Towards Creating a Test Collection for European Patents

Erik Graf
Dept. of Computing Science
University of Glasgow
graf@dcs.gla.ac.uk

Leif Azzopardi
Dept. of Computing Science
University of Glasgow
leif@dcs.gla.ac.uk

ABSTRACT

In this poster, we discuss issues regarding the creation of a test collection for European patents for the task of Prior Art Search. Our approach is based on inferring relevance assessments from references extracted from patent documents. This should enable the creation of high quality, realistic judgements in a cost effective manner. Our future work will be directed towards refining and implementing the proposed methodology.

1. TEST COLLECTION CREATION

Test collections form the central element in obtaining objective and quantitative evaluation of the effectiveness of Information Retrieval (IR) systems for particular tasks such as document retrieval [5]. To serve that purpose a test collection consists of a representative set of documents (corpus), a representative set of task specific topics (formalized information needs), and corresponding relevance assessments [3].

The task for the patent test collection, we propose to construct, is the identification of prior art, because (1) it is one of the most commonly executed patent retrieval tasks, and (2) the task is ingrained in the very creation of a patent document [2] (i.e. an examiner explicitly marks relevant prior art for a specific patent application). To build a test collection for patents, we first assume that the corpus will be a subset of patent documents from the European Patent Office (EPO) issued between 1978 and 2008. For each examined patent specification a set of annotated references is available that refers to other patents or non-patent literature in, and outwith, the corpora. The subset of references to patents within the corpora, we posit, can be used in the creation of topics for prior art search.

The justification for reverse engineering relevance assessments from the references within a patent is based on the following:

- The patent references found on patent documents issued by the European Patent Office are set by its patent examiners. The subject and legal expertise of the examiner at the patent office allows for qualified assessment of relevance from his or her side with respect to the prior art search task.
- The specification of the European Patent Convention [4] (Rule 44, article 92(1), and article 54) sets the criteria for valid reference matter and can thus be interpreted as definition of relevance for an information need.
- The reference categories given in the 'Guidelines for Examination in the EPO B X 9.2' [1] provide a precise description of the nature of the stated form of relevance.

Thus, we believe it is feasible to create a methodology for prior art test collection creation which is not only credible and realistic, but also cost effective. The method could be repeatedly applied to different patent documents in order to create numerous topics in a range of different domains of the patent classification system (such as chemistry; metallurgy , electricity, etc). While a number of other issues remain in order to formulate the methodology, such as defining subtasks, queries, and more pragmatic issues concerning the extraction of the references, it is anticipated that these can be resolved satisfactorily to create a reliable and high quality test collection.

2. OUTLOOK

Although this work is still in progress our initial analysis looks promising in terms of cost effective creation of numerous topics. Further work will be directed towards, examining the issues above and proposing a general method for creating prior art topics and relevance judgements. Once formulated, we shall apply the methodology to form a pilot patent test collection. On which we shall compare standard retrieval models in order to obtain a baseline in terms of performance, and to identify any problems with the application of the methodology.

Acknowledgements The authors would like to thank Matrixware Information Services[1] and the Information Retrieval Facility[2] (IRF) for their support of this work.

3. REFERENCES

[1] http://www.epo.org/patents/law/legal-texts/guidelines.html.
[2] N. J. Akers. The referencing of prior art documents in european patents and applications. *World Patent Information*, 22(4):309–315, Dec. 2000.
[3] C. Cleverdon. The cranfield tests on index language devices. pages 47–59, 1997.
[4] D. Visser. *The annotated European Patent Convention. (10th revised ed.)*. H. Tel Publisher, Veldhoven, The Netherlands, 2003.
[5] E. M. Voorhees. The philosophy of information retrieval evaluation. In *In Proceedings of the The Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, pages 355–370. Springer-Verlag, 2002.

[1] http://www.matrixware.com
[2] http://www.ir-facility.org/

CIKM 2008 PaIR workshop, Napa Valley USA