So many topics, so little time

Giovanna Roda , Veronika Zenz
Matrixware
{*g.roda,v.zenz*}*@matrixware.com*

Mihai Lupu
Information Retrieval Facility
m.lupu@ir-facility.org

Kalervo Järvelin
University of Tampere
Kalervo.Jarvelin@uta.fi

Mark Sanderson
University of Sheffield
m.sanderson@shef.ac.uk

Christa Womser-Hacker
University of Hildesheim
womser@uni-hildesheim.de

Abstract

In the context of creating large scale test collections, the present paper discusses methods of constructing a patent test collection for evaluation of prior art search. In particular, it addresses criteria for topic selection and identification of recall bases. These issues arose while organizing the CLEF-IP evaluation track and were the subject of an online discussion among the track's organizers and its steering committee. Most literature on building test collections is concerned with minimizing the costs of obtaining relevance assessments. CLEF-IP can afford to have large topics sets since relevance assessments are generated by exploiting existing manually created information. In a cost-benefit analysis, the only issue seems to be the computing time required by participants to run (tens or hundreds of) thousands of queries. This document describes the data sets and decisions leading to the creation of the CLEF-IP collection.

1 Introduction

In 2008 an Intellectual Property (IP) track within the CLEF evaluation campaign was launched. The goal of the CLEF-IP track[1] is to investigate IR techniques in the IP domain.

The main task of the track is *prior art* search. This is one of the most common types of searches for patent examiners. The purpose of a patent is to protect a novel idea, method or product and a prior art search aims to identify any document that could prove that the idea is not novel. An ad-hoc search is often the starting point for a prior-art search.

In order to reproduce a real-life scenario where a patent examiner seeks prior art items for a given patent, the track uses as topics entire patent documents. This is unusual in IR

[1] http://www.ir-facility.org/the_irf/clef-ip09-track

experiments but a realistic task in the patent domain as the starting point for prior art search typically is an entire patent application. While the results of prior art searches include in general non-patent literature, for the first year of the track we restricted the scope of the search to patents. Patents for which no prior art was available - no citations at all or none in the target dataset - were excluded from the set of potential topics, as these would require special considerations for evaluation.

CLEF-IP also offers three facultative subtasks that use parallel monolingual queries in English, German, and French. The goal of these subtasks is to evaluate the impact of language on retrieval effectiveness.

The CLEF-IP track utilizes a collection of 2.8 million patent documents corresponding to 1.6 million patents. For each patent in the collection, a prior art search has already been conducted, the results of which are listed as *citations* in each patent document. We extracted these citations and used them to compile our IR test suite. By using citations from patent documents as relevance assessments, we exploit the manual work of highly qualified personnel (be they patent applicants, patent office searchers or employees of companies owning competing patents) to build, automatically, an IR test collection.

Out of the available data, we defined a pool of *518000* patents that are potential topics for the collection. Even after reducing the number of potential topics according to some criteria (e.g. minimum number of citations > 3, completeness of the document) we were still left with almost 16,000 patents.

How to choose a reasonable set of topics out of this pool was the subject of a discussion amongst the CLEF-IP steering committee on which we report here. Such discussion finds a starting point in previous works by Voorhees and Buckley [13], Sanderson and Zobel [11], Ritchie et al. [9], Sakai and Kando [10], Carterette et al. [4].

2 Target data and topics pool

The CLEF-IP track has at its disposal a collection of all patent documents published between 1978 and 2006 at the European Patent Office (EPO). Patents published prior to 1985 were excluded, as before this year many documents were not filed in electronic form. As for the upper limit, 2006, our data provider, a commercial institution, has not made more recent documents available.

The collection consists of 2.8 million XML files corresponding to 1.6 million individual patents adding up to 100GB of uncompressed data, or 20GB of compressed data.

Based on some earlier work on patent test collections ([8]), we split this data into two parts

1. the **test collection corpus** (or target dataset) - all documents with publication date between 1985 and 2000 (1,958,955 patent documents pertaining to 1,022,388 patents, 75GB)
2. the **pool for topic selection** - all documents with publication date from 2001 to 2006 (712,889 patent documents pertaining to 518,035 patents, 25GB)

The main task of the track is intended to mimic a real-life scenario of an IP search professional. Typically, a prior art search for an application is carried out at the EPO before a patent is granted in order to determine its novelty (in this case one also talks about *novelty search*). Since in our dataset a single patent corresponds to several files, and some of the

original application files were missing important fields, we were faced with the problem of how best to represent a topic.

Note that here we are concerned with the representation of patents and not with the issue of query formulation. On this topic, see for instance [3] on the issue of long queries and the recent study presented at the IRF Symposium in 2008 [7] on using machine learning for identifying best features for building queries.

In general, to one patent are associated several patent documents generated at different stages of the patent's life-cycle. Each publication is marked with a *kind code* that specifies the stage and gives supplementary information the kind of the publication. Each stage is denoted by a letter code, where "A" denotes patent's application stage and "B" the patent's granted stage. This code might be followed by a one-digit numerical code that gives additional information. For example the EPO uses "B1" to denote a patent specification and "B2" to mark a later, amended version of the patent specification. At each stage, the file may contain a set of fields (title, abstract, description, claims) that can be used, alone or in combinations, as topics.

To represent a topic as a set of files would have created some additional difficulties for the generation of queries. So we concentrated on simulating a real search task of a patent examiner, who initiates a prior art search with a full patent application.

Taking the highest version of the patent application's file would not have been possible because of some missing fields. For instance, for EuroPCTs patents (currently about 70% of EPO applications are EuroPCTs) whose PCT predecessor was published in English, French or German, the application files contain only bibliographic data (no abstract and no description or claims).

Often patent application documents are missing important fields, thus we decided on using the B-documents, i.e. granted patents, as the source for our topics. Even these documents are not always complete (for instance, the abstract is missing). In the end, we decided to assemble a virtual "patent file" to be used as topic by

- taking the highest version of the "B" file available;
- adding the abstract from the highest "A" file;
- and removing all citation information.

In addition to the main task, three optional subtasks dedicated to multi-lingual search are also offered. According to Rule 71(3) of the European Patent Convention [1], European granted patents must contain claims in the three official languages of the European Patent Office (English, French, and German). This data appears to be well-suited for investigating the effect of languages in the retrieval of prior art. In the three parallel multi-lingual subtasks topics are represented by title and claims extracted from the same "B" patent document.

3 Using citations to create a patent test collection

The method for generating relevance assessments for a patents' collection is described in [8]. This idea had already been exploited at the NTCIR workshop series[2]. Further discussions within the 1st IRF Symposium in 2007 [3] led to a clearer formalization of the method.

[2] http://research.nii.ac.jp/ntcir/

[3] http://www.ir-facility.org/symposium/irf-symposium-2007/the-working-groups

Our data provider delivered a list of *extended citations* for each of the patents in the pool for topic selection. Extended citations are not limited to those patents that are cited in the search report of the target patent, but also include citations from patent-family members, which are a set of related patents either via parallel applications to different patent offices (in order to obtain IP protection in different countries) or via a split of one patent application into two or more patents.

Citations are extracted from several sources [12]:

1. applicant's disclosure : some patent offices (e.g. USPTO) require applicants to disclose all known relevant publications when applying for a patent
2. patent office search report : each patent office will do a search for prior art to judge the novelty of a patent
3. opposition procedures : often enough, a company will monitor granted patents of its competitors and, if possible, file an opposition procedure (i.e. a claim that a granted patent is not actually novel).

What is to be noted when using citations lists as relevant judgments is that:

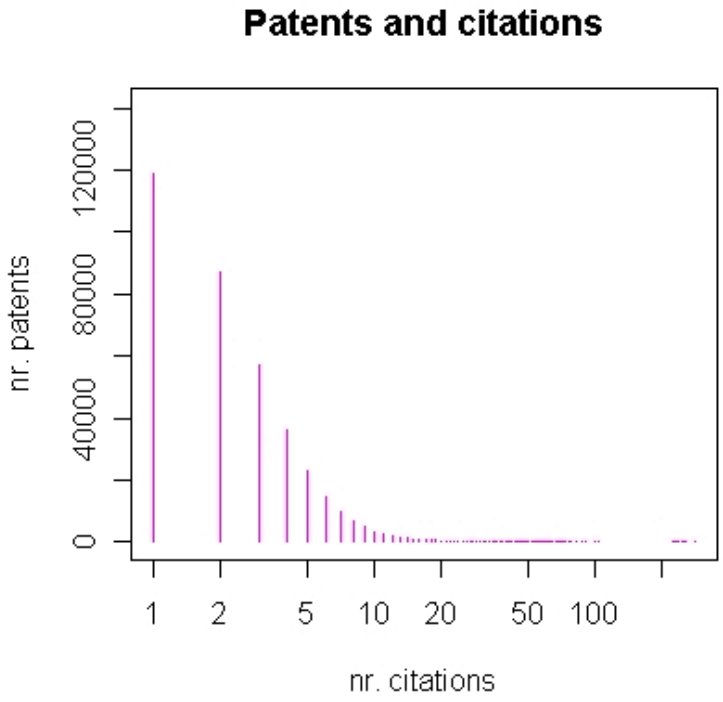
- citations have different degrees of relevancy (e.g. sometimes applicants cite not really relevant patents) This can be spotted easily by labeling citations as coming from applicant or from examiner and patent experts advise to chose patents with less than 25 - 30 citations coming from the applicant.
- the lists are incomplete: Even though, by considering patent families and opposition procedures, we have quite good lists of judgments, the nature of the search is such that it often stops when it finds one or only a few documents that are very relevant for the patent. The Guidelines for examination in the EPO [2] for example define that if the search results in several documents of equal relevance, the search report should normally contain no more than one of them. The magnitude of incompleteness will be estimated by using expert patent searchers. This means that we will have incomplete recall bases and should be taken into account in evaluation.

A number of other features of patents can be used to identify potentially relevant documents: *co-authorship* (in this case "co-inventorship"), if we assume that an inventor generally has one area of research, *co-ownership* if we assume that a company specializes in one field, or *co-classification* if two patents are classified in the same class according to one of the different classification models at different patent offices. Again, these features would require intellectual effort to consider. Therefore they are not used in the construction of the test collection for CLEF-IP 2009, but are interesting avenues to consider in the future.

4 Criteria for topics selection

These are in short the criteria used for the selection of candidate topics

1. availability of granted patent
2. full text description available
3. at least three citations
4. at least one highly relevant citation



The first criteria restricts the pool of candidate topics to those patents for which a granted patent is available. This restriction was imposed in order to guarantee that each topic would include claims in the three official languages of the EPO: German, English and French. In this fashion, we are also able to provide topics that can be used for parallel multi-lingual tasks.

Still, not all patent documents corresponding to granted patents contained a full text description. Hence we imposed this additional requirement on a topic.

Finally, we choose out of the remaining candidates only those patents that have at least 3 citations.

The plot in Figure 1 shows the number of potential topics and their citations. The pool for topic selection contains approximately 120,000 patents with one citation (=1 relevance assessment), 87,000 patents with two citations, etc.

Starting from a topics pool of approximately 500,000, we reduced the candidate topics to 80,000 by applying criteria (1). The reason why the number of topics shrank so much is that our initial pool covers applications filed in recent years (2001 to 2006) and many of those applications are still in their examination phase. Criteria (2) did not reduce this number that much, as almost all granted patents have a description available.

Criteria (3) reduced the number of potential topics from 80,000 to 31,000 patents.

Finally, we required the topics to have at least one highly relevant document in our data collection, i.e. at least one citation with category X or Y. In the end, we were left with almost 16000 patents in the pool for topic selection fulfil all these criteria.

At the time of this writing, a training set containing 500 topics with the corresponding relevance assessments has already been released. Ten of these 500 topics were chosen to be patents who have at least a citation originating from an opposition procedure. We had to

add these topics with highly relevant citations manually because they would not have been likely to appear in a random choice of patents (being less than 1% of the pool).

5 How many topics?

Evaluation methods should be

- robust (or resistant to errors in the results produced by deviations from assumptions (e.g. the of normality of data)
- reliable (yielding consistent results over repeated tests of the same subject under identical conditions)

While large number of topics should guarantee robustness (by the central limit theorem), reliability is also dependent on the quality of relevance judgements ([14]).

Limiting the number of topics and of relevance judgements while guaranteeing sufficient robustness and reliability has been the main concern of test collection builders (see [13], [11], [5], [4]). This is due to the high cost associated with obtaining relevance assessments.

The CLEF-IP test collection gets relevance assessments for free! So why not just use them all?

Well, with so many topics the time needed to run the queries, transfer and process the data might become an issue. A good compromise would be to release 10,000 topics. One week to process queries at 1 minute per topic is not terribly unreasonable and scientifically one would potentially gain a data set on which to conduct experiments on whether this number of topics is too much.

On the other hand, if it is cheap to produce a much larger recall base, say for 100,000 topics, why not make that available as well for those who would like to play with it? Or even make available to track participants different topics sizes: S, M, X, XL ? After all, some participants may use distributed systems to process queries and it will be interesting to see how is it best to distribute the collection in order to maximize the likelihood that all answers are on the same machine. A potential avenue for investigation in this sense is, again, using the manual labor that patent offices put into classifying patents.

However, if we do decide on a fixed number, a random sample of 10,000 would contain a reliable distribution of topics with varying sizes of recall bases. If there are several topic languages, the combined distribution of languages and recall bases sizes would still be reliable. The advantage in having a set of topics of the order of several thousands is that one does not need to impose strict requirements on the number of citations (e.g. > 5), as it was done at NTCIR. Even if performance on individual topics with only some citations varies, their great number guarantees reliable results on performance.

Based on the experience from NTCIR Patent task, it appears that the number of "judgment points" of a test collection defined as the "average number of relevant documents per topic" times the number of topics is a good way of measuring how feasible a test collection is for a stable and sensitive evaluation. For example, in traditional TREC/CLEF/NTCIR test collections using news documents, the average number of relevant documents per topic

is about 30 or more. If it is, for example 30 and the collection has 50 topics, the number of judgment points is 30 x 50 = 1500. However, for the patent retrieval, the majority of the topics has only one citation, i.e. one relevant document and this results in just 1 x 1000 = 1000 judgement topics for a collection with 1000 topics.

Choosing a large number of topics for a patent test collection is scientifically sound. The large number of topics would make up for a low average in citations. Selecting a set of topics with larger average number of citations would make less as longer lists of citations tend to include less relevant documents..

This is less of an issue for European patents, where there is no duty of candor for the applicant to supply references, and therefore the average "relevance" of the citations is higher (see [6]).

6 Conclusion

For all IR research, creating a meaningful, extensive and complete test collection is still a work in progress. We found that there exists a large repository of manually created data from the field of intellectual property protection, and in particular from patents. A collection of patents, while not covering abstract facts, covers a very wide array of industries, in many languages, and benefits from careful manual processing from inventors, patent officers and lawyers. In principle, the IR community only has to collect this already existing information and use it.

Unfortunately, nothing is as easy as it seems. This large repository is distributed in different countries, with little or no standardization. Even within the same patent office, standards have changed over the years and data that existed only in hard copies is, at best, provided only as facsimiles. Furthermore, the test collection based on patents citations is potentially incomplete because a search is stopped as soon as a document is found to invalidate an application.

We must then rely on two methods to make good use of this test collection: We must either use a large number of topics, to make sure that comparing different IR systems is statistically relevant; or, we must ask again patent experts to search for a complete set of relevant documents for each (rejected) patent.

This paper discusses such methods, as we plan to design and implement them for the CLEF-IP 2009 track. We must balance the desire to run as many queries as possible in order to get statistical significance, with the inherent hardware and software limits of each participant's system to handle massive data.

We must also aim to improve and overcome the existing limits of the patent test collection. Therefore we are planning to have at least a small portion of the test collection reviewed by patent experts. This will allow us to draw conclusions on how incomplete the automatically generated recall base really is.

7 Acknowledgements

The authors would like to thank all the members of the CLEF-IP steering committee. Their comments on topic selection were the subject of an email conversation that constitute the basis of this paper. Particular thanks go to Gianni Amati, Noriko Kando, John Tait and

Erik Graf. Thanks also to Florina Piroi who has recently joined the CLEF-IP team for her help with XML data manipulation.

References

[1] *European patent convention (EPC)*, http://www.epo.org/patents/law/legal-texts.

[2] *Guidelines for Examination in the European Patent Office*, http://www.epo.org/patents/law/legal-texts/guidelines.html, 2009.

[3] M. Bendersky and W. B. Croft, *Discovering key concepts in verbose queries*, Proceedings of the 31st ACM SIGIR Conference, 2008.

[4] B. Carterette, V. Pavlu, E. Kanoulas, J. A. Aslam, and J. Allan, *If I had a million queries*, Proceedings of ECIR, 2009.

[5] B. Carterette and M. D. Smucker, *Hypothesis testing with incomplete relevance judgments*, Proceedings of the CIKM, 2007.

[6] P. Criscuolo and B. Verspagen, *Does it matter where patent citations come from? inventor versus examiner citations in european patents*, Research Memoranda 017 **11** (2005), no. 5.

[7] W. B. Croft and X. Xue, *Prior art searching*, IRF Symposium Exposition, 2008.

[8] E. Graf and L. Azzopardi, *A methodology for building a patent test collection for prior art search*, Proceedings of the Second International Workshop on Evaluating Information Access (EVIA), 2008.

[9] A. Ritchie, S. Teufel, and S. Robertson, *Creating a test collection for citation-based IR experiments*, Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, 2006.

[10] T. Sakai and N. Kando, *On information retrieval metrics designed for evaluation with incomplete relevance assessments*, Inf. Retr. **11** (2008), no. 5.

[11] M. Sanderson and J. Zobel, *Information retrieval system evaluation: effort, sensitivity, and reliability*, Proceedings of the 28th ACM SIGIR Conference, 2005.

[12] Henk Tomas, personal communication, 2008.

[13] E. M. Voorhees and C. Buckley, *The effect of topic set size on retrieval experiment error*, Proceedings of the 25th ACM SIGIR Conference, 2002.

[14] J. Zobel, *How reliable are the results of large-scale information retrieval experiments?*, Proceedings of the 21st ACM SIGIR Conference, 1998.