Building Queries for Prior-art Search

Parvaz Mahdabi, Mostafa Keikha, Shima Gerani
Monica Landoni, and Fabio Crestani
{parvaz.mahdabi, mostafa.keikha, shima.gerani
monica.landoni, fabio.crestani}@usi.ch

Faculty of Informatics, University of Lugano, Lugano, Switzerland

Abstract. Prior-art search is a critical step in the examination procedure of a patent application. This study explores automatic query generation from patent documents to facilitate the time-consuming and labor-intensive search for relevant patents. It is essential for this task to identify discriminative terms in different sections of a query patent, which enable us to distinguish relevant patents from non-relevant patents. To this end we investigate the term distribution of words occurring in different sections of the query patent and compare them with the rest of the collection using language modeling estimation techniques. We experiment with term weighting based on the Kullback-Leibler divergence between the query patent and the collection and also with parsimonious language model estimation. Both of these techniques promote words that are common in the query patent and are rare in the collection. We also incorporate the classification assigned to patent documents into our model, to exploit the available human judgements in the form of a hierarchical classification. Experimental results show the effectiveness of generated queries particularly in terms of recall while patent description showed to be the most useful source for extracting terms.

1 Introduction

The objective of prior-art search in patent retrieval is identifying all relevant information which can invalidate the originality of a claim of a patent application. Therefore all patent and non patent literature that have been published prior to the filing date of a patent application in question need to be searched. An invention is patentable when it is found to be an original creation and no record of similarities with already patented inventions is found. As shown in [2] the most executed type of search in patent domain is novelty and patentability search. In this type of search a patent examiner is required to find all previously published materials on a given topic, because even missing one relevant document can lead to a multi million Euro law suit due to patent infringement. Thus patent retrieval is considered as a recall-oriented application.

There are few issues which make prior-art search a challenging task and different compared to other search tasks such as web search. The first issue is that the starting point of the prior-art task is a patent document in question. Since

in this task the information need is presented by a patent document rather than short queries, the major challenge is how to transform the patent application into search queries [5, 21, 17]. A variety of different techniques have been employed in previous studies for identifying effective query terms mainly looking into distribution of term frequency. The second issue is that the vocabulary used in patent applications is very diverse. Writers tend to purposely use many vague terms and expression and non-standard terminology in order to avoid narrowing down the scope of their invention [3]. They also develop their own terminologies to increase their acceptance chances in patent examination procedure. Previous works [5, 9] show the effectiveness of incorporating International Patent Classification (IPC) classes for knowledge extraction. Another special characteristic of patent documents is their structural information. Patent documents have different fields such as title, abstract, description, and claim. Different fields use different type of language for describing the invention. Abstract and description use a technical terminology while claim field uses a legal jargon [21].

In this paper we explore generating queries from different fields of the patent documents. Our contribution is building an effective term selection and weighting technique using a weighted log likelihood based approach to distinguish words which are indicators of the topic of the query and are not extensively used in the collection. We also investigate query modeling based on parsimonious language model for building the topic of the query patent. Furthermore we utilize the knowledge embedded in IPC classes in our model. This will address the vocabulary mismatch since we include words in query which are not present in the query topic itself.

The rest of the paper is structured as follows. We first explain the CLEP-IP 2010 collection and some recent work on patent retrieval. We then present an overview of our approach. We define the query generation problem and describe three query modeling approaches for estimating the topic of the query patent. Finally, we describe the empirical comparison we performed between different query modeling methods for the prior-art task of CLEF-IP 2010.

2 CLEF-IP 2010 Collection

Patent collection released for prior-art search of CLEF-IP 2010 constitutes of 1.3 million patent documents from EPO (European Patent Office). Collection has multilingual nature in which patent documents can be in English, French and German. Each patent application have one or more IPC classes assigned to them, where International Patent Classification (IPC) exhibits a hierarchical order consisting of more than 70,000 subdivisions. This classifications show the technological aspects of the described invention. These assignments are performed by patent examiners and are used by all patent offices [9]. Patent documents come in different versions which corresponds to the different stages of the patents's life cycle and are referred to as kind codes [15]. In the relevance judgements released for this task, different kind codes of a patent are expected to be found.

The structure of a prior-art task topic is as follows:
<topic>
<num>PAC-*number*</num>
<narr>Find all patents in the collection that potentially invalidate patent application *patentNumber*. </narr>
<file>*fileName.xml* </file>
</topic>

As mentioned before, the information need is represented by a document rather than a query, so participants have to first generate the query from the patent document.
The CLEF-IP training set contains documents and relevance judgements for 300 topics. The test set consists of two sets, one with 500 and one with 2000 topics referred to as small and large test set. We performed our experiments on the english subsection of the collection and on the large topic set.

3 Related Work

In the third NTCIR workshop [7], a patent retrieval track was first introduced and few patent test collection were released. Starting from the fourth NTCIR, a search task related to the prior-art search was presented which was referred to as invalidity search run. The goal was to find prior-art before the filing date of the application in question which conflicts with the claimed invention. This type of search task is performed when a party is accused of infringement. For the purpose of the search task in NTCIR, queries were generated from the claim fields of the patent documents. Participants were asked to find the patents and passages associated with the query topic.

CLEF-IP is another important evaluation platform for comparing performance of patent retrieval systems. CLEF-IP has been running since 2009 and participants have explored standard information retrieval approaches in this domain. In prior-art Task defined in Clef-IP 2010, participants were asked to find the prior-art for a given patent application [16]. Definition of the task is closely related to the actual prior-art searches of patent examiners, where they search for patent documents relevant to submitted patent applications [9]. The citation parts of the applications are removed and counted as relevant documents used for evaluation of results [4, 17]. This type of automatic evaluation has been performed in NTCIR as well. Results were evaluated using three metrics: mean average precision (MAP), Recall, and the Patent Retrieval Evaluation Score (PRES) [11]. The last evaluation metric is suitable for the recall-oriented applications by taking into account both recall and the users's expected search effort.

In the following we provide an overview over the two better performing and more innovative approaches seen in CLEF-IP 2010. The first group [9] constructed a restricted initial working set by exploiting the citation structure and IPC metadata. They showed that by combining the citation-based information

with the text-based information better performance can be achieved. The second ranked group [10] used a simple technique for query generation by using the most frequent unigrams and bigrams. They reported that this simple approach outperforms their previous attempt to use the structural information of the patent document. Similar approaches were employed by other groups [1, 20].

4 System Architecture

The retrieval system starts with a query patent document, which we aim to find relevant documents for, and generates a rank list of patent documents. Figure 1 shows the overall architecture of our system for prior-art search. In the first step we need to generate a query from the patent document. In the second step we formulate the query by selecting top k terms from the term distribution of the query model. In the third step we retrieve documents relevant to the generated query. We then filter this ranked list by excluding documents which do not have any IPC class in common with the query patent document. In the next

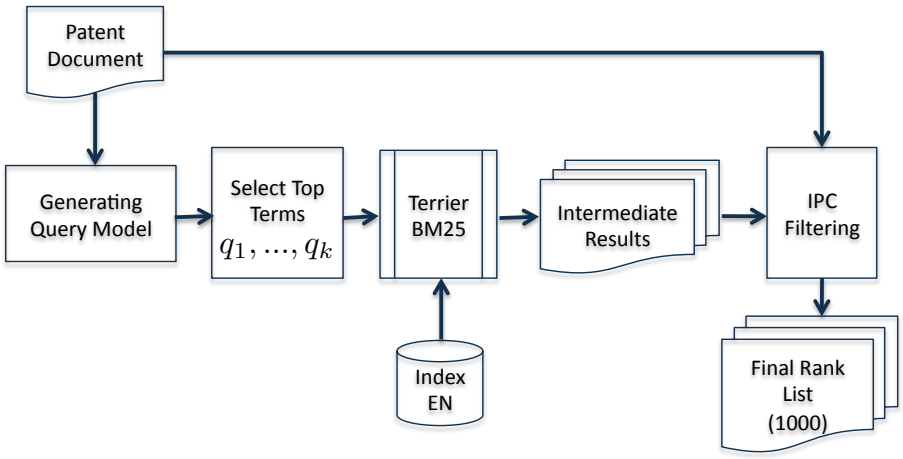

Fig. 1. Overall architecture of the proposed system

section we focus on the query generation problem and propose three methods for performing this step. We limit our experiments to the English subset of the collection. We do not take advantage of the citation information to see what can be gained using only the text information of patent documents.

5 Query Generation for a Query Patent Document

The *query generation for prior-art search* problem is the problem of developing an effective algorithm that selects the best terms from the whole patent document to form an effective query. An effective query is defined as a query that can better distinguish relevant patents from non relevant patents. We implemented two different approaches based on Weighted Log-Likelihood [13] and one approach based on parsimonious language modeling [6] to estimate the query model of a patent document. The goal of all these approaches are to select the most informative terms for representing the topic of the query patent. These approaches will be discussed in more details in the following.

We utilize the structural information of a patent document in our model by building a query model for each field separately. A patent document in CLEF-IP 2010 collection contains the following fields: the title (ttl), the abstract (abs), the description (desc), and the claim (clm). Our aim is to investigate and compare the quality of the extracted terms according to the query model of each field. In an attempt to take into account the full structure of the document, we also explore merging different ranked lists.

5.1 Query Model Based on Weighted Log-Likelihood

In the first approach we build a query model for the field f of the patent document (denoted θ_{Q_f}), where f belongs to {title, abstract, description, claim}. We estimate the query model θ_{Q_f} by calculating the relative frequencies for terms in the query document. To have a better representation, we smooth the θ_{Q_f} estimate with the topic model of a relevant cluster. This cluster consists of documents with at least one IPC class in common with the query document (denoted $RIPC$). The intuition is that patent documents with similar IPC classes are assumed to have similar topics. This smoothing the parameters away from their maximum likelihood estimates is necessary and it helps us to exploit the knowledge embedded in the IPC hierarchy into our model. In other words, this can be seen as expanding the document model with the IPC metadata. This will assign non zero probability to words which are associated with the topic of a document and are not mentioned in the document itself.

$$P(w|\theta_{Q_f}) = \lambda \frac{tf(w, Q_f)}{|Q_f|} + \frac{(1-\lambda)}{N} \sum_{d \in \texttt{RIPC}} \frac{tf(w, D_f)}{|D_f|} \tag{1}$$

$tf(w, Q_f)$ denotes the term frequency of the word w in the patent document Q_f, $|Q_f|$ is the length of the patent document Q_f, and N denotes the size of the relevant cluster $RIPC$. In order to estimate a query model for the patent in question it is necessary to highlight words from the term distribution of θ_{Q_f} which are rare in the collection. To this end, we weight term probabilities in θ_{Q_f} with the following formula:

$$P(w|LLQM_f) \propto p(w|\theta_{Q_f}) log \frac{p(w\theta_{Q_f})}{p(w|\theta_{C_f})} \tag{2}$$

where $P(w|\theta_{C_f})$ shows the probability of a word in the collection and is estimated as follows:

$$P(w|\theta_{C_f}) = \frac{tf(w, C_f)}{\sum_{d \in \mathtt{C}} |d_f|} \tag{3}$$

where $tf(w, C_f)$ denotes the collection term frequency for the field f. We refer to this model as the Log-Likelihood Query Model $LLQM_f$. This is a slight variation of the standard Weighted Log-Likelihood ratio [13]. The value in Equation 2 is normalized by a constant which is the Kullback-Leibler divergence [12] between Q and the collection language model. This measure quantifies the similarity of the query document with the topical model of relevance and the dissimilarity between the query document and the collection model. Terms which cause high divergence are therefore good indicators of the patent document and show the specific terminology of the patent document.

In the second approach, in order to incorporate the knowledge of the hierarchical classifications of IPC into our model, we estimate a slightly different formulation of the query model referred to as Cluster Based Query Modeling $CBQM_f$ by weighting term probabilities in θ_{Q_f} by their relative information in cluster language model and the collection language model. So this model assigns a high score to query terms which are similar to the cluster model while dissimilar to the collection background model [13]. We base this estimate on the divergence between θ_{Q_f} and cluster language model and we measure this divergence by determining the log-likelihood ratio between θ_{Q_f} and cluster language model, normalized by the collection C. This formulation gives another way of constructing query model based on the relevant cluster derived from IPC classes.

$$P(w|CBQM_f) \propto p(w|\theta_{Q_f}) log \frac{p(w|\theta_{Cl_f})}{p(w|\theta_{C_f})} \tag{4}$$

5.2 Parsimonious Query Modeling

In the third approach we estimate a query model that differentiate the language use of the query patent from the collection model. As suggested by Hiemstra *et al.* [6], we estimate the topic of the query patent following the parsimonious language modeling, by concentrating the probability mass on terms that are indicator of the topic of the query patent and are dissimilar from the collection model. We use EM-algorithm for estimating the query model of different fields of a patent document. Parsimonious Query Model PQM_f is estimated according to the following iterative algorithm:

E-step:

$$e_t = tf(t, Q_f).\frac{\lambda P(w|PQM_f)}{(1-\lambda)P(t|C_f) + \lambda P(w|PQM_f)} \tag{5}$$

M-step:

$$P(w|PQM_f) = \frac{e_t}{\sum_t e_t}, \text{ i.e. normalize the model} \quad (6)$$

$P(t|C_f)$ is the maximum likelihood estimate for the collection and is calculated according to Equation 3. The initial value for $P(w|PQM_f)$ is based on the maximum likelihood estimate for the query as in Equation 1, ignoring the smoothing part. The advantage of this estimation model is that it can discard field specific stopwords automatically. This is because we estimate the query model for each field separately. For example for the abstract field the set of words such as "system", "device", "apparatus", and "invention" are identified as stopwords.

6 Experimental Results

In this section we first explain our experimental setup for evaluating the effectiveness of our proposed methods. We then explain the experiments that we conducted in order to evaluate the effectiveness of different setting of the proposed methods in section 5.

6.1 Experimental Setup

We index the collection using Terrier[1]. Our pre-processing is minimum and involves stop word removing and stemming using Porter stemmer. For all our retrieval experiments we use the BM25 implementation of Terrier. In our experiments, we compare our term selection techniques with other participants of CLEF-IP 2010 [14].

6.2 Parameter Settings

We select the top k terms from generated query models and submit them as weighted query to Terrier using BM25 retrieval function. We then filter the retrieved results based on IPC classes. The proposed models have two parameters: the field f of query patent used for building the query model and the parameter k which shows the number of selected terms from each field. We first tune this parameters on the training set. Note that title has on average 10 words and abstract is limited to 50 words, while the descption and claim can be lengthy. So the range of the query length parameter for fields are different. Smoothing parameter λ in LLQM and PQM is experimentally set to 0.9.

6.3 Effect of Query Length and Field

Tables 1 to 4 show the result of selecting different number of terms from different sections of the patent document with three approaches introduced in previous section on the training set.

[1] http://terrier.org/

PQM(desc)	25	50	75	100	125	150
MAP	0.08	0.09	0.09	**0.10**	0.10	0.09
Recall	0.56	0.57	0.59	**0.59**	0.58	0.57
CBQM(desc)	25	50	75	100	125	150
MAP	0.08	0.09	0.10	**0.11**	0.10	0.09
Recall	0.58	0.59	0.60	**0.60**	0.59	0.59
LLQM(desc)	25	50	75	100	125	150
MAP	0.08	0.08	0.11	**0.12**	0.12	0.11
Recall	0.59	0.62	0.62	**0.63**	0.61	0.60

Table 1. Evaluation scores of LLQM, CBQM, and PQM, showing the effect of the number of selected terms extracted from the description field on the training set for the English subset

The results on all four tables show that increasing the query length improves the evaluations scores. But when the query length exceeds 100, adding more candidate query terms does not improve the performance anymore. This fact is valid for all the three query estimation methods. So based on these experiments we limit the length of the generated queries from description, claim, abstract, and title by 100, 100, 50, 10, respectively. We see that LLQM outperforms CBQM and PQM in terms of MAP and Recall.

PQM(clm)	25	50	75	100	125	150
MAP	0.04	0.05	0.06	**0.07**	0.07	0.07
Recall	0.48	0.50	0.52	**0.54**	0.53	0.52
CBQM(clm)	25	50	75	100	125	150
MAP	0.05	0.06	0.06	**0.07**	0.06	0.06
Recall	0.49	0.52	0.53	**0.56**	0.54	0.52
LLQM(clm)	25	50	75	100	125	150
MAP	0.06	0.08	0.10	**0.10**	0.09	0.09
Recall	0.51	0.53	0.56	**0.57**	0.56	0.55

Table 2. Evaluation scores of LLQM, CBQM, and PQM, showing the effect of the number of selected terms extracted from the claim field on the training set for the English subset

Table 5 reports the performance of the three term selection techniques over different fields on the training set with the optimized query length. Experiments show that extracting terms from description field has the best performance over all other fields. The reason for this is the technical language used in description as opposed to the legal wrapping of sentences which is the characteristic of the claim field. We believe the short length of titles are the reason why selecting terms from the title performs worse compared to other fields. Prior work [22] suggests that the abstract and description both use technical terminology, but

PQM(abs)	10	20	30	40	50
MAP	0.05	0.05	0.06	0.06	**0.07**
Recall	0.47	0.48	0.50	0.52	**0.54**
CBQM(abs)	10	20	30	40	50
MAP	0.05	0.05	0.06	0.06	**0.07**
Recall	0.48	0.52	0.54	0.56	**0.56**
LLQM(abs)	10	20	30	40	50
MAP	0.05	0.05	0.06	0.07	**0.07**
Recall	0.50	0.52	0.54	0.55	**0.56**

Table 3. Evaluation scores of LLQM, CBQM, and PQM, showing the effect of the number of selected terms extracted from the abstract field on the training set for the English subset

PQM(tit)	5	10
MAP	0.03	**0.03**
Recall	0.48	**0.50**
CBQM(tit)	5	10
MAP	0.04	**0.04**
Recall	0.52	**0.53**
LLQM(tit)	5	10
MAP	0.04	**0.05**
Recall	0.52	**0.53**

Table 4. Evaluation scores of LLQM, CBQM, and PQM, showing the effect of the number of selected terms extracted from the title field on the training set for the English subset

our results based on abstract section are less effective. Further investigation is needed to understand why query terms extracted from the abstract field are not as effective as the ones extracted from the description.

The other observation is that LLQM outperforms CBQM and PQM on all settings. The reason that CBQM performed slightly worse than LLQM, is perhaps due to the fact that we consider all documents which have IPC classes in common with the query as feedback documents. This generated cluster of relevant documents is very big, therefore we loose the specific terms which are representative of the topic of the query document.

In an attempt to merge results of different sections we tried CombSUM and CombMNZ [18] but it did not improve the performance of the best setting. Similar results were found when building a single query by combining the selected query terms from different fields, therefore we did not report the results.

6.4 Comparison with the CLEF-IP 2010 participants

We fix our two parameters for the estimation method of query model, namely the query length and the query field, to the value which has been shown to

Run	MAP	R
PQM(tit)	0.03	0.50
PQM(abs)	0.07	0.54
PQM(desc)	**0.10**	**0.59**
PQM(clm)	0.07	0.54
CombSUM(all)	0.05	0.55
CombMNZ(all)	0.04	0.54
CBQM(tit)	0.04	0.53
CBQM(abs)	0.07	0.56
CBQM(desc)	**0.11**	**0.60**
CBQM(clm)	0.07	0.56
CombSUM(all)	0.09	0.57
CombMNZ(all)	0.07	0.56
LLQM(tit)	0.05	0.53
LLQM(abs)	0.07	0.56
LLQM(desc)	**0.12**	**0.63**
LLQM(clm)	0.10	0.57
CombSUM(all)	0.09	0.57
CombMNZ(all)	0.08	0.56

Table 5. Comparison of performance of LLQM, CBQM, and PQM over different fields of a patent document

achieve the best performance on the training set. Now we present our results following this setting on the test set. Our results on the training set show that LLQM and CBQM perform better than PQM. Thus we only present the results of these two approaches on the test set. If we would have submitted the results of LLQM approach, it would have ended up on the top-3 for the prior-art task in terms of Recall and PRES. In terms of MAP it would have been placed at rank 4. While CBQM would have been placed two ranks below LLQM.

Table 6 shows our position with respect to other CLE-IP 2010 participants according to the evaluation results. In our techniques we did not look into citations proposed by applicants and among the top ranked participants only two other approaches by Magdy and Jones [10] and Alink et al. [1] were similar to us in this aspect, which are indicated by dcu-no and spq, respectively. This is the main reason behind the well performance of the first two ranked approaches in Table 6. Our two approaches are shown with bolded fonts.

Although previous works [19, 8] mainly use claims for query formulation, our results suggest that building queries from description field can be more useful. This results are in agreement with [22] in which query generation of US patents were explored and background summary of the patent was shown to be the best source for extracting terms. Since background summary in US patents uses a technical terminology for explaining the invention, it can be considered as the equivalent of description field in European patents.

Run	MAP	R	PRES
humb[9]	0.2264	0.6946	0.6149
dcu-wc[10]	0.1807	0.616	0.5167
LLQM	**0.124**	**0.60**	**0.485**
dcu-nc[10]	0.1386	0.5886	0.483
CBQM	**0.124**	**0.589**	**0.477**
spq[1]	0.1108	0.5762	0.4626
bibtem[20]	0.1226	0.4869	0.3187

Table 6. Prior-art results for best runs in CLEF-IP 2010, ranked by PRES, using the large topic set for the English subset

7 Discussion and Future Work

Prior-art task is one of the most performed search tasks in patent domain. The information need in this task is presented by a document. Therefore converting the document into effective search queries is necessary. In this work, we presented three query modeling methods for estimating the topic of the patent application. We integrate the structural information of a patent document and IPC classification into our model. Our study suggests that description is the best section for extracting terms for building queries. Based on our experiments, combining different fields in query formulation or merging the results afterwards, did not show to be useful. In the future work, we can explore the advantage of using the citation structure and noun phrases in the proposed framework. Using a smaller cluster of similar IPC classes for estimating the topical model should be explored in an attempt to avoid adding general terms to the query and selecting more specific terms.

8 Acknowledgements

Authors would like to thank Information Retrieval Facility (IRF) for the support of this work. We would like to also thank Mark Carman for helpful discussions and valuable suggestions.

References

1. W. Alink, R. Cornacchia, and A. P. de Vries. Building strategies, a year later. *CLEF (Notebook Papers/LABs/Workshops)*, 2010.
2. L. Azzopardi, W. Vanderbauwhede, and H. Joho. Search system requirements of patent analysts. *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010*, pages 775–776, 2010.
3. S. Bashir and A. Rauber. Improving retrievability of patents in prior-art search. *32nd European Conference on IR Research, ECIR 2010*, pages 457–470, 2010.

4. E. Graf and L. Azzopardi. A methodology for building a patent test collection for prior art search. *In Proceedings of the Second International Workshop on Evaluating Information Access (EVIA)*, 2008.
5. E. Graf, I. Frommholz, M. Lalmas, and K. van Rijsbergen. Knowledge modeling in prior art search. *Advances in Multidisciplinary Retrieval, First Information Retrieval Facility Conference, IRFC 2010*, pages 31–46, 2010.
6. D. Hiemstra, S. E. Robertson, and H. Zaragoza. Parsimonious language models for information retrieval. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2010.
7. M. Iwayama, A. Fujii, N. Kando, and A. Takano. Overview of patent retrieval task at ntcir-3. *In Proceedings of NTCIR-3 Workshop*, 2002.
8. K. Konishi. Query terms extraction from patent document for invalidity search. *Proc. of NTCIR 2005*, 2005.
9. P. Lopez and L. Romary. Experiments with citation mining and key-term extraction for prior art search. *CLEF (Notebook Papers/LABs/Workshops)*, 2010.
10. W. Magdy and G. J. F. Jones. Applying the kiss principle for the clef- ip 2010 prior art candidate patent search task. *CLEF (Notebook Papers/LABs/Workshops)*, 2010.
11. W. Magdy and G. J. F. Jones. Pres: a score metric for evaluating recall-oriented information retrieval applications. *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010*, pages 611–618, 2010.
12. C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.
13. E. Meij, W. Weerkamp, and M. de Rijke. A query model based on normalized log-likelihood. *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, 2009.
14. F. Piori. Clef-ip 2010: Prior art candidate search evaluation summary. 2010.
15. F. Piori. Clef-ip 2010 track guidelines. 2010.
16. F. Piroi. Clef-ip 2010: Retrieval experiments in the intellectual property domain. *CLEF (Notebook Papers/LABs/Workshops)*, 2010.
17. G. Roda, J. Tait, F. Piroi, and V. Zenz. Clef-ip 2009: Retrieval experiments in the intellectual property domain. *Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009*, pages 385–409, 2009.
18. J. A. Shaw and E. A. Fox. Combination of multiple searches. *In TREC 1994*, 1994.
19. T. Takaki, A. Fujii, and T. Ishikawa. Associative document retrieval by query subtopic analysis and its application to invalidity patent search. *Proceedings of the 2004 ACM CIKM International Conference on Information and Knowledge Management*, pages 399–405, 2004.
20. D. Teodoro, J. Gobeill, E. Pasche, D. Vishnyakova, P. Ruch, and C. Lovis. Automatic prior art searching and patent encoding at clef-ip '10. *CLEF (Notebook Papers/LABs/Workshops)*, 2010.
21. X. Xue and W. B. Croft. Automatic query generation for patent search. *CKIM*, 2009.
22. X. Xue and W. B. Croft. Transforming patents into prior-art queries. *SIGIR*, 2009.