Published in EMBnet.journal , 2012, vol.18b, p. 44-46 which should be cited to refer to this work

Development of a text search engine for medicinal chemistry patents

Emilie Pasche[1]✉, Julien Gobeill[2], Fatma Oezdemir-Zaech[3], Therese Vachon[3], Christian Lovis[1], Patrick Ruch[2]

[1]Division of Medical Information Sciences (SIMED), University Hospitals of Geneva and University of Geneva, Geneva, Switzerland
[2]Bibliomics and Text-Mining Group (BiTeM), Information Science Department, University of Applied Sciences, Geneva, Switzerland
[3]Novartis Institute for BioMedical Research – Text Mining Services (NIBR-IT/TMS), Novartis Pharma AG, Basel, Switzerland

Motivation and Objectives

Over the last decades, the size of patent collections has strongly increased. Thus, in 2009, it was estimated that there are globally about 50 millions patents (Bonino et al., 2010) with about 15-20 millions related to medicinal chemistry, which represent a corpus of knowledge comparable to the content of MEDLINE. These collections represent an important and high-quality source of knowledge. However, while the past years have seen the development of a wealth of search engines and text mining instruments to navigate the bibliome – a term coined by (Grivell, 2002) to refer to the post-omics biomedical literature – with applications such as EBIMed, EAGLi, GoPubMed and Twease to cite only a few of them, text-mining applications dedicated to patents of the biomedical field remain rare.

Recently, the development of such specialized patent retrieval engines has benefited from the effort of dynamic communities of researchers, encouraged by the emergence of several evaluation campaigns, such as the Text REtrieval Conferences (TREC), one of the most popular competitions to evaluate and compare search engines. Prestigious universities, but also corporate research centres have regularly participated to these competitions. Lately, a task of information retrieval dedicated to patent search for chemistry, called TREC-Chem, has been set up. The objective was to model a prior art search task on a sizeable patent collection (two millions patents).

Based on the experience we have acquired during TREC competitions, we have developed and tuned an original search engine dedicated to patent search in the pharmaceutical domain. This paper describes the indexing and tuning of the engine to perform different types of search in a corporate patent collection.

Methods

A set of 1'004'868 patents has been randomly selected out of a collection of more than 13 millions of patents stored in an Oracle Database provided and maintained by IBM Almaden for Novartis. The content of the patents has been extracted using SQL queries and stored in files using an ad hoc XML format.

Evaluation of our methods is based on three sets of queries and relevance judgments. The first benchmark (B1) is used to evaluate the related patent search using the same methodology as proposed by TREC-Chem 2009 for the Prior Art Search task (Lupu et al., 2009). It is constituted of 96 topics or queries. Each topic corresponds to the title, abstract and claims of a given patent. For these experiments, the relevance judgments are generated out of the set of patents cited as prior-art by the given query. Only patents that are cited in the set of 1'004'868 patents are selected as many citations may concern patents not covered by the sample. The second benchmark (B2) is used to evaluate the engine in an *ad hoc* search task. In *ad hoc* search queries are usually limited to a few keywords. It is constituted of 24 topics, corresponding to the TREC-Chem 2010 and 2011 Technical Survey topics. Relevance judgments are provided by TREC and have been pre-processed to filter out patents not available in the 1'004'868 patent collection we are using. The last benchmark (B3) is used to evaluate a variant of the *ad hoc* search, where a single patent is targeted, using a *known-item search* methodology (Allen, 1989). It is constituted of 514 topics. Each topic contains ten words randomly selected from the title, abstract or claims of a given patent. In this set of experiments, the relevance judgments for each topic correspond to the patent from which the words are extracted. In that setting, a unique patent is considered as relevant for each query. The tuning of the system

Table1: P0 of the runs with the different strategies tested for each of the three benchmarks.

	B1	B2	B3
Baseline	2,20%	15,87%	23,63%
Remove description	2,87%	19,51%	33,59%
Remove description from metadata	3,63%	30,30%	35,02%
Use another weighting schema (BM25)	5,36%	20,05%	40,86%
Re-ranking based on citation network	6,76%	21,24%	40,87%
Injection of IPC codes	5,88%	23,28%	46,02%

is based on the maximization of the top precision, also called P0 or mean reciprocal rank. This measure evaluates the precision of the first returned result by the search engine. In our preliminary experiments, we focus on this metric since it provides a sound estimate of the retrieval effectiveness of the system for the three benchmarks. Indeed, other measures such as mean average precision cannot be applied to *known-item search* tasks.

We perform the indexing of the patent collection, using the Terrier search platform. Indexing is performed using baseline settings, with Porter stemming. First, we attempt to evaluate the impact of the description field – a time-expensive field to normalize and index – on the search effectiveness of the engine with the three use cases. Indeed, for sake of efficiency (in particular indexing time), we attempt to select only the most content-bearing sections of the patent. Second, we perform an ontology-driven normalization of the patent content. Three terminologies are used: Medical Subject Headings (MeSH), Gene Ontology (GO) and Caloha (Duek et al., 2012). Main terms and identifiers of mapped terms are stored as metadata. We evaluate the impact of the metadata field, to determine whether our onto-terminological normalization strategies bring useful additional information. Third, we evaluate the impact of the search models. Two search models are tested: the Okapi BM25 (Robertson et al., 2000) and PL2, a model based on Poisson estimation from randomness (Amati et al., 2002). Fourth, we evaluate the use of co-citation networks to improve our strategy. This approach consists in favouring the patents that are the most cited ones in the collection. We rank all patents by the number of time each patent is cited by the others; thus building a large co-citation matrix. Then, we combine through linear combination this ranking with the results of the query as originally returned by the retrieval engine. Fifth, we attempt to evaluate the impact of the use of IPC classes. Some authors (Sternitzke, 2009) reported that using IPC codes with four values allowed retrieving the totality of the state of the art. Our method consists simply to add IPC codes to the topics and execute a new run.

Results and Discussion

The main, as well as most surprising result is that the *description* field did not improve our results for any of the three benchmarks (Table 1), but rather decreased significantly the precision at high ranks (P0). We thus decided to remove *description* fields from the engine's indexes, which resulted in faster indexing and reduced the size of indexes.

Second, we observed that the use of metadata, which was generated based only on the content of the title, abstract and claims, improved the precision of our results compared to the metadata including the description (Table 1). Thus, we can assume that descriptions should simply be discarded not only for indexing as mentioned in previous section, but also from the onto-terminology-driven normalization, which should result in a significant gain of time for the normalization process. It is to be noted that the next experiments are not based on this observation and use the onto-terminology-driven normalization of the full patent content (including description). As a further experiment, it would be interesting to evaluate the impact of the normalization by entity types. Indeed, (Ruch et al., 2005) reported that normalization and ex-

pansion of genes and gene products degraded the precision of search in MEDLINE during the TREC Genomic competition, while normalizing chemical, pathological, organism-related and anatomical concepts was moderately effective.

Third, concerning the weighting model, we observed that BM25 performed better than the deviation from randomness weighting schema we tested (Table 1). Our experiments focused on the feature selection and combination steps; therefore we assume the results reported here are rather weighting schema-independent; in particular because BM25 can be regarded as a strong baseline in the domain.

Fourth, we observed that the re-ranking based on citation networks improved results for the three benchmarks, but mainly for the related patent search (Table 1). We thus can assume that it is an appropriate functionality for prior art tasks (+26%, $p<0.01$). The improvement for the *ad hoc* search task with the TREC benchmark was also significant (+5.9%).

Finally, we observed that IPC codes improved *ad hoc* search, but not related patent search (Table 1). Thus, we can assume that using an interactive IPC classifier (Teodoro et al., 2010) for *ad hoc* search could have a beneficial effect on the effectiveness of the search engine. In contrast, the length of the input for the prior art search makes obviously the use of IPC descriptors of less value

We have thus presented the development of a search engine dedicated to patent search, based on the state of the research methods applied to patents. We have showed that a proper tuning of the system clearly increases the effectiveness of the system. We can also conclude that different search tasks, such as related patent search and *ad hoc* search, do demand to set up specific information retrieval models to be optimal.

Acknowledgements

Funding: Novartis Pharma AG.

References

1. Allen B (1989) Recall cues in known-item retrieval. J Am Soc Inf Sci 40(4), 246-252. doi:10.1002/(SICI)1097-4571(198907)40:4<246::AID-ASI4>3.0.CO;2-Z
2. Amati G and Van Rijsbergen CJ (2002) Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Trans Inf Syst 20(4), 357-389. doi: 10.1145/582415.582416
3. Bonino D, Ciaramella A and Corno F (2010) Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics. World Patent Information 32(1), 30-38. doi:10.1016/j.wpi.2009.05.008
4. Duek PD, Gleizes A, Zwahlen C, Mottaz A, Bairoch A et al. (2011) CALOHA: A new human anatomical ontology as a support for complex queries and tissue expression display in neXtProt. Bio-Ontologies 2011. http://bio-ontologies.knowledgeblog.org/196 (accessed 06 October 2012).
5. Grivell L (2002) Mining the bibliome: searching for a needle in a haystack? New computing tools are needed to effectively scan the growing amount of scientific literature for useful information. EMBO Rep 3(3), 200-203. doi: 10.1093/embo-reports/kvf059
6. Lupu M, Piroi F, Huang XJ, Zhu J and Tait J (2009) Overview of the TREC 2009 Chemical IR Track. In proceedings of TREC 2009. http://trec.nist.gov/pubs/trec18/papers/CHEM09.OVERVIEW.pdf (accessed 06 October 2012).
7. Robertson SE, Walker S and Beaulieu M (2000) Experimentation as a way of life: Okapi at TREC. Information Processing & Management 36(1), 95-108. doi: 10.1016/S0306-4573(99)00046-1
8. Ruch P, Müller H, Abdou S, Cohen G and Savoy J (2005) Report on the TREC 2005 Experiment: Genomics Track. In proceedings of TREC 2005. http://trec.nist.gov/pubs/trec14/papers/uhospital-geneva.geo.pdf (accessed 06 October 2012)
9. Sternitzke C (2009) Reducing uncertainty in the patent application procedure – Insights from invalidating prior art in European patent applications 31(1), 48-53. doi: 10.1016/j.wpi.2008.04.007
10. Teodoro D, Pasche E, Vishnyakova D, Gobeill J, Ruch P et al. (2008) Automatic IPC Encoding and Novelty Tracking for Effective Patent Mining. In proceedings of NTCIR-8 Workshop Meeting. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings8/NTCIR/03-NTCIR8-PATMN-TeodoroD.pdf (accessed 06 October 2012)