview of the TREC 2011 Chemical IR Track. In Proceedings of the Text REtrieval Conference. 2011.
[9] Gobeill J, Teodoro D, Pasche E, Ruch P. Exploring a Wide Range of Simple Pre and Post Processing Strategies for Patent Searching in CLEF-IP 2009. CLEF. 2009.
[10] Gobeill J, Gaudinat A, Pasche E, Teodoro D, Vishnyakova D, Ruch P. BiTeM group report for TREC Chemical IR Track 2011. In proceedings of the Text Retrieval Conference. 2009.
[11] Corbett P, Murray-Rust P. High-Throughput Identification of Chemistry in Life Science Texts. Computational Life Sciences II. 2006:107-118.

[12] Ruch P. Automatic assignment of biomedical categories: toward a generic approach. Bioinformatics. 2006;22(6):658-664.