Simple vs. Sophisticated Approaches for Patent Prior-Art Search

Walid Magdy[1], Patrice Lopez[2], and Gareth J. F. Jones[1]

[1] CNGL, School of Computing, Dublin City University, Dublin 9, Ireland
[2] INRIA & Humboldt Universitat zu Berlin, Germany
[1]{wmagdy, gjones}@computing.dcu.ie, [2]patrice.lopez@inria.fr

Abstract. Patent prior-art search is concerned with finding all filed patents relevant to a given patent application. We report a comparison between two search approaches representing the state-of-the-art in patent prior-art search. The first approach uses simple and straightforward information retrieval (IR) techniques, while the second uses much more sophisticated techniques which try to model the steps taken by a patent examiner in patent search. Experiments show that the retrieval effectiveness using both techniques is statistically indistinguishable when patent applications contain some initial citations. However, the advanced search technique is statistically better when no initial citations are provided. Our findings suggest that less time and effort can be exerted by applying simple IR approaches when initial citations are provided.

1 Introduction

Prior-art search task in patent retrieval is concerned with finding all prior-art patents that are relevant to a patent application. Relevant prior-art patents have common technical aspects with a patent application, and include patents that can invalidate the novelty of the invention and patents that describe the state-of-the-art in the field of the invention on which the patent application is building [4, 5]. Identified relevant patents are cited in a search report which is part of the publication of the patent application. A typical patent application when filed to a patent office will include some initial patent citations describing the state-of-the-art. These citations are considered useful for patent examiners to understand the key aspects of an application and to start a search for relevant existing patents. However, large proportions of these initial citations are ultimately not found to be relevant, and are not included by patent examiners in the search report. Moreover, patent examiners usually identify a large amount of additional relevant patents.

Patent prior-art search task was addressed in the CLEF-IP task in both 2009 [5] and 2010 [4]. In 2010, relevant documents identified by the European Patent Office (EPO) in the search reports acted as the relevance set, and the initial patent application filed to the EPO acted as the topic [4]. The objective was to identify the relevant documents for each patent topic automatically. Submitted runs from two participants achieved considerably higher retrieval effectiveness than the other submitted runs [1, 3]. These two participants used IE techniques to extract the patent citations provided in the patent application. Later they utilized these citations in two different ways to improve the retrieval effectiveness. The participant group who

achieved the best run used an advanced search approach for the retrieval process including key-term extraction, multiple retrieval models, multiple indexes, and post-ranking techniques [1]. The other participant group used a simpler IR technique [3].

In this paper, the best two runs in the CLEF-IP 2010 are revisited and compared after using the same extracted citations for both runs. The comparison was applied on the English topics and divided into two sets. The first set contains topics for which patent citations can be extracted from its text, and the second set contains topics that do not include any patent citations in their description. The comparison results show that when patent citations can be extracted from the text of a patent topic, the simple search approach achieves results comparable to the more sophisticated method. However, when no citation could be extracted from the patent topic, the advanced search approach achieves significantly better results. This finding suggests that using simpler approaches could be used effectively for patent search when applicants provide initial citations, which is the situation for more than half of the filed patents. Otherwise, following this hypothesis more complex approaches could be used to improve the retrieval effectiveness when no initial citations are provided.

2 Retrieval Methodologies in CLEF-IP 2010

Different retrieval methodologies were used in the 25 runs submitted for the CLEF-IP 2010 prior-art patent search task. Excluding the best two runs, the other 23 runs achieved retrieval effectiveness ranging from 0.007 to 0.14 MAP, and 0.01 to 0.21 PRES@100. PRES (Patent Retrieval Evaluation Score) is a new evaluation metric used in CLEF-IP 2010 [4] which emphasises the quality of the system in retrieving a larger portion of the relevant documents at relatively high ranks according to a user given cut-off (N_{max}) [2]. The best two runs used a citation extraction methodology to achieve significantly higher scores. The second ranked run achieved 0.2 MAP and 0.32 PRES@100 while the first ranked run achieved 0.26 MAP and 0.39 PRES@100, which are considerably higher score than those for the other runs.

A comparison of the best two runs was carried out using 1348 English topics provided by the CLEF-IP 2010 to search a collection of 1.35M patents from the European Patent Office (EPO) [4]. The comparison is based on the same set of extracted citations, which are those extracted by the first participant [1]. This set is used because it included more citations, since it used additional external resources to improve the results by using patent family look-up [1]. The extracted patent citations included 7706 citations extracted from 728 topics. For the experiments, the 1328 topics were divided into two sets: 728 topics that have citations extracted from their text and 620 topics that have no citations.

2.1 Simple Search Approach

The approach presented in [3] uses a straightforward IR technique to retrieve a ranked document list, then appends it to the extracted citations list to create the final results list. In this approach patent documents are treated as plain text neglecting their structure. The query is constructed from terms in the description section of the patent topic after filtering out terms that appeared once, in addition to bigrams (two consecutive terms) that appeared in the title and abstract sections of the patent topic

more than one time. The Indri search toolkit was used for the retrieval process [6]. The retrieved results are then filtered based on the patent classification, where each patent document or topic has a classification according to the scope of the invention. This filtering process guarantees that the patent topic and the retrieved results share the same first three levels of classification [3]. The produced results list is then simply appended to the extracted citations after removing the duplicates.

2.2 Advanced Search Approach

The approach presented in [1] uses a more sophisticated retrieval method:

1. Creating a working set for each patent topic for pruning the search space. This working set is built recursively from patents which share common classification, inventor, or citations with the patent topic and extracted patent citations.
2. Applying multiple retrieval models (BM25 and Indri) using different indexes (English lemma, phrases, and concepts) for producing several sets of ranked results, and then merging them based on multiple SVM regression models and a linear combination of the normalized ranking scores.
3. Post-ranking the results based on an SVM model exploiting patent metadata.

This system used several complementary indexes, including phrase and conceptual indexes. The phrase index is based on the extraction of key terms from the patent topics using vast ranges of metrics and features. The conceptual index is based on a large scale database resulting from merging various terminological resources (METS, UMLS, the Gene Ontology, Wikipedia, etc.).

3 Results

Figure 1 shows the retrieval effectiveness of the simple and advanced IR techniques for the patent prior-art search task. Results are reported for the two topic sets when citations could and could not be extracted. The retrieval effectiveness when only extracted citations are used without any kind of IR is reported as a baseline. From Figure 1 it can be seen that the extracted citations achieve higher retrieval effectiveness than either IR approach when no citations could be extracted. Also it is clear that both systems achieved nearly double the performance level for topics that contain patent citations within their text. Comparing both systems on their performance, it can be seen that when citations exist, the simple IR approach is as effective as the more complex approach when compared using PRES and even better when compared by MAP. However, when no citations could be extracted as an initial step, the complex approach is significantly better. This observation leads to the hypothesis that when a patent application includes directly cited prior-art patents, simple search approaches are sufficient to achieve good retrieval results.

To further support this finding, the results list of the complex approach was appended to the extracted citations list after removing duplicates in the same way as the final step of the simple approach. This approach gave MAP of 0.3 and PRES@100 of 0.43 which is statistically indistinguishable from the results for the simple approach shown in Figure 1.

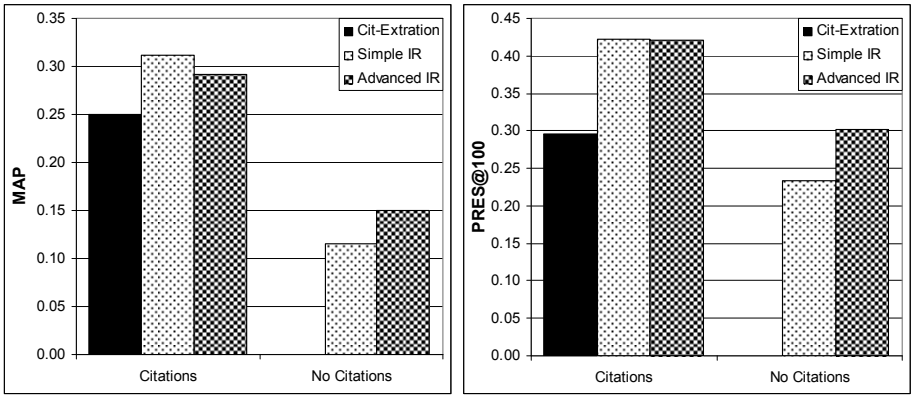

Fig.1. Retrieval results for simple and complex IR approaches with CLEF-IP prior-art search task, when citations could be and could not be extracted

4 Conclusion

In this paper, we have presented a comparison between two approaches for the patent prior-art search task. The first approach is characterized by its simplicity and low resources requirement, while the second one is more sophisticated, using an advanced level of content analysis. The results show that the simple search approach is as effective as the sophisticated one when initial citations are provided. This is the situation for 54% of the test collection used in our experiments, but also a legal requirement of the EPO (Rule 27(b) of the EPC, European Patent Convention). The observation that simple IR approaches can in many cases achieve similar results to current sophisticated ones, suggests that further investigation is required to better understand the retrieval process in patent prior-art search in order to develop more effective methods for this task.

5 References

1. Lopez P. and L. Romary. Experiments with citation mining and key-term extraction for Prior Art Search. In *Proceedings of CLEF 2010.* (2010)
2. Magdy W. and G. J. F. Jones. PRES: a score metric for evaluating recall-oriented information retrieval applications. In *SIGIR 2010.* (2010)
3. Magdy W. and G. J. F. Jones. Applying the KISS Principle for the CLEF-IP 2010 Prior Art Candidate Patent Search Task. In *Proceedings of CLEF 2010.* (2010)
4. Piroi F. CLEF-IP 2010: Retrieval Experiments in the Intellectual Property Domain. In *Proceedings of CLEF 2010.* (2010)
5. Roda G., J. Tait, F. Piroi, and V. Zenz. CLEF-IP 2009: retrieval experiments in the Intellectual Property domain. *In Proceedings of CLEF 2009.* (2009)
6. Strohman T., D. Metzler, H. Turtle, and W. B. Croft. Indri: A language model-based search engine for complex queries. In *Proceedings of ICIA.* (2004)

This research is partially supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (CNGL) project at Dublin City University.