CLEF-IP 2010: Retrieval Experiments in the Intellectual Property Domain

Florina Piroi

The Information Retrieval Facility (IRF)
Vienna, Austria
f.piroi@ir-facility.org

1 Introduction

In the recent decade that research in IR methods for Intellectual Property domain has increased. The first efforts in observing how information retrieval is done in patent domain were done with the series of Nist workshops (see for example [2]). Lately, more workshops and conferences are dedicated to bringing together IR and IP specialists [3,7].

In 2008, the Irf obtained the agreement to coordinate two evaluation campaigns with emphasis on patent documents and prior art retrieval: Clef–Ip and Trec–Chem.

The Clef–Ip track was launched in 2009 to investigate IR techniques for patent retrieval and it was part of the CLEF 2009 evaluation campaign. In 2010, the track continued as a benchmarking activity of the Clef 2010 conference.

The track utilizes a collection of more than 1.3 million patent documents derived from Epo (European Patent Office) sources. The collection covers English, French and German with at least 150,000 documents in each language.

There were two tasks in the 2010's track. The first one is to find patent documents that are candidates to constitute prior art for a given document. The second task is to classify a given document according to the International Patent Classification system (Ipc). Relevance judgements will be produced using the patent citations for the Prior Art Candidates search task and using the recorded classification codes for the Classification task.

This notebook gives a report on the Clef–Ip activity in 2010. The paper is structured as follows: Section 2 describes the test collection used this year, section 3 presents the participating teams and gives an overview of the methods the teams involved. In the same section we also present the main measurements done in this track.

2 The 2010 Clef–Ip Data Collection

2.1 The Objects in the Collection

The Clef–Ip collection contains patents, physically stored as a collection of patent documents. A patent document may be an application document, a search report, or a granted patent document. We describe in the following some of the key terms and steps in a patent's life–cycle.

A patent is a set of exclusive legal rights for the use and exploitation of an invention in exchange for its public disclosure. The exclusive rights are given by a governing authority and are limited in time. The requirements for granting patents vary widely among patent offices, but a common first step is to file a patent application request with a patent office. For this, the applicant must supply a written specification of the invention—also called an *application document*—where the background of the invention, a description of the invention, and a set of claims which define the scope of protection, should the patent be granted, are given. The *application date*, or *filing date* of a patent refers to the date when the patent application was filed.

In order to be granted, a patent application is examined by professionals who will analyze wether it meets certain patentability criteria and wether the application complies with the relevant patent law. The most important patentability criteria are *novelty*, *inventiveness*, and *practicality*. Of relevance to the Clef–Ip benchmarking activity is the novelty criteria. A patent application satisfies the novelty requirement if no earlier patent or other kind of publication describing (parts of) the invention can be found in a reasonable amount of time. Such a search for novelty–relevant documents is called *a prior art search*. Results of a prior art search are recorded in a *search report*, and are a basis for further communication with the applicant which may result in modifications of the patent specifications before the patent is granted. The the relevant documents listed in a search report of a patent are referred to as *patent citations*. Usually, the search report and the application document are published within 18 months from the application date.

When a patent application is found to meet all the necessary legal and patentability requirements, a decision to grant the patent is made and, after further fees and procedural steps, the granted patent is published. An important procedural step at the Epo is that a translation of the claims in all three official Epo languages (English, German, French) is provided [1].

Patent documents generated at the different stages of the patent's life-cycle are identified by a country code (denoting the patent office analyzing/granting the patent), a unique numeric identifier, and by a kind code together with a version number[1]. In the case of Epo the "A" in the kind code denote a patent document published in the application phase, the "B" kind code marks a granted patent document.

It is possible to file a patent application at more than one patent office. When the same invention is granted a patent by different patent offices, the two patents are said to belong to the same *patent family*.

[1] For the EP patents, documents at different stages have the same numeric identifier. For other patent offices this is not always the case. For example, the patent document US-6689545-B2 represents a US granted patent with its application document publication number US-2003011722-A1

An important tool in organizing the large amount of patent data which patent offices regulate is the *classification system*. A patent classification system 'sorts' the patents according to the technical area they belong to, and it is a basis for a quick investigation of the state of the art in a field[2]. There are several patent classification systems, the most used being the International Patent Classification system (IPC), the European Classification System (ECLA), the US Classification System.

2.2 Technical Elements

Compared to the CLEF–IP 2009 data collection, this year there has been an increase in the number of patent documents to be included in the CLEF–IP data collection. The total number of patent documents is over 3.5 million, almost one million more documents than in 2009.

The documents in the CLEF–IP 2010 collection are extracted from the MAREC[3] data corpus, and are patent documents published by the EPO.

Following the same procedure as last year, we split the available data into two parts

1. the **collection corpus** (or target data set) contains documents with application date prior to 2002. This set contains over 2.6 million documents, representing over 1.9 million patents.
2. the **topic pool** contains documents with application date between 2002 and 2009. This set contains over 0.8 million patent documents, representing over 0.6 million patents.

The same as in 2009, the Test Collection Corpus was delivered to the participants "as is", without merging the documents related to the same patent into one document. Each patent is identified by a unique patent number-Ůa string starting with "EP" and followed by 7 digits. Corresponding to each patent is a directory containing the patent documents related to that patent. The layout is nnnnnn/nn/nn/nn/*.xml.

For example, to patent EP 0981201 corresponds the directory containing files EP-0981201-A2.xml, EP-0981201-A3.xml, and EP-0981201-B1.xml:

```
> pwd
/000000/98/12/01
> ls
EP-0981201-A2.xml EP-0981201-A3.xml EP-0981201-B1.xml
```

All documents in the CLEF–IP collection contain the following main XML fields: bibliographic data, abstract, description, and claims. Not all documents actually have content in these fields. This happens because certain EPO patent applications are internationally filed under the Patent Cooperation Treaty (PCT[4])

[2] See http://www.wipo.int/classifications/ipc/en/

[3] The MAREC data corpus is a collection of over 19 million patent documents, in XML format, made available by the IRF for research purposes.

[4] http://www.wipo.int/pct/en/

in which case, the Epo does not republish the whole patent application, but only a bibliographic entry which refers to the original application.

2.3 Tasks and Topics

There were two tasks in Clef–Ip 2010. A Prior Art Candidates Search task (Pac) and a Classification task (Cls).

The first task in this track (Pac) consisted in finding patent documents in the target collection that may invalidate a given patent application. The participants were provided with two sets of patents from the topic pool (a small set of 500 topics and a large set of 2000 topics). The task didn't restrict the language used for retrieving the documents, but participants were encouraged to use the multilingual characteristic of the collection (namely, that claims in granted patent documents are provided in three languages).

The second task in the Clef–Ip track (Cls) is a newly introduced one, and required to classify a given patent document according to the Ipc system. The classification was to be given at the subclass level. The set of classification topics contained 2000 patent documents, a different set than the one used in the Pac task.

Differently from the last year's topics, where a virtual patent document was composed with a description and claims in German, English and French, this year we have used patent application documents as topics. This means that the topic documents contain claims in only one of the three languages, with about 67% of the documents having English content, 26% German content, and 7% French content. We have placed no constraints on the choice of topics, other than one: the application documents must have content in the abstract, description and claims sections of the Xml document. The patent documents released as topics had the citation records (for the Pac task) and classification records (for the Cls task) removed from the documents.

2.4 Relevance Assessments

The relevance assessments used to evaluate the Pac submissions were obtained automatically from the patent citations stored in the collection documents. Since the average number of citations per patent in the Clef–Ip collection is low (below 4), we have looked for methods to extend the set of relevant documents per topic. For this we used an extended list of citations, where to the patents listed in the patent's search report (the direct citations), we added also the patent citations listed in the family members of the topic patent, as well as the family members of the cited patents. For a detailed explanation of the citation extraction procedure, we point the reader to the last year's track overview article [6].

The relevance assessments used to evaluate the Cls submissions were also obtained automatically from the documents that originated the Cls topics. We have extracted the Ipc codes, restricted to the subclass level, from the patent documents.

3 Submissions and Results

For both tasks, a submission consisted of a single text file with at most 1,000 answers per topic, in the standard format used for the Trec submissions. 12 participating groups have submitted a total of 25 runs to the Pac task and 27 runs to the Cls task (see Table 1). The submissions were sent to us via a ftp server.

Table 1. List of participants and runs submitted

ID	Institution		Cls	Pac
bitem	BiTeM, Service of Medical Informatics, Geneva University Hospitals	CH	7	2
dcu	Dublin City Univ. - School of Computing	IE		3
hild	Hildesheim Univ. - Information Science	DE		4
humb	Humboldt Univ. - Dept. of German Language and Linguistics	DE	1	1
insa	Lci – Institut National des Sciences Appliquées de Lyon	FR	5	
jve	Industrial Property Documentation Department, JSI Jouve	FR	3	
run	Information Foraging Lab, Radboud University Nijmegen	NL	2	2
spq	Spinque	NL	1	1
ssft	Simple Shift	CH	8	
uaic	Al. I. Cuza University of Iaşi - Natural Language Processing	RO		1
ui	Information Retrieval Group, Universitas Indonesia	ID		3
uned	UNED - E.T.S.I. Informatica, Dpto. Lenguajes y Sistemas Informaticos	ES		8

3.1 Description of the Submitted Runs

This section is based on the descriptions provided by the participants. We present here which Xml fields were used in document processing, what kind of pre– and post–processing was done, the retrieval and ranking system that was used to obtain the results, cross–language techniques involved.

⋆ The **bitem** participant has submitted runs to both Pac and Cls tasks. For both tasks, the Porter stemmer was applied, and stopwords were eliminated in a document preprocessing step.

In the Pac task, the participant has used the following fields both for index creation and query generation: title, abstract, claims, Ipc codes, applicants and inventors information. Using the Terrier platform, only one English index was created, and retrieval results were ranked using Terrier's PL scheme. Fields in other language than English were translated into English before adding them to the index using the Google translator. Topic documents in a different language than English were also translated into English with the Google translator. For the run that simulated the examiner search a post–processing step was applied where the citations provided by the applicant in the text of the document were used. The participant also experimented with using the geographical location of the applicant in the post–processing phase.

The document fields used for indexing within the CLS task are title, abstract, claims, description, applicant and citations. First a retrieval step is done, where the Google translator is used as in the PAC task. The retrieved document are given as input to the k–NN algorithm which maps them to IPC codes, which are then re–ranked.

⋆ The **dcu** group submitted runs to the PAC task. The English index used in the retrieval was created from the following fields: title, abstract, description, claims, and classification tags. The document pre–processing phase included stopwords removal, stemming and number removal. The non–English topics were translated into English using the Google translator, the IR system used for retrieving results was Indri which ranked the results using a language model and inferred networks. The post–processing step for one of the runs added the citations extracted from the topic document descriptions (i.e. applicant citations) to the list of results.

⋆ The University of Hildesheim group (**hild**) participated in the PAC task and experimented with various types of queries in the frame of an Apache Lucene based system. One English language index was created based on the patent number, title, abstract and IPC XML fields. Stopwords were removed and a Porter stemmer was also applied on both corpus and topic documents. Phrase queries are extracted from various fields in the topic document, like phrases from the title only or from title, main claim, first part of the description. The IPC codes are also used in filtering the results, by looking for results that share at least one IPC code with the topic document.

⋆ The **humb** group took part in both track's tasks with the same custom–made system (PATATRAS). The pre–processing step included citation identification in the patent's text, cleaning the inventor and applicant names, language–based tokenizations, POS–tagging, concept–tagging, key–term extraction and lematization. All patent document fields were used in the index creation. One lemma index per language and a concept index based on a self–developed terminological database GRISP were used in the retrieval experiments. The PATATRAS system combines the Lemur, Okapi BM25, and Indri retrieval engines, each acting on certain index files. Result ranking is done by BM25, Indri and SVM. The classification results were obtained with the same system, but involving the KNN classifier, and the conceptual tagging is eliminated from the pre–processing phase.

⋆ All classification results sent in by the **insa** group were obtain using the LCS2[5] classification system, using a balanced Winnow method. The text fed into the classifier was considered as a bag of words or as a bag of linguistic triples obtained by preprocessing selected fields with AGFL[6] built–in linguistic parsers (EP4IR for English, and FR4IR for French). The various training experiments were done by choosing different document fields to be considered in the training process: a)abstract and titles, b) abstracts, titles, names and addresses, c) description.

⋆ **jve** participated in the CLS task with three runs. Both in training and in the test phase the title, claims, description, and abstract (when available) were used. The first run was obtained with a SVM classifier, where documents were pre–processed by tokenization, POS–tagging, lemmatization, and a "key–phrase" tagging step, which in a patent–oriented context detects the terms that best explain the subject of the patent application. WordNet (a lexical resource for the English language) was also used in this step. The second run submitted by JSI Jouve made use of the Lemur system to index the data corpus, generating one index per language. Lemur was also used to retrieve relevant documents to the given topics. From the returned patent documents, only the classification codes were kept. The third run combined the two methods used for the first two runs.

⋆ The Radboud University group (**run**) participated in both PAC and CLS tasks. The retrieval system used for the PAC task was Lemur/Indri based. The index was created out of the title, abstract, claims, description and IPC code fields, in English only. Per topic, one hundred documents were retrieved, which were then re–ranked using regression models. The system used for the CLS task is LCS Winnow with a Lucene analyzer. Only the English abstracts were fed into the classifier. The abstracts were first processed to remove punctuation, numbers and to put all letters into lowercase, then a simple tokenizer was applied. Experiments with dependency triples in the abstract were done using the AEGIR hybrid parser.

⋆ The retrieval system used by Spinque (**spq**) is an in–house retrieval system that contains a graphical search strategy builder, and an own indexer based on MonetDB. The same system, with the same generic index was used in both PAC and CLS tasks. In both tasks, the topic documents were passed through a Snowball stemmer, with English stopwords removed. The first 26 terms given by the tf-idf algorithm were considered to be part of the query. Also, the IPC codes were used in the query creation. The retrieval step returned a list of ranked patent documents, for the PAC task, and, for the CLS task, the IPC codes attached to the ranked patent documents.

⋆ The classifier used by the Simple Shift participant (**ssft**), myClass, is a winnow like in–house implementation. The fields used in the classification process are: inventor, applicant, title, abstract, claims, and description reduced to a size of

[5] http://www.phasar.cs.ru.nl/LCS/

[6] http://www.agfl.cs.ru.nl

maximum 4k. During the training phase **ssft** made use of additional patent corpora to increase the classification precision. In other training experiments, over–sampling (copying the documents in the respective category a certain number of times) was used. A collocation[7] extraction step was done during the indexing phase.

⋆ The run submitted by the **uaic** participant to the PAC task was obtained by a Lucene based system. The English only index was created using the invention title, claims, and abstract or description (when the abstract was missing) document fields. The data corpus was split into 20 and the indexing was done in parallel on 20 machines, one split per machine. Lucene was used to extract a query from the topic document using the same document fields as for the index creation. Various boost factors were applied to the document fields used in the query.

⋆ In the PAC task, the University of Indonesia **ui** participant used a simple Lemur Indri setting. A large number of XML fields were used in the English index creation, among which we list abstract, applicant, applicant's address, abstract, claims, classification codes, descriptions. The three submitted runs extracted the query from different document fields in the topic documents: invention title and description; claims; invention title, description, and claims. The query extraction is done by the tf-idf term weighting algorithm, keeping the first 10 terms.

⋆ The **uned** group participated in the PAC task with a retrieval system based on BM25. Before indexing or retrieving, the patent documents in the collection have been joined (at XML field level) into one patent document. Four XML fields, to which stemming and stopword removal was applied, have been considered for the indexing process: title, abstract, description and claims. A separate index per language was created. The query terms are extracted from the topic documents by computing the Kullback-Leibler divergence (KLD) between the language model of the topic document and the language model of the patent collection. Experiments with field boost values were also made.

3.2 Evaluation Results

We have evaluated the submitted experiments using the most common metrics in IR. Before we ran the evaluation software, some simple clean–up of the data was done. A further important data correction was done on the experiments submitted to the CLS task. Here, we noticed that several participants have made use of IPC versions that were not used in the CLEF–IP data corpus. (The data feeds that originated the CLEF–IP documents did not carry classification symbols that were eliminated when the IPC system got revised over the time.) For this reason, we removed all entries in the result files where classification codes not occurring in the CLEF–IP 2010 corpus were listed.

For each submitted PAC experiment we computed the following measures:

[7] Collocations are concept expressed by more than one word.

- Precision, Precision@5, Precision@10, Precision@50, Precision@100
- Recall, Recall@5, Recall@10, Recall@50, Recall@100
- Map
- Ndcg
- Pres

For each submitted Cls experiment we computed the following measures:

- Precision@1, Precision@5, Precision@10, Precision@25, Precision@50
- Recall@5, Recall@25, Recall@50,
- Map
- F_1 at 5, 25 and 50.

All computations were done using the `trec_eval` 9.0 software provided by Nist, with the exception of the Pres measure, which we computed using a script provided to us by the measure's authors [8]. Figures 1 through 5 show some of the calculated measures. Detailed values for each of the mentioned measures are given in [5] and [4].

4 Final Observations

We have presented here an account of the benchmarking activities done within the Clef–Ip lab, organized in the frame of the Clef 2010 conference. The time was, unfortunately, too short to be able to do an in–depth analysis of the found results, we leave this as future work. Lack of resources was also an impediment in following some of the lines drawn at the end of the 2009 track. One of such lines to follow is intensifying contacts with IP professionals. However, we were not able to pursue this goal. We didn't forget about the conclusions drawn at the end of Clef–Ip 2009, we only postponed their realization.

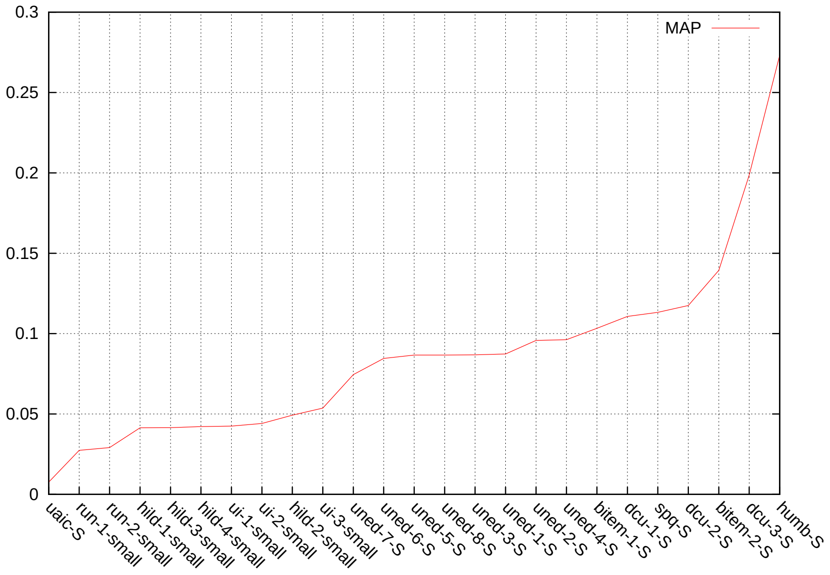

Fig. 1. MAP measures for the Pac runs

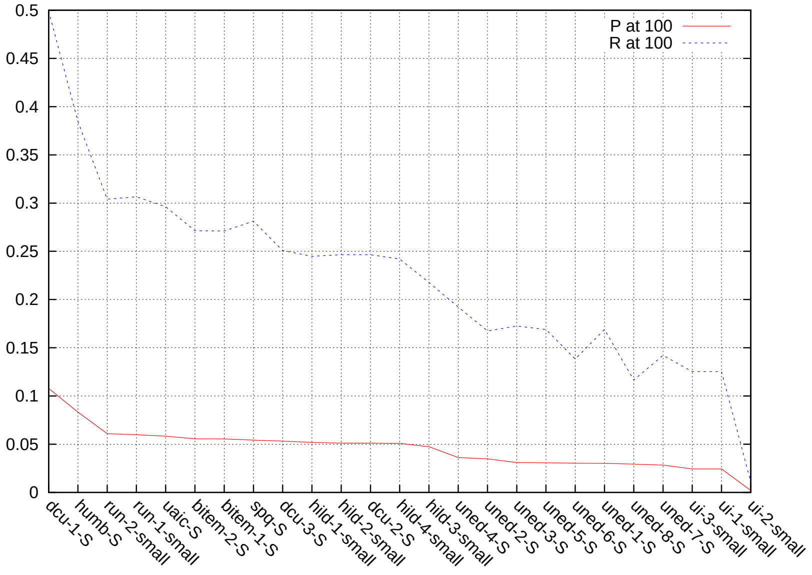

Fig. 2. Precision and Recall at 100 measures for the Pac runs

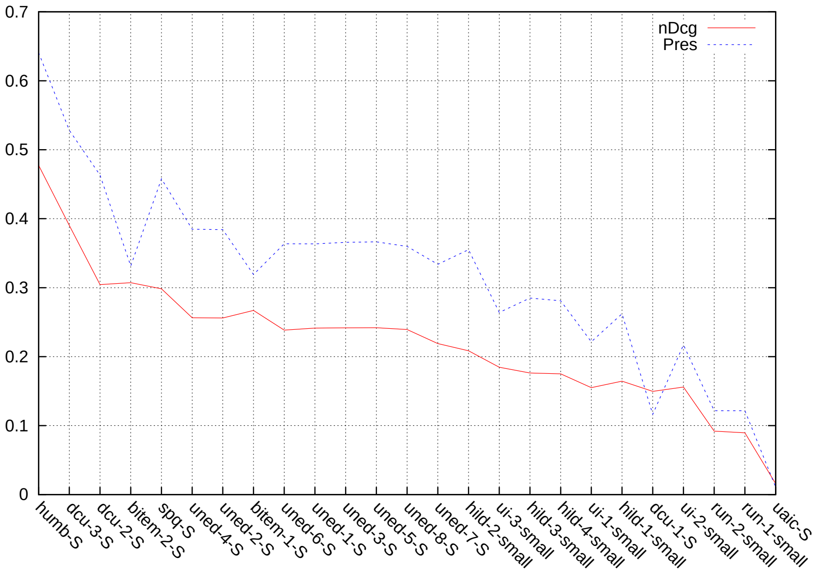

Fig. 3. Pres and Ndcg for the Pac runs

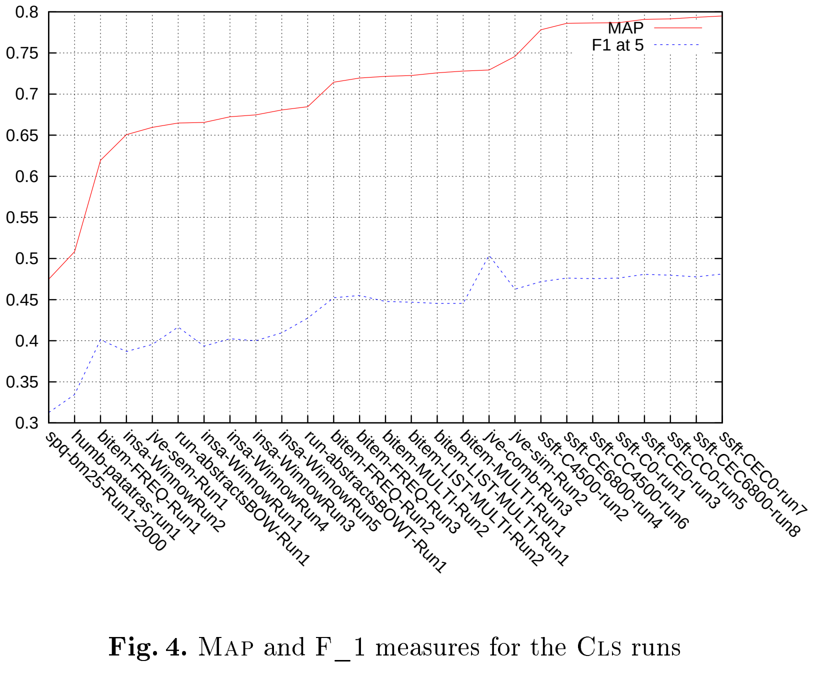

Fig. 4. Map and F_1 measures for the Cls runs

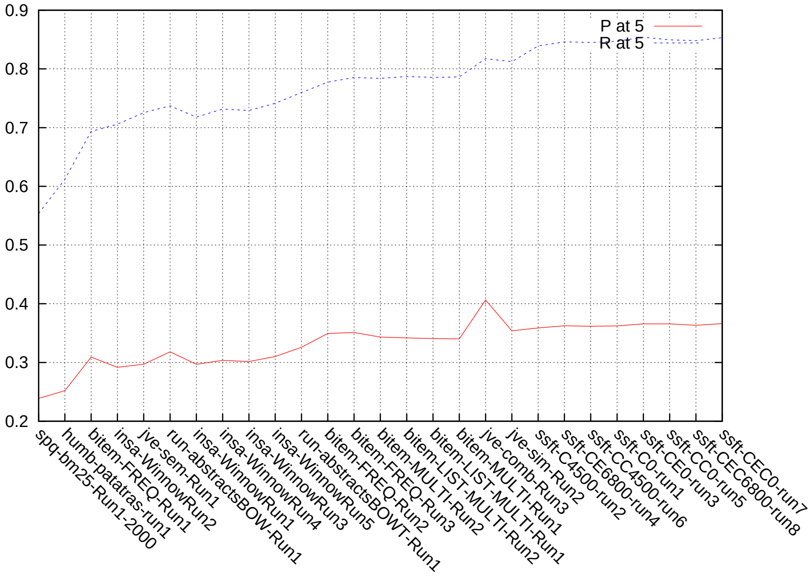

Fig. 5. Precision and Recall at 5 measures for the CLS runs

References

1. *European Patent Convention (EPC).* http://www.epo.org/patents/law/legal-texts.
2. Atsushi Fujii, Makoto Iwayama, and Noriko Kando. Overview of the Patent Retrieval Task at the NTCIR-6 Workshop. In Noriko Kando and David Kirk Evans, editors, *Proceedings of the Sixth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*, pages 359–365, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan, May 2007. National Institute of Informatics.
3. A. Hanbury, V. Zenz, and H. Berger. 1st international workshop on advances in patent information retrieval (AsPIRe'10). March 2010.
4. Florina Piroi. CLEF-IP 2010: Classification task evaluation summary. August 2010.
5. Florina Piroi. CLEF-IP 2010: Prior art candidates search evaluation summary. July 2010.
6. G. Roda, J. Tait, F. Piroi, and V. Zenz. CLEF-IP 2009: Retrieval Experiments in the Intellectual Property Domain. To appear. In *Proc. of CLEF, Revised Selected Papers*. Springer, 2010.
7. J. Tait, C. Harris, and M. Lupu. The 3rd international workshop on patent information retrieval (PaIR 2010). October 2010.
8. Magdy W. and G. J. F. Jones. PRES: A score metric for evaluating recall-oriented information retrieval applications. In *SIGIR 2010.*