CLEF-IP 2010; Building strategies, a year later

W. Alink, R. Cornacchia, and A.P. de Vries

Centrum Wiskunde & Informatica,
Science Park 123, 1098 XG Amsterdam, Netherlands
{alink,cornacchia}@spinque.com,arjen@cwi.nl
http://www.cwi.nl/, http://www.spinque.com/

Abstract. After participating in last year's CLEF IP (2009) evaluation benchmark, our scores were rather low. The CLEF IP 2010 PAC task enabled us to correct some experiments and obtain better results, basically using the same techniques (almost the same BM25-category strategy as used last year) and improved strategy builder software, and less computing hardware at our disposal. The results are now comparable with other participants. Similar to last year, no feature extraction techniques have been applied; and queries only used the structural information provided in the XML-format of the patent-documents. Furthermore we participated in the new CLS task, which, although scores were rather low, shows again the flexibility of our approach. The low scores can be explained by the straight-forward method applied searching the patent-document collection using keywords from the topic-patent, and using the IPCR-classifications extracted from the documents as results.

1 Introduction

The main objective of this research is to demonstrate the importance of flexibility in expressing strategies for patent-document retrieval. While last year's submission focussed on flexibility, this year also scalability, and retrieval quality have been taken into account. The results of our system are comparable with other participants, while it can also be operated interactively, which makes it a powerful tool. The paper gives an overview of the system (Section 2) and the techniques used to generate the runs (Section 3). Afterwards the results are evaluated (Section 4). Finally a conclusion is drawn (Section 5).

2 System overview

We created our submission for the CLEF-IP 2010 evaluation benchmark using Spinque's strategy builder interface. The setup is similar to last year's setup, only the system has matured. The hardware requirements have lowered from a supercomputer to a single server, and querying could be performed on a desktop machine, whereas query-speed has even increased. Last year the strategies were still optimized by hand; this year, there was no performance tweaking done between strategy definition and performing the benchmark runs. The reader is pointed to last year's paper [1] for more details about the setup.

2.1 Index creation

The index was created on a high-end server: 2.4Ghz 4 core processor, 36GB RAM, 5x 2TB SATA disks in RAID-5 configuration. All querying was done on a 3-year old desktop with average computing specs: 2.4Ghz processor, 4 cores, 8GB ram, 2x500GB SATA disks in RAID-0. Creating a generic index for the whole collection took about 3 days. Creating a full run over 2000 topics took about 12 hours. A SQL dump of the resulting database is roughly 300GB uncompressed in size (81GB bzipped).

3 CLEF-IP Experiments

This Section reports on the experiments conducted for the official submission. Fine-tuning of all parameters used for the PAC task was performed on the training set provided. The parameters for the CLS task haven't been trained.

Instead of merging patent-documents belonging to the same patent into a single document (suggested in the CLEF-IP instructions), we have indexed (similar to last year's submission) the original documents, and, aggregate scores from different patent-documents into patents as part of the search strategies.

Two runs have been submitted for CLEF-IP 2010.

3.1 Prior-Art Candidate Search Task

The strategy can be explained as follows: first make a selection of patents in the corpus that have at least one classification code in common with the topic patent, or have the same assignee. Search this selection using 26 keywords from topic-patent using the BM25 model. As a last step in the strategy, and due to the evaluation measures that are used for CLEF-IP 2010, all patent-documents within the same simple-family have been given the same score as the best scoring patent-document of that family. See figure 1.

Difference with last year is the searching for patents by same assignee, and the hard selection before keyword search based on classification and assignee, instead of a mixture between keywords, classifactions, and assignees.

3.2 Classification Task

Used same strategy builder as for the PAC run, no additional coding, or re-configuring. The strategy was to search documents using 26 keywords from the topic-patent, and then extract the classifications of these documents as results. See figure 2.

4 Evaluation and analysis

We learned from the problems found during last year's participation. Similar to last year's contribution, the strategy building contribution, shows flexibility without re-programming, re-indexing, or re-configuring a system. Last year

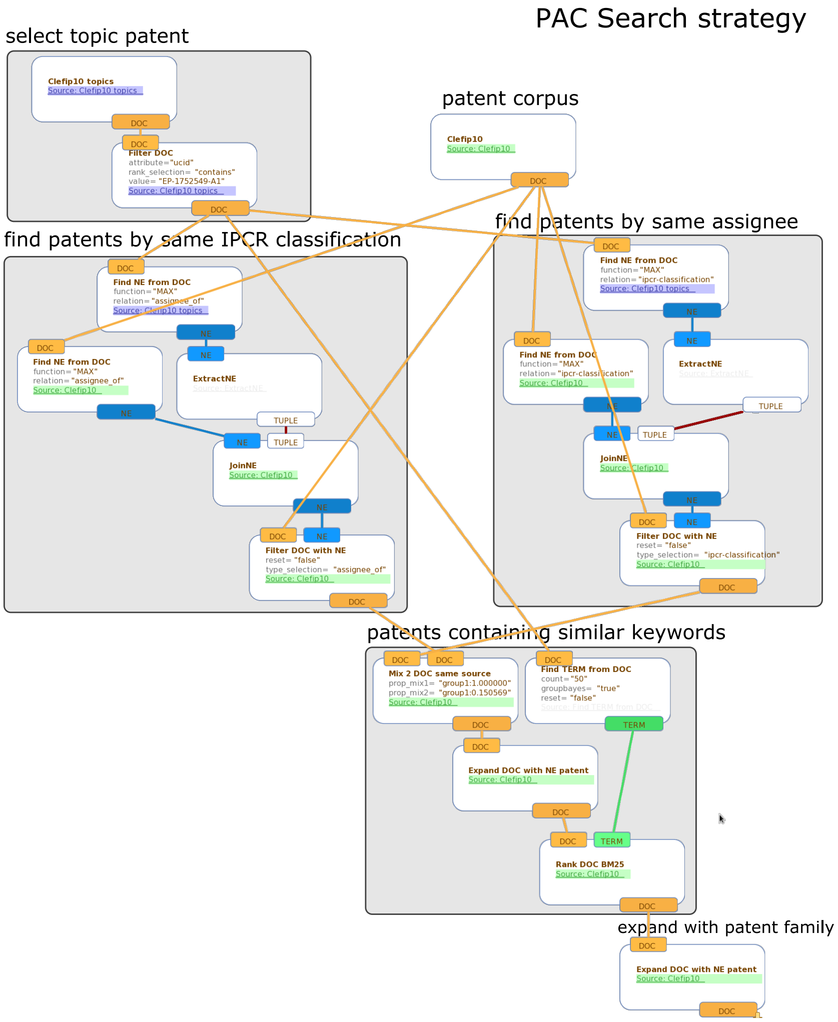

Fig. 1: Prior-Art Candidate search strategy

(CLEF-IP 2009) the system was still under heavy development. This caused the software to perform far from optimal. This year, most of the issues have been resolved.

4.1 Prior-Art Candidate Search Task

The results are more in line with the expected results than last year. It looks like the participants with similar strategies perform similar.

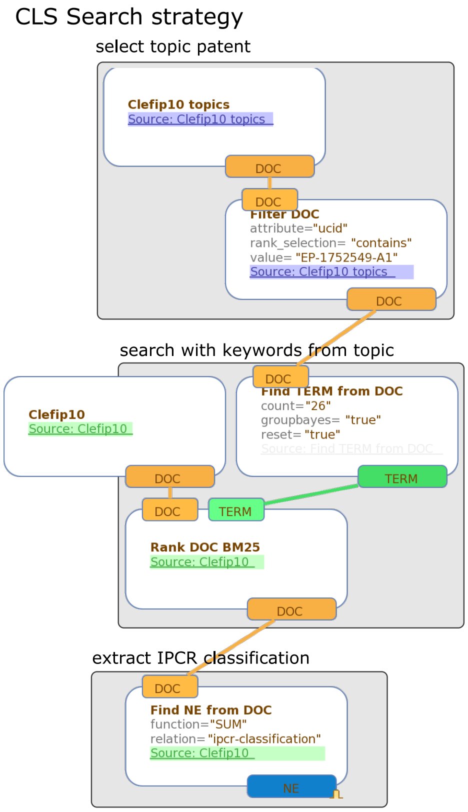

Fig. 2: IPCR Classification strategy

Our run did not use structural information outside the given structural information in the XML documents in which the patent-documents were provided. It was shown by Lopez and Romary [2] that an increase of MAP by .10 (absolute) can be achieved when using citations extracted from the topic-patent.

Notice that the MAP, recall, and PRES are all quite stable over the different languages of the topic-patent, while some other participants seem to have much higher fluctuations between them. A possible explanation for this phenomenon is the lack of language specific optimizations.

In contrast to last year, we did not merge the results of the classification search, with the results of the keyword search, but filtered the results of the classification search with the results of the keyword search. This had a major consequence; documents not having any classification in common with the topic document were not extracted. Recall is therefore likely to be lower than when the results would have been mixed; notice that the average number of results per topic is not the maximum (1000), but a little lower. Results in [3] confirm this effect of hard (facet) selections. However, it has been much easier to define a balance between the weight of the classification and the weight of the keywords, which took a lot of time in last year's submission. More work on automatically tuning the weights for multiple components of the search to a specific task has to be done, and would likely yield better results, both in terms of MAP and recall.

4.2 Classification Task

Our classification run got low scores for both precision and MAP. The recall (at 25 and at 50) was reasonable to high compared to other participants. This is most probably due to the fact that for each topic we tried to provide the full 1000 results. Other dedicated systems do clearly outperform the strategy we have build. It would be interesting to see what happens to the precision scores if a cut-off was applied to the results based on the computed probabilities (and using a fixed cut-off value for all topics). Other improvements could probably be made by using more aspects of the patent, perhaps using the patent citations used in the topic patent, and information on the inventor and assignee. Due to the limited time available, such runs have not been created.

5 Conclusion

Participation in the CLEF-IP 2010 evaluation track has been easier than last year. Compared to last year's results seem to have improved relative to other participants for the PAC task. The results for the CLS task show that there is still a lot of room for improvement. Also, more work is needed on the automatic tuning of strategy parameters.

Acknowledgements: We would like to thank the MonetDB kernel developers and Spinque for their support.

References

1. Wouter Alink, Roberto Cornacchia, and Arjen de Vries. Running CLEF-IP experiments using a graphical query builder. In *Lecture Notes in Computer Science (to appear)*, 2009.
2. Patrice Lopez and Laurent Romary. Multiple retrieval models and regression models for prior art search. In *CLEF Working Notes*, 2009.
3. Lanbo Zhang and Yi Zhang. Interactive Retrieval based on Faceted Feedback. In *Proceedings of the 33rd ACM SIGIR Conference*, 2010.