A user-friendly interface for health-related patent retrieval

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, Geneva, Switzerland*
[b] *BiTeM, Information Science Department, University of Applied Sciences, Geneva, Switzerland*

Abstract. Health-related information retrieval is complicated by the variety of nomenclatures available to name entities, since different communities of users will use different ways to name a same entity. We present in this report the development and evaluation of a user-friendly interactive web application aiming at facilitating health-related patent search. Our tool, called TWINC, relies on a search engine tuned during several patent retrieval competitions, enhanced with intelligent interaction modules, such as chemical query normalization and expansion. While the functionality of related article search showed promising performances, the ad hoc search results in fairly contrasted results. Nonetheless, TWINC performed well during the PatOlympics competition and was appreciated by intellectual property experts. This result should be balanced by the limited evaluation sample. We can also assume that it can be customized to be applied in corporate search environments to process domain and company-specific vocabularies, including non-English literature and patents reports.

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

Introduction

Over the last decades, the corpus of patents' publications have greatly increased and in 2009 contained over 50 millions of patents [1]. Such collection provides an important and high-qualified source of knowledge by the fact that it contains exclusive, detailed and validated information. It has been shown that a significant number of patents contains unique information not available in other sources [2]. A subset of these patents presents an important interest for the biomedical fields, such as patents related to drugs. Therefore the use of such corpus is essential for information retrieval in biomedical domain.

Nevertheless, several studies [3] have focused on features requirements for patent search and showed obvious lack of useful functionalities in current tools. To be effective, a tool should at least be able to deal with different ways of naming entities,

[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.

while different communities of users will use different nomenclatures. For example, physicians could name drugs with commercial names, while chemists would preferably use chemical formulas and chemical identifiers.

To promote the development of information retrieval systems based on patent corpus and allow their evaluation, several competitions have emerged. The Text REtrieval Conference (TREC) has set up the TREC-Chem track [4] proposing a Prior Art task and a Technical Survey search task. While during the first two years, the focus was put on chemistry, the TREC-Chem 2011 track turned to biomedical and pharmaceutics subjects. In parallel, the competition PatOlympics [5] has been focusing on the development of interface-based tools and has been providing, among others, a qualitative assessment of patent search in the domain of the chemistry

We have developed TWINC, a web-based interactive and user-friendly tool dedicated to patent search to assist life and health specialists. It provides two search modes: ad hoc search, to retrieve a set of documents that best fulfill an information need; and related article search, to retrieve a set of related patents. It also includes three main interactive features: an International Patent Classification (IPC) classifier to automatically attribute IPC codes to a query [6]; a chemical query expansion to cope with various naming entities issue; and a Rocchio feature to refine the query according to relevant results. In this report, we present the main features of TWINC.

1. Data and Methods

1.1. Data

A collection of patents is provided by the TREC campaign [4]. This collection contains 1.2 millions of patents related to chemistry from EPO (European Patent Office), USPTO (United States Patent and Trademark Office) and WIPO (World Intellectual Property Organization) patent offices.

A set composed of 1000 topics is provided for the Prior Art (PA) task of the TREC-Chem 2009 track. Relevance judgments are constructed based on the original citations of the patents used as topics.

A second set of 18 topics, in natural language, (Figure 1) is authored by experts for the Technical Survey (TS) task of the TREC-Chem 2009 track. These topics are accompanied by a set of relevance judgments obtained by stratified sampling approach.

Finally, two sets of 3 queries, one for 2010 and one for 2011, (Figure 2) are defined by IP (Intellectual Property) experts in the field of chemistry to evaluate qualitatively the patent search tools during the chemical track (ChemAthlon) of the PatOlympics competition. The experts define the relevance judgments during the live session [5].

Betaines for peripheral arterial disease
Cardiovascular uses of betaines, especially peripheral arterial disease

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

Bismuth subsalicylate use in stomach relief aids-control of nausea, heartburn, indigestion, upset stomach

Figure 2. Example of a ChemAthlon topic related to biomedical domain

1.2. Methods

1.2.1. Ad hoc search and related article search

Both ad hoc search and related article search follow a similar pipeline of three steps: pre-processing of the collection, information retrieval and post-processing of the results. More detailed information on the tuning can be found in [7]. The pre-processing step consists mainly to select the relevant sections of the patent on which the search will be performed. Based on the results obtained in the Intellectual Property Evaluation Campaign CLEF-IP 2009 [8], only title, abstract, claims and IPC codes are stored for indexing, which is done using the platform Terrier. The information retrieval step is based on the weighting schema Okapi BM25, with the default settings. Finally, the post-processing step re-ranks the results based on two strategies. The first strategy exploits the citation network in order to re-rank the results according to their frequency of citations in other patents. The second strategy is based on the information stored in claims. Results returned in the first place are re-indexed using only the claims, and the query is recomputed, resulting in a new ranking of those results [9]. These two search modes are evaluated respectively with the PA topics and TS topics of the TREC-Chem 2009 track.

1.2.2. Chemical query expansion

The chemical (including drugs) query expansion feature takes place in a three-stages pipeline. First, Oscar3 [10], an open source Named Entity Recognition tool dedicated to chemistry, detects the boundaries of the chemical entities. Then, a Medical Subject Headings (MeSH) categorizer normalizes these entities [11]. Finally, we perform the query expansion by adding synonyms found in different thesaurus, in our case MeSH and PubChem. An additional run including this functionality is also evaluated on the TREC-Chem results to determine the impact of such expansion.

1.2.3. TWINC platform

The TWINC's Graphical User Interface, including the features described above, is developed with Flex technology. It is assessed during the ChemAthlon task of the PatOlympics in 2010 and 2011. The evaluation consists of three sessions of 20 minutes each. During these sessions, three IP experts perform a search on one of the topics in collaboration with a member of our team. Results returned by systems are analyzed manually and relevant documents are submitted. For each of these topics, up to 200 documents can be submitted. The tool performance is evaluated with two criteria: the recall, i.e. the number of hits among these 200 documents and the user-friendliness, i.e. the global appreciation of the tool by the experts.

2. Results

2.1.1. Ad hoc search and related article search

The related article search obtained top-performing results, as we were ranked first out of eight participants in TREC-Chem 2009. The best tuning obtained a mean-average precision (MAP) of 16.9%.

For the TS task, TWINC was ranked as fifth out of six participants. However, the query sample was regarded too small to derive statistical significance difference in the results.

2.1.2. Chemical query expansion

The chemical expansion feature, evaluated during the PA task of TREC-Chem 2009, has shown an improvement of 2%, leading to an improvement of the MAP from 17.9% to 18.2%.

2.1.3. TWINC platform

The TWINC platform is freely available for non-commercial use at http://casimir.hesge.ch/TWINC/index.html (Figure 3). It obtained two consecutive years the jury's choice of the PatOlympics for its user-friendliness (Table 1). Concerning the performances, we were ranked first in 2010.

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 the ChemAthlon 2010 and 2011

3. Discussion

We obtained in TREC-Chem 2009 competitive results for the related article search, but less encouraging results for the ad hoc search. Our re-ranking approaches based on citations and claims seem to be relevant strategies. Regarding the chemical query expansion, the improvement obtained by our approach is quite modest (+2%). To avoid an overload of synonyms bringing an important noise, we decided to limit the query expansion to a subset of the supposed most common synonyms, with potentially a loss of relevant synonyms. Thus, we can assume that this feature will be more powerful in

an interactive usage, through the validation of relevant synonyms by the user. Therefore we can expect that ad hoc search performs better in the TWINC platform due to the user interaction.

Despite the promising results obtained by TWINC, we are aware that it should be balanced due to the fact that the qualitative evaluation has been performed on a very limited sample of queries and compared to a restricted number of tools. Nevertheless, such competitions are currently the best available evaluation platforms for academic research in collaboration with IP experts.

Moreover, TWINC is a useful tool, targeting a variety of communities of users, because of its ability to cope with different manners to name medicinal substances. While query normalization is at this stage limited to chemical-related entities, retrieved patents are normalized with various types of entities, such as pathology, and relevant MeSH descriptors are indicated for each document. Finally, we also include features such as keywords highlighting to facilitate results analysis. Our tool can also be customized to be applied in corporate search environments to process domain and company-specific vocabularies, including non-English literature and patents reports (e.g. Chinese, Japanese).

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 information in this document reflects solely the views of the authors and no guarantee or warranty is given that it is fit for any particular purpose. The European Commission, Directorate General Information Society and Media, Brussels, is not liable for any use that may be made of the information contained therein.

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] Azzopardi L, Joho H, Vanderbauwhede W. A Survey on Patent Users Search Behavior, Search functionality and System requirements. IRF Report. 2010.

[4] 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.

[5] Lupu M. PatOlympics : an infrastructure for interactive evaluation of patent retrieval tools. In Proceedings of the DESIRE '11 Conference. 2011.

[6] 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.

[7] Gobeill J, Teodoro D, Pasche E, Ruch P. Report on the TREC 2009 Experiments: Chemical IR Track. In proceedings of the Text Retrieval Conference. 2009.

[8] 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.

[9] Mase H, Matsubayashi T, Ogawa Y, Iwayama M, Oshio T. Two-stage patent retrieval method considering claim structure. Working notes of the fourth NTCIR workshop meeting. 2004:256-261.

[10] Corbett P, Murray-Rust P. High-Throughput Identification of Chemistry in Life Science Texts. Computational Life Sciences II. 2006:107-118.

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