Exploring a wide Range of simple Pre and Post Processing Strategies for Patent Searching in CLEF IP 2009

Julien Gobeill[1], Douglas Theodoro[2], Patrick Ruch[1]
BiTeM group

[1] University of Applied Sciences, Geneva, Switzerland
[2] University and Hospitals of Geneva, Switzerland
julien.gobeill@hesge.ch

Abstract

Patent processing is the rising research field in the Western Information Retrieval community. The objective of the 2009 CLEF-IP Track was to find documents that constitute prior art for a given patent; in other words, participants had to re-build the patent citations field of a given patent. We explored a wide range of simple pre-processing and post-processing strategies, using Mean Average Precision (MAP) for evaluation purposes. For determining the best document representation, we evaluated the impact of each field, among title, abstract, description, claims and IPC codes. Despite our efforts to design a specific stopwords list, the description field had a negative impact on the retrieval (- 14%) and had to be discarded, while the claims field seemed to be the most informative one (+ 86%). Then, we tuned a classical Information Retrieval engine in order to perform the retrieval step; the chosen weighting scheme finally was BM25. Finally, we explored two different post-processing strategies. Filtering retrieved patents that didn't share at least one IPC code with the query led to a significant improvement (+10 % when using complete IPC codes); as for the document representation, using the complete IPC codes led to greater improvements than using 4-digits IPC codes. The second post-processing strategy was to exploit the citations of retrieved patents in order to boost scores of cited patents. A light use of direct citations led to a small improvement (+ 3%), but despite our efforts we were not able to take benefit from the citation network for this task. Combining all selected strategies, we computed optimal runs that reached a MAP of 0.122 for the training set, and a MAP of 0.129 for the official 2009 CLEF-IP XL set, that makes our team having submitted the best run after – a far away from – the Humboldt University run. The 2009 CLEF-IP Track provided us a first approach of patent searching techniques; however, we need know to investigate more advanced techniques, by drawing our inspiration in particular from works that were conducted in the previous NTCIR campaigns.

Categories and Subject Descriptors
H.3.3 Information Search and Retrieval; I.2.7 - Natural Language Processing

Free keywords
Cross-Language Retrieval, Patent Searching, Patent citations, IPC codes, Patent Representation

1 Introduction
According to the European Patent Office (EPO), 80% of the world technical knowledge can be found in patent documents [1]. Moreover, patents are the only tool for companies to protect and take benefit from their innovations, or to check if they are free to operate in a given field or technology. As patent applicants have to provide a prior art search describing the field and the scope of their invention, and as a single missed document can invalidate their patent, patent searching is a critical field for the technical, scientific and economic worlds.

A Patent Track is proposed in NTCIR [2] since its third edition in 2002. As the NTCIR workshops took place in Japan and dealt with Asian languages, they did not retain all the attention of the Western Information Retrieval community. At the instigation of the Information Retrieval Facility, Patent Tracks appeared in 2009 in Europe (the CLEF IP competition [3]) and in North America (the TREC Chemistry competition [4]). These tracks aim at bridging the gap between the Information Retrieval community and the world of professional patent search.

The 2009 CLEF-IP Track was defined by the official guidelines as being a prior art search task: the goal was to find patents that constitute prior art for a given patent, in a collection of patent documents from EPO sources [5]. As there were more than 1M patent documents, and as these patent documents were huge files (often several megabytes), the task was firstly to be considered as a very large scale Information Retrieval task. The preprocessing strategies hence are essential in order to work with a manageable but efficient collection. On the

other hand, the different structured fields in patents make possible several post-processing strategies in different domains, such as text categorization with IPC codes, or cocitations networks with references.

Thanks to a well designed training set, with 500 patents used as queries, we were able to explore and evaluate a wide range of the strategies we mentioned above. In the following sections, we present and discuss the different strategies in the same order than we explored them during our work on the 2009 CLEF-IP Track.

2 Data and Strategies

The CLEF-IP 2009 collection contained around 1'950'000 patent documents from the EPO. As several patent documents could belong to a same patent, there were actually around 1 million patents. Each patent document was a XML file containing structured data; different fields were delimited by specific tags. Fields that retained our attention were :

- *Title*
- *Description* : the complete description of the invention, that is the longest field.
- *Abstract* : a summary of the description field.
- *Claims* : the scope of protection provided by the patent.
- *IPC codes* : codes belonging to the International Patent Classification and describing technological areas
- *Citations* : patents cited in the prior art.

Inventor and *Applicant* fields were not retained, as we assumed they were not informative. We now think that we should have included these fields in the experiments. Moreover, we used IPC codes in two different formats: 4-digits codes (e.g. D21H) and complete codes (e.g. D21H 27/00). Citations were not used for building the patent representation, but were investigated for post processing purposes.

The task was to find patents that constitute the prior art for a given patent; in other words, participants had, from a given patent for which organizers had discarded the *Citations*, to re-build the *Citations* field. A training set of 500 patents was provided. In the *Citations* field, another patent can be cited because it can potentially invalidate the invention, or more generally because it is useful for the understanding of invention. Thus, two ways were possible in order to define what citations have to be re-build: a stringent qrel or a liberal qrel. All results reported in this Working Note were evaluated with the liberal qrel. More information is available in the official guidelines [5].

During our experiments, we could explore and evaluate a wide range of strategies. Indeed, as queries can be generated only by discarding the *Citations* field, organizers were able to generate a large training set. We chose to firstly develop a complete pipeline with default settings, in order to be able to evaluate a baseline run; thus, we were able to evaluate any strategy we explored by comparing it to the baseline run. Runs were evaluated with Mean Average Precision (MAP). The Information Retrieval step was performed with Terrier [6]. Thus, our approach can be seen as a gradient descent approach.

The first run we computed, with all mentioned patent fields representing the document and the queries, with standard Terrier settings and without any post-processing strategy, reached a MAP of 0.074.

3 Patent Representation

The first step was to decide how to merge several patent documents belonging to the same patent into a unique file. The official guidelines proposed several strategies, but we decided to keep all information contained in the different files and to concatenate it in a unique patent file.

3.1 Document Representation

The second step was to determine which fields to keep in the indexed patent files. Our priority was to keep the *Description*, as we hypothesized that it was the more informative field. However, the *Description* fields in patents are often huge, so we had to take care not to generate an unmanageable collection. Hence, our strategy was to light the *Description* field, by discarding a massive list of the most frequent words in the collection. Experiments showed that the best performances were obtained by using a list of 500 stopwords. Thus, using this 500 stopwords list was the optimal setting but still let a huge mass of data. Worst, we observed that discarding the *Description* field for document representation led to a MAP of 0.097, which was a + 30% improvement.

Despite all our efforts, the *Description* field as we used it contained more noise than information, and we had to discard it for the patent representation.

Table 1 shows some supplementary results on how much each field contributed to the final performance. From the new baseline run obtained by discarding the Description field (MAP 0.097), we discarded each field separately and observed the improvement induced by including this field.

Discarded field	MAP	Improvement
Baseline	0.097	
Title	0.096	+ 1 %
Abstract	0.091	+ 7 %
Claims	0.052	+ 86 %
IPC 4-digits codes	0.0791	+ 1%
IPC complete codes	0.0842	+ 3%

Table 1: Mean Average precision (MAP) for different Document Representation strategies.

Results show that the *Claims* are the most informative field, as using them led to a + 86 % improvement. This result contradicts the remarks of the patent expert provided by the official guidelines [5], that suggested that "*claims don't really matter in a prior art searches […] whereas it would be significant for validity or infringement searches*", unless the task finally must be seen as a validity search task. Another result is that the *Title* seems to be poorly informative. This result is coherent with what Tseng and Wu wrote in their study describing search tactics patent engineers apply [7]: "*It is noted that most patent engineers express that title is not a reliable source in screening the search results […] [as] the person writing up the patent description often chooses a rather crude or even unrelated title*". Finally, we chose to keep all fields excepted from *Description* in order to build the document representation.

3.2 Query representation

Experiments showed that, for query representation, keeping the *Description* field led to slightly better performances than discarding it (+ 3%). Hence, we chose to keep all fields in order to build the query representation.

4 Retrieval Model

Once we fixed the Patent Representation, we tuned the Information Retrieval system in order to find the best settings. As mentioned above, we used the Terrier 2.2.1 platform in order to make the retrieval.

Firstly, we evaluated several available weighting models in Terrier with their default settings, to make the conclusion that we didn't need to change the default BM25. Results are presented in Table 2. Please refer to the Terrier documentation in order to obtain more information about mentioned weighting schemes [8].

Weighting model	MAP
BM25	0.097
DFR_BM25	0.095
TF IDF	0.095
BB2	0.084
IFB2	0.088
In_expB2	0.089
In_expC2	0.089
InL2	0.093
PL2	0.093

Table 2: Mean Average precision (MAP) for different available weighting models in Terrier.

We then tuned the BM25 weighting model by setting the b parameter; we finally reached a MAP of 0.105 with b=1.15. Finally, we observed that using query expansion with the available Bo1 model (Bose-Einstein inspired), set with default parameters, led to a final MAP of 0.106.

5 Post Processing strategies

Once we fixed the best retrieval model, we focused on how additional information contained in patent document could be used for re-ranking and improving the computed run. We chose to explore two different strategies: whether to filter out-of-domain patents regarding to IPC codes, or to boost related patents regarding to the citations of the retrieved patents.

5.1 IPC filtering

In an expert patent searching context, Stemitzke [9] assumed in his abstract that "*patent searches in the same 4-digits IPC class as the original invention reveal the majority of all relevant prior art in patent*". Another study assumed that it is between 65% and 72% – whether citations were added by the applicant of the examiner – of European patent citations that are in the same technology class [10]. Moreover, dealing with what IPC granularity – whether 4-digits or complete codes – using in patent searches, the EPO best practices guidelines indicate that "*for national searches […] the core level is usually sufficient*" [11].

Hence, we decided to explore IPC filtering strategies that consisted in filtering (i.e. simply discarding in the ranked list) retrieved patents that did not share any IPC code with the query. We evaluated this strategy for both 4-digits and complete codes. Moreover, another strategy could consist in, for each query, only indexing documents that share at least one IPC code with the query. Thus we evaluated both strategies, respectively named *IPC filtering* and *IPC indexing* strategies, with both IPC granularities, 4-digits and complete. Results are presented in Table 3. *IPC filtering* strategy was applied in the previous baseline run that reached a MAP of 0.106.

MAP	***IPC filtering* strategy**	***IPC indexing* strategy**
Baseline	0.106	0.106
4-digits IPC codes	0.111 (+5%)	0.112 (+6%)
complete **IPC codes**	0.118 (+11%)	0.115 (+8%)

Table 3: Mean Average precision (MAP) for different filtering strategies using IPC codes.

Results show that both strategies led to improvements, but none was significantly better than the other. However, the *indexing* strategy needs to re-index a specific part of the collection for each query, which is a time-consuming process. Thus we preferred to apply the *filtering* strategy. Moreover, using the complete IPC codes let to a bigger improvement than using 4-digits codes (+11% comparing to +5%). Working on the patent representation, we also observed that complete codes seemed to be more informative (see Table 1). These results, and the designed strategy for automatic prior art searches, seem to run counter to the state of the art for expert prior art searches.

5.2 Citations boosting

Finally, we explored post-processing strategies dealing with patent citations. Few studies addressed the cocitation issue in the patent domain. Li and al. [12] used citations information in order to design a citation graph kernel; evaluating their work with a retrieval task, they obtained better results exploiting citation network rather than only direct citations.

We computed the citation network for the collection, and we explored a range of post-processing strategies, from citation graphs to weighting schemes based on the number of citations. Making a slightly use of the direct citations, we reached the MAP of our run from 0.118 to 0.122 (+3%). Another interesting result was the improvement of Recall at 1000 from 0.53 to 0.63. Unfortunately, we never were able to exploit the citation network, i.e. more than direct citations.

6 Official results

Hence, the final set of strategies we applied performed a MAP of 0.122 on the training set. As we explored all strategies we wanted to, we chose to just submit one official run for the CLEF-IP 2009 official test set. Evaluated on the XL set (10'000 queries), our official run reached a MAP of 0.129. These results make us one of the team leading the chase, far away from the leading team, from the Humboldt University, who submitted an outperforming run evaluated at 0.28 for MAP.

7 Multilingual tasks

CLEF aims at proposing cross-lingual challenges, thus a multilingual task was proposed in the CLEF-IP 2009 Track. The objective was to compare results for test sets in different languages: French, German and English [5]. To address this problem, we chose to keep the same pipeline than for the main task, and to simply translate the fields written in French or German into English, via Google translator [13]. Evaluated on the M test set (500 queries), we achieved a MAP of 0.111 for English, 0.095 for German and 0.1 for French. Strategies relying on IPC codes are language-independent; it would be interesting to evaluate their impact in these performances for multilingual runs.

8 Conclusion and future work

Finally, we explored a wide range of simple strategies, aiming at choosing the best document representation, at choosing the best information retrieval platform, and at applying some efficient post-processing tactics. The results were satisfying, as our run was one of the leading ones. Unfortunately, strategies that improved the performances were quite simple, and we need know to design more advanced winning strategies in order to still be competitive in the CLEF-IP 2010 evaluation. We probably need to improve our semantic representation of the patents, and to deal with the problem and the solution aspects of the invention. In particular, we have to pay attention to the works produced on this domain by Asian teams for the previous NTCIR competitions.

Limitations in the CLEF-IP 2009 evaluation were that retrieved documents were considered as relevant only if they were cited by the patent given as query. Yet, it does not imply that these retrieved documents were not relevant with regard the prior art of the invention. Indeed, if several documents are equally relevant regarding to a given part of the prior art, the examiner needs to cite only one of them, choosing less or more arbitrarily. Other variables such as geographical distance, technological distance or strategic behavior of the applicant have an influence on the citations and can induce additional biases in cited patents [10]. Thus, some retrieved documents can be judged non relevant in this evaluation, because another document was chosen in the citations; but these documents could be judged relevant and useful by a professional searcher in a semi automatic process. Nevertheless, the CLEF-IP 2009 evaluation let us to start working on patent searching and to compare our strategies in a very pleasant framework.

9 References

1. Augstein J., “Down with the Patent Lobby or how the European Patent Office has mutated to controlling engine of the European Economy”, Diploma Thesis, University of Linz, 2008
2. http://research.nii.ac.jp/ntcir
3. http://www.clef-campaign.org
4. http://www.ir-facility.org/the_irf/trec_chem.htm
5. Piroi F., Roda G. and Zenz V., “CLEF-IP 2009, Track Guidelines”, 2009.
6. Ounis I., Lioma C., Macdonald C. and Plachouras V.. “Research Directions in Terrier: a Search Engine for Advanced Retrieval on the Web”, Novatica/UPGRADE Special Issue on Next Generation Web Search, vol 8, pp 49-56, 2007.
7. Tseng Y.-H. and Wu Y.J., “A Study of Search Tactics for Patentability Search – a Case Study on Patent Engineers”, Proceedings of the 1st ACM Workshop on Patent Information Retrieval, 2008.
8. http://ir.dcs.gla.ac.uk/terrier/doc/configure_retrieval.html
9. Sternitzke C, “Reducing uncertainty in the patent application procedure – insights from malicious prior art in European patent applications”, World patent Information, vol.31, pp 48-53, 2009.
10. Criscuolo P and Verspagen B, “Does it matter where patent citations come from? Inventor versus examiner citations in European patents”, Research Policy, vol.37, pp 1892-1908, 2008.
11. http://www.epo.org/patents/patent-information/ipc-reform/faq/levels.html
12. Li X., Chen H, Zhang Z. and Li J., “Automatic patent classification using citation network information: an experimental study in nanotechnology”, Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, pp 419-427, 2007.
13. http://translate.google.com.
14. Waguespack D. and Birnir J., “Foreignness and the diffusion of ideas”, J. Eng. Technol. Manage. vol. 22, pp 31–35, 2005.