CLEF-IP 2011: Retrieval in the Intellectual Property Domain

Florina Piroi[1], Mihai Lupu[1], Allan Hanbury[1], Veronika Zenz[2]

[1]Vienna University of Technology
Institute for Software Technology and Interactive Systems
Favoritenstr. 9–11/188, 1040 Vienna, Austria
{name}@ifs.tuwien.ac.at
[2]max-recall Information Systems OG
Pulverturmgasse 17/3, 1090 Vienna, Austria
veronika.zenz@max-recall.com

1 Introduction

The patent system is designed to encourage disclosure of new technologies and novel ideas by granting exclusive rights on the use of inventions to their inventors, for a limited period of time. Before a patent can be granted, patent offices around the world perform thorough searches to ensure that no previous similar disclosures were made. In the intellectual property terminology, such kind of searches are called *prior art searches*. In some industries, the number of granted patents a company owns has a high impact on the market value of the company. This underlines the importance of well-performed prior art searches.

Together with the Trec–Chem track [5], also organized by our institution, the Clef–Ip effort comes to complete the work that is being done in the series of Ntcir workshops (see for example [4]). The first Clef–Ip track ran within Clef 2009[1]. The purpose of the track was twofold: to encourage and facilitate research in the area of patent retrieval by providing a large clean data set for experimentation; to create a large test collection of patents in the three main European languages for the evaluation of cross–lingual information access. The Clef–Ip data set includes documents published by the European Patent Office (Epo) which contain a mixture of English, German and French content. The track focused on the task of prior art search.

In 2010 and 2011, the Clef–Ip track was organized as a benchmarking activity (lab) in the Clef conference. In these years, the main goal of the Clef–Ip effort remained the same – to foster research in the patent retrieval area, and provide a large clean data set. To this end, the number of tasks in the track was increased and the data set was enlarged.

Recognizing the importance of patent classifications in the daily activity of an intellectual property professional, in 2010 the Clef–Ip benchmarking activity included a patent classification task. The participants were asked to classify

[1] http://www.clef-campaign.org

given patent application documents according to the International Patent Classification (IPC) system [6].

This year (2011), in addition to the (now classic) prior art search task and the patent classification task two patent image related tasks were offered. Often, patent applications contain images that clarify details about the invention they describe. Images in patents may be drawn by hand, by computer, or both, may contain text, and are generally black-and-white (i.e. not even monochrome). Depending on the technological area of a patent, images may be technical drawings of a mechanical component, or an electric component, flow–charts if the patent describes, for example, a work–flow, chemical structures, tables, etc. When a patent expert browses through a list of search results given by a search engine, he or she can very quickly dismiss irrelevant patents to the patent application by just glancing at the images in the retrieved patents. The number of documents to be looked at in more detail is thus greatly reduced. With the Image-based Document Retrieval and Image-based Classification tasks in CLEF–IP we try to make this aspect of an IP expert's daily work familiar to the research communities

12 international teams have participated this year, we present here an overview their work and research results. 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 2011 CLEF–IP Collection

2.1 The Documents in the Collection Corpus

The CLEF–IP collection contains patents, physically stored as a collection of XML files encoding patent documents. A patent document may be an application document, a search report, or a granted patent document. Each patent document is assigned a kind code, which appears as a suffix to the patent identifier (e.g. EP-nnnnnnn-A1, WO-nnnnnnnnnn-A2). In the case of the EPO, patent application documents that include a search report carry the code A1, patent application search reports carry the code A3, granted patent documents carry the code B1, etc.[2] For a description of key terms and steps in a patent's life–cycle see [6].

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 [1,6]. The mostly used patent classification systems are the International Patent Classification system (IPC), the European Classification System (ECLA), the US Classification System. In the CLEF–IP lab, the patent classification tasks make use only of the IPC system.

[2] A list of kind EPO kind codes is listed at https://register.epo.org/espacenet/help?topic=kindcodes.
Kind codes used by the WIPO are listed at http://www.wipo.int/patentscope/en/wo_publication_information/kind_codes.html

The 2011 CLEF–IP data collection is based on the 2010 data, and is extracted from the MAREC[3] data corpus. The CLEF–IP collection contains mainly patent documents published by the EPO.

Two important additions were made to the 2010 collection corpus. The first one was to include in the distributed corpus certain patents published by the World Intellectual Patent Organization (WIPO). A high percentage of the EPO patents contained in the CLEF–IP corpus are patent applications internationally filed under the Patent Cooperation Treaty (PCT [2]) in which case, the EPO does not republish the whole patent application, but only a bibliographic entry referring to the original application. For these patents we added their WIPO equivalent to the CLEF–IP collection, in order to provide the participants a collection that is both larger and more realistic. The second addition to the CLEF–IP collection regards one of the new patent image–based tasks, the Patent image–based retrieval task. For this task we have added to the CLEF–IP target collection the patent images for three IPC subclasses: A43B, A61B, and H01L.

In number of documents, adding the WIPO patent documents to the collection corpus increased with 1.2 million patent documents, to a final number of approximately 3.5 million XML documents, referring to approximately 1.5 million patents. The images corresponding to the 47 thousands XML documents in the three IPC subclasses in the IMG–PAC task occupy 5.4 GB for 290880 tiff files.

The same as in the previous years, 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– a string identifying the publishing office ("EP" for the EPO and "WO" for the WIPO) followed by a series of digits. Corresponding to each patent is a directory containing the XML documents representing the patent documents related to that patent. For the EP patents, the layout is 00000n/nn/nn/nn/*.xml, where the sequence of digits in the directory layout corresponds to the one in the patent number. For the WO patents, the layout is 00nnnn/nn/nn/nn/*.xml, where the first 4 digits (after '00') represent the document publication year.

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

To patent WO 1994030029 corresponds the directory containing the file WO-1994030029-A1.xml:

```
> pwd
WO/001994/03/00/29
> ls
WO-1994030029-A1.xml
```

[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.

The patent image files are stored as tif files in one separate folder, the correspondence between the image file and its XML file is established with a small set of rules.

All textual 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. The content of the various XML fields can be English, German, or French. Some fields may occur more than once, each time with a different language. The XML patent file have also a document language (English, German or French), this not excluding that its subfields occur with a different language attribute than the document language. For example, granted EP patent documents (EP-nnnnnnn-Bn.xml) must contain claims in three languages (English, German and French).

2.2 Tasks and Topics

5 tasks were proposed to the participants: Prior Art Candidates Search (PAC), Patent Classification (CLS1), Refined Patent Classification (CLS2) Patent Image-based Retrieval (IMG–PAC), and Patent Image-based Classification (IMG–CLS). The topics for each of the tasks were chosen from the same topic pool we have used in 2010. We will detail each of the proposed tasks in the following.

Prior Art Candidates Search. 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 one set of 3973 topics. The topics were formulated as 'Find all patents in the collection that potentially invalidate patent application EP-nnnnnnn-An.', where the XML file storing the patent application document EP-nnnnnnn-An (which we call the *topic file* or *topic patent*) was given in an attached archive. A third of the topic files' document language was English, another third was German, and the last third was French. The task did not 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). A small set of 300 training topics was also provided, and participants were allowed to use the CLEF–IP 2010 topics sets in their system training.

Patent Classification. The second task in the CLEF–IP track (CLS1) required to classify a given patent document according to the IPC system. the topics were formulated as 'Classify patent document EP-nnnnnnn-An according to the IPC system.', where the XML file storing the patent application document EP-nnnnnnn-An was given in an attachment. The IPC system is hierarchically organized into sections, classes, subclasses, groups and subgroups. The classification was to be given at the subclass level. The set of classification topics contained 3000 patent documents, a different set than the one used in the PAC task. Again, the task didn't restrict the language used for classification, but the topic language was English for one third of the topics, German for the next third, and French for the last third of the topics. Participants could use the Clef–Ip collection corpus as a training set.

Refined Patent Classification. This task is, at least in formulation, very similar to the Cls1 task. It required the participants to classify given patent documents according to the Ipc system. The subclass was given, the participants had to return the group/subgroup classification levels.

Patent Image-based Prior Art Search. This task (Img–Pac) was introduced as a pilot task. It has the same aim as the Pac task, except that the images and Xml files corresponding to the patents were available and were to be used. Only the three Ipc sub–classes listed above were used, as for these classes, patent searchers often rely on visual comparison of images in the patents to find relevant prior art. The queries consisted of the text and complete set of images of 211 patents, with the topics formulated in the same way as for the Pac task.

Patent Image-based Classification. The aim of the image classification task (Img–Cls) was to automatically classify patent images based on visual content. For the Img–Cls task, only images extracted from patents, not full patents, were provided. Participants to this task did not need the full Clef–Ip corpus. The classification was into nine classes: drawing, chemical structure, program listing, gene sequence, flow chart, graph, mathematics, table, and symbol. Training data with between 300 and 6,000 training images for each of these classes was provided, and only these data were to be used to train image type classification techniques. The task was to train a classifier using the provided training data, and test the resulting classifier on a set of 1,000 patent images.

2.3 Relevance Assessments

The relevance assessments used to evaluate the Pac and Img–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 detailed explanations of the citation extraction procedure, we point the reader to the overview article [7].

The relevance assessments used to evaluate the Cls1 and Cls2 submissions were also obtained automatically from the documents that originated the Cls topics. We have extracted the Ipc codes, restricted to the subclass level, and group/subgroup level respectively, from the patent documents.

The relevance assessments for the Img–Cls topics were done manually by us.

3 Submissions and Results

With the exception of the Img–Cls task, 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. The submissions to the Img–Cls task consisted of a single text file with one entry per topic, each entry containing the topic id and nine space separated values one for each class used in the classification. 12 participating groups have submitted a total of 77 runs, (unequally) distributed over the five proposed task. Table 1 shows the list of participating groups, marking the tasks where runs were submitted. The submissions were uploaded to the Direct system [3][4].

Table 1. List of participants and runs submitted

ID	Institution		Pac	Cls1	Cls2	Img–Pac	Img–Cls
chemnitz	Chemnitz University of Technology, Retrieval Group	DE	x				
hildesheim	Hildesheim Univ. - Information Science	DE	x				
hprussia	Hewlett-Packard Labs, Russia	RU	x				
hyderabad	International Institute of Information Technology - SIEL	IN	x				
joanneum	Joanneum Research, Institute for Information and Communication Technologies	AT					x
lugano	University of Lugano	CH	x				
nijgmenen	Radboud University Nijgmenen, Information Foraging Lab	NL	x	x	x		
spinque	Spinque	NL	x				
tuwien-1	Vienna University of Technology, Inst. for Computer-Aided Automation	AT					x
tuwien-2	Vienna University of Technology, Inst. for Software Technology and Interactive Systems	AT	x				x
wisenut	WISEnut Inc.	KR	x	x	x		
xerox-sas	Xerox Research Centre Europe	FR				x	x
	Total:		30	16	9	10	12

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

[4] The Direct system is currently developed under the Promise project

obtain the results, cross–language techniques involved, as well as any other relevant details.

⋆ The **hildesheim** and **chemnitz** groups collaborated in try to identify how patent IR is affected by the query length and the use of linguistics in the retrieval process. Their method was developed on top of the Xtrieval framework, which provides a common interface to several search engines. In preprocessing, they used a specific stopword list, especially created for the patent domain. Subsequently, each language was indexed separately. In addition to the text, the Ipc classes were also added to the index. Like other groups this year, they extracted different types of phrases in order to improve the precision of the results. However, their method was not purely statistical, but also used a rule based dependency parser.

The search process considered three types of queries: term-based, phrases-based and a combination of the two. They also compared very long queries with short (and precise) queries. Their results would indicate that using long queries is better than using small precise ones. It is arguable however, how indicative of the content of the given patent application a short query can be. Furthermore, they show that using linguistic phrases did not increase the effectiveness of the retrieval system.

⋆ At the time of writing this text, we have no information on how the experiments submitted by the **hprussia** participant were obtained.

⋆ The team from **Hyderabad** combines three methods in order to retrieve and rank prior art. At the base, their method relies on the Lemur toolkit and the translation of queries into English. First, they apply a key phrase extraction method in order to create queries from the topic patent documents. Then, they identify references to other patens within the text of the document and use them in two ways: first, to create a document vector based on the Ipc codes assigned to the current patent application and to all the referenced patents. Second, to add them directly to the result list.

Without using this citation list, the best results appear to be those using both the Ipc information and the text search. When using citations, the Ipcinformation seems to reduce the quality of the results.

⋆ The **joanneum** group participated in the Img–Cls task. They use features such as Local Binary Patterns (LBP), MPEG-7 features as well as OCR'd text extracted from the images. Support Vector Machines are used for the classification. Runs either use a single type of feature, or combine the classification results of different types of features using late fusion. The best run is the one using the LBP only.

⋆ The **lugano** and **tuwien-2** teams took up on the Pac task by focusing on generating patent summaries that improve the query formulation. Topic patent documents are summarized by the PatTextTiling technique, which is an adaptation of the TextTiling summarization algorithm. The topics' Ipc codes are used to define a relevance set (corpus documents that share Ipc codes with the topic) which is used to get closer to a better relevance model. Query models can be built summary-based or description-based (from the description document sections). After removing stop-words and stemming, the patent document corpus was indexed using Terrier. The ranking model used is BM25. The experiments are filtered by the topic's Ipc code and patent citations extracted from the topic's text are added to the results. Only English topics were considered, for this reason the plot values for the German and French languages are missing in Figure 3.

⋆ In its approach to tackle the Patent Classification tasks, the **nijgmenen** team use the Linguistic Classification System (Lcs) to implement three classifiers: Naive Bayes, Winnow, and SVM^{lihgt}. Various experimental settings were compared to gradually improve classification results. Among such settings are the use of different document sections, of patent citations, different document representations, and different training data sets. Parameter tuning during training also contributed to better classification results. In the training phase only English abstracts and descriptions were used. As metadata the Ipc codes, and the applicant, inventor, and address fields were extracted. It turned out that applicant, inventor and address information did not contribute much to the classification results. To the abstracts and (first 400 words of) the description thus extracted the Aegir dependency parser was applied, its results being added to the document representations. Patent citations in the topic files were used to re–rank the classification results.

For the Prior Art Candidates search task, the **nijgmenen** group teamed up with the **spinque** group, focusing on using bag-of-words approach enriched with syntactic–semantic information. Only the English content of the titles, abstracts, claims and descriptions was used. (For this reason, Figure 7 doesn't have plot entries for the **nijgmenen** experiments.) In a separate processing step, the Ipc information and the first 400 words of the description were extracted. The extracted English content was cleaned up by removing image references and claim headers, sentenced, and parsed with Aegir. Both the corpus and the topics were processed in this way. Selecting the query terms was done based on their relevance to Ipc classes, relevance computed with the (Lcs) software. Finally, the retrieval was done using with the Spinque framework, which allows the definition of search strategies via a graphical user interface.

⋆ The **wisenut** group participated both in the Prior Art Candidate search and Classification tasks, although the Pac participation was due to implementing a KNN-like classification using Pac search results. This solution was chosen based on the experience that training and classifying documents suffer from having few training documents, not enough memory for the (usually) large classification model, and from using too much processing time. To obtain the Pac search results, the Xml files were processed in the following way: weighted keywords were selected out of the title, abstract, description, and claims field. Then, after Pos tagging, functional words, stop words and non–content words are removed and co–occurrence terms are added. Terms and text content in German and

French were translated into English using the MyMemory translation service[5]. The system used to index, search and classify is based on Lucene, to which a simple Java-based interface was implemented. The interface included also an application for classification that applied a improved weighting scheme in the KNN classification to obtain the Cls1 and Cls2 results.

⋆ The **xerox-sas** group participated in both the Img–Pac and Img–Cls tasks. Images for both tasks are represented using Fisher vectors. For the Img–Cls task, linear classifiers are used. The best results are obtained for the run in which the images were artificially rotated and added to the training set, to take into account that the images are sometimes rotated. For the Img–Pac task, different strategies are investigated to compare one set of images to another (as patents consisting of a group of images, not single images, are to be retrieved). Runs are also created in which the predicted Img–Cls image classes are used. For the retrieval of the patent text, different sections of the patent are weighted differently. Similarities are also calculated based on shared Ipc categories and similarities of the patent citation graph, with late fusion used to combine the similarities. A weighted late fusion strategy was again used to combine the text and image ranking, with the text rankings weighted higher than the image rankings. While the visual retrieval performs poorly, when combined with text retrieval it outperforms the text-only retrieval.

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. This included replacing whitespace sequences with only one blank space, filtering out experiment entries which did not belong to a given topic patent documents.

For each submitted Pac experiment we computed the following measures:

- Precision, Precision@5, Precision@10, Precision@20, Precision@50, Precision@100
- Recall, Recall@5, Recall@10, Recall@20, Recall@50, Recall@100
- Map
- Ndcg

For each submitted Cls experiment we computed the following measures:

- Precision@1, Precision@5
- Recall@1, Recall@5
- F_1 at 1 and 5.

All computations were done using the `trec_eval` 9.0 software provided by Nist. Figures 1 through 8 show some of the calculated measures. Detailed values for each of the mentioned measures were sent to the lab participants and are soon to be published into a technical report.

[5] MyMemory, http://mymemory.translated.net

The figures below use a shortened version of the original run names, whose length would have made the pictures less legible. The mapping between the original run name and the shortened versions is shown in the appendix. Note that, although run names might be the same, the experiment files are different between tasks.   For each submitted Img–Cls experiment, we computed Equal Error Rate (EER) and Area Under Curve (AUC) of a ROC curve, and True Positive Rate (TPR) per class averaged over all classes, as well as confusion matrices. This was done using a custom-written script running in Octave[6]. The results of all runs are summarized in Table 2.

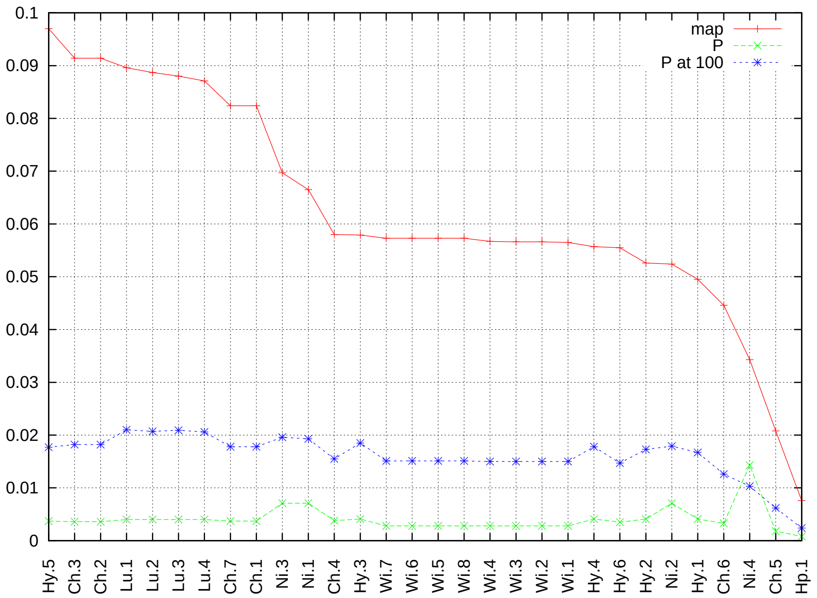

Fig. 1. Map, P measures for the Pac runs

4 Final Observations

We have given an account of the benchmarking activities done in the frame of the Clef–Ip 2011. Compared to the last year, collaborations between research groups has intensified. Another positive observation is that participants are drawing on the research results obtain in the previous years to improve their retrieval methods. The coagulation of research groups leads to a consolidation of the methods used for patent retrieval and allows them to reach a maturity

[6] http://www.gnu.org/software/octave/

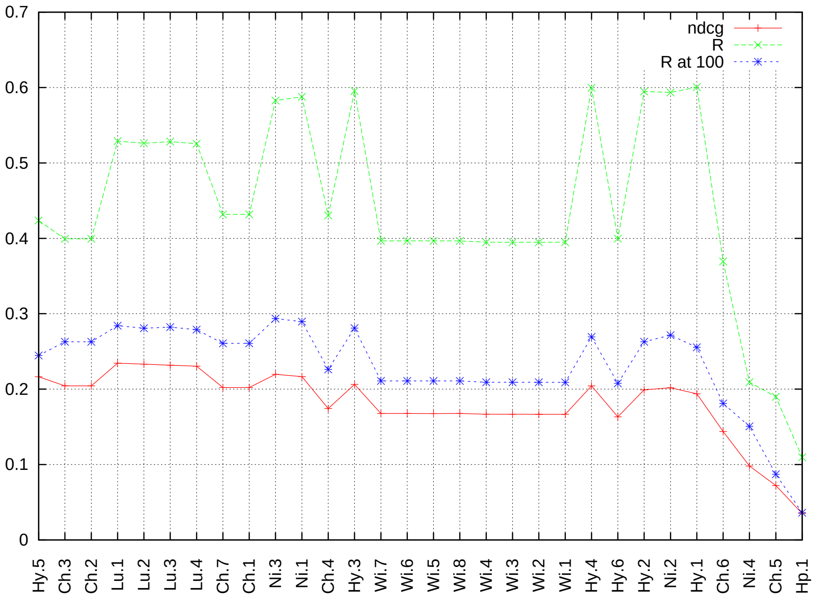

Fig. 2. NDCG R measures for the PAC runs

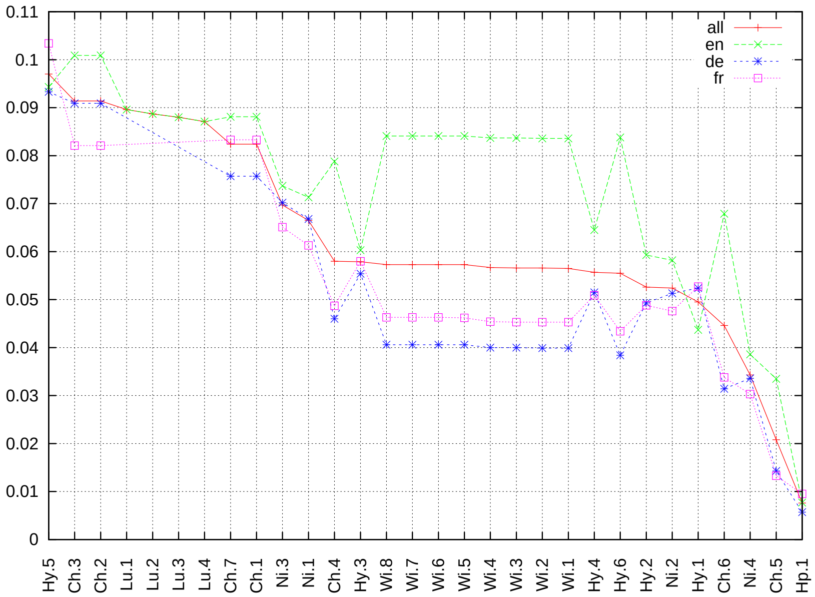

Fig. 3. MAP measures per languages for the PAC runs

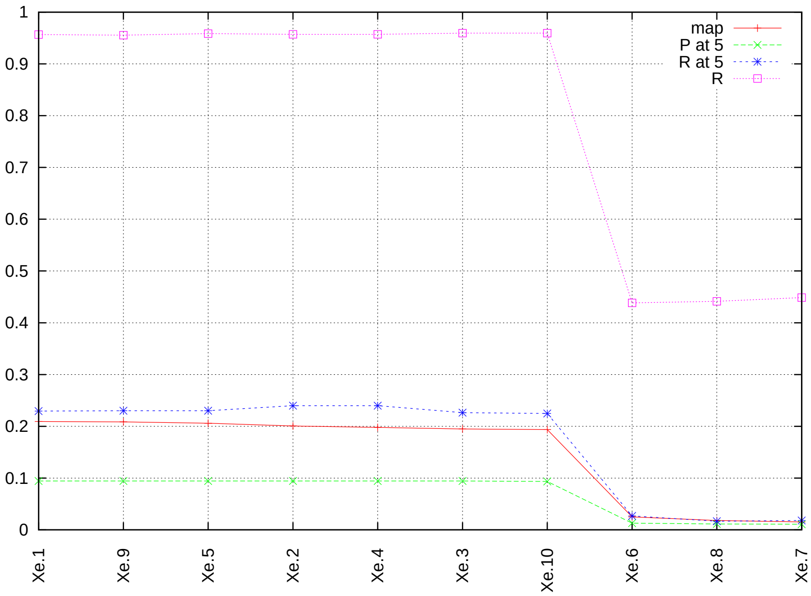

Fig. 4. Measures for the Img–Pac runs

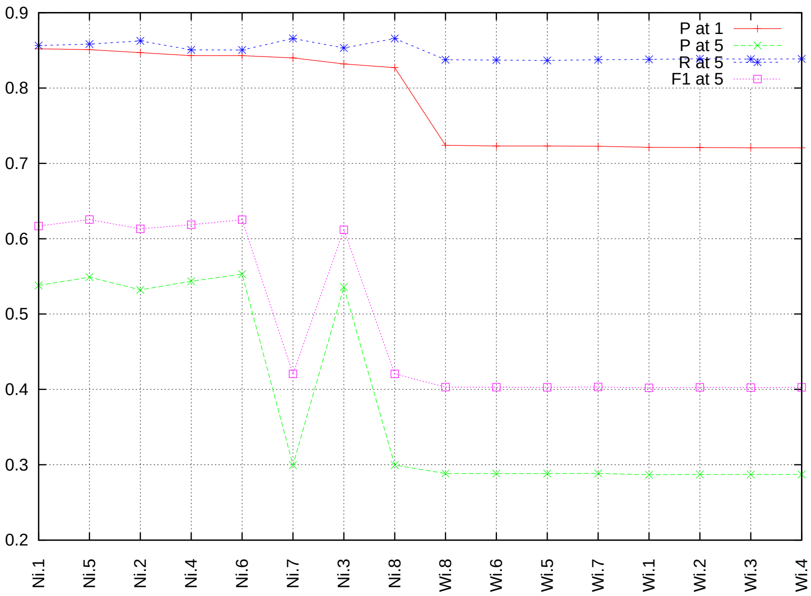

Fig. 5. Measures for the Cls1 runs

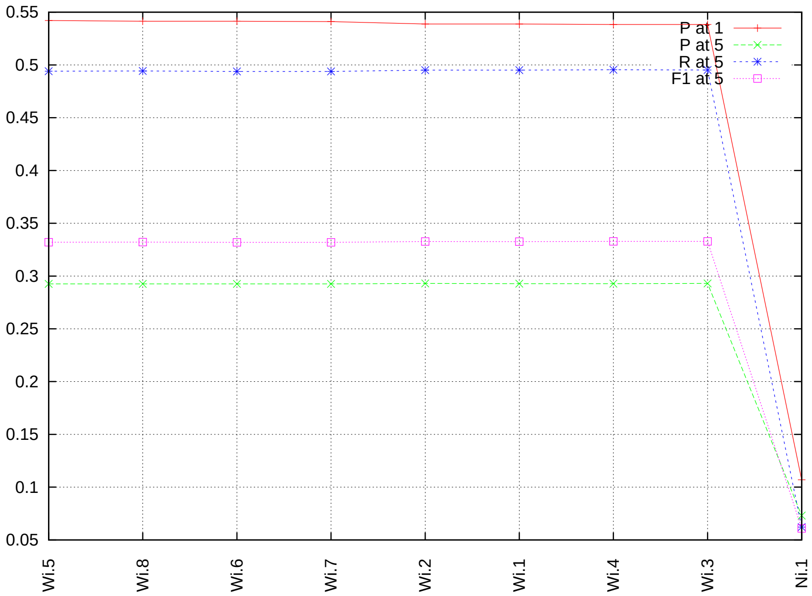

Fig. 6. Measures for the Cls2 runs

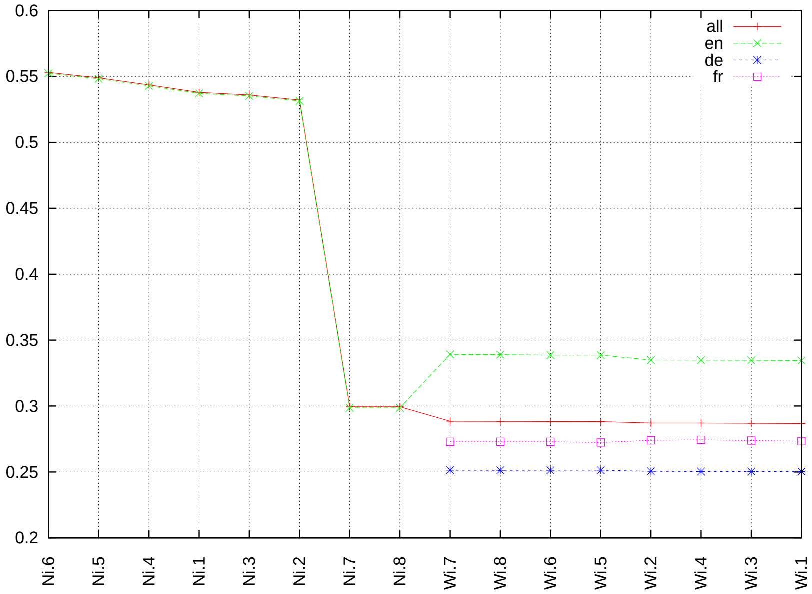

Fig. 7. Measures by language for the Cls1 runs

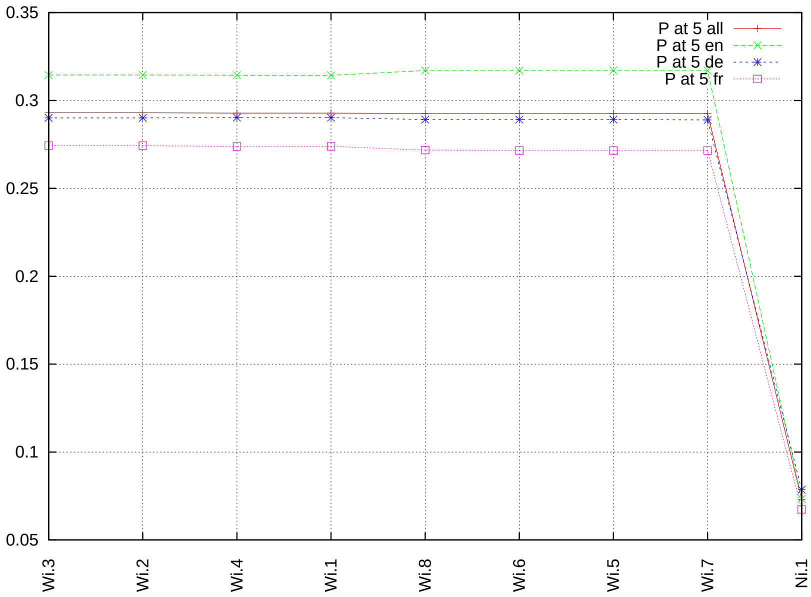

Fig. 8. Measures by language for the Cls2 runs

run	EER	AUC	TPR
joanneum.alphacentauri	0.15	0.91	0.66
joanneum.arcturus	0.24	0.81	0.50
joanneum.betelgeuse	0.18	0.90	0.62
joanneum.canopus	0.16	0.91	0.65
joanneum.procyon	0.37	0.67	0.27
joanneum.rigel	0.16	0.90	0.63
joanneum.sirius	0.16	0.91	0.64
joanneum.vega	0.32	0.72	0.28
xerox-sas.RUNORH	0.06	0.98	0.85
xerox-sas.RUNORH_ROTRAIN	0.04	0.99	0.91
xerox-sas.FV_ORH_SP	0.08	0.92	0.85
xerox-sas.MEAN_ALL	0.08	0.91	0.85

Table 2. Summary of the Img–Cls run results.

level which would make them a candidate for commercial exploitation. However, it also reduces the diversity of the submissions. The number of participating groups was lower this year and at the workshop we need to explore the ways in which the evaluation of patent retrieval tools needs to go ahead.

One such was was identified as of last year. Image retrieval is extremely important for many technologies patented. However, participation in the image tasks was low in the first. The Img–Pac pilot task is challenging due to the multimodal nature of the retrieval task, the large amount of data and the full patent retrieval (containing a set of images) as opposed to single image retrieval. It is hoped that this task can lead to more collaboration between image and text retrieval groups in the next years. The Img–Cls task was planned so as to have a lower threshold of entry for groups with image classification expertise. While 6 groups registered to obtain the data, only 2 participated.

Acknowledgements This work was partly supported by the EU Network of Excellence PROMISE[7] (FP7-258191) and the Austrian Research Promotion Agency (FFG) FIT-IT project IMPEx[8] (No. 825846).

References

1. International Patent Classification (IPC). http://www.wipo.int/classifications/ipc/en/. last retrieved: August, 2011.
2. Patent Cooperation Treaty (PCT).
3. Marco Dussin and Nicola Ferro. DIRECT: Applying the DIKW Hierarchy to Large-Scale Evaluation Campaigns. *Bulletin of IEEE Technical Committee on Digital Libraries (IEEE-TCDL)*, 5(1), 2009.
4. 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.
5. M. Lupu, J. Huang, J. Zhu, and J. Tait. TREC-CHEM: large scale chemical information retrieval evaluation at TREC. *SIGIR Forum*, 43(2), December 2009.
6. F. Piroi and J. Tait. CLEF-IP 2010: Retrieval experiments in the intellectual property domain. Technical Report IRF–TR–2010–00005, Information Retrieval Facility, Vienna, September 2010. Also available as a Notebook Paper of the CLEF 2010 Informal Proceedings.
7. 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] http://www.promise-noe.eu

[8] http://www.joanneum.at/?id=3922

Appendix

Original id	Short id
NIJMEGEN.RUN_ADMWCIT	Ni.1
NIJMEGEN.RUN_ADMWOWCIT	Ni.2
NIJMEGEN.RUN_ADMWOW	Ni.3
NIJMEGEN.RUN_ADMW	Ni.4
NIJMEGEN.RUN_ADWOWCIT	Ni.5
NIJMEGEN.RUN_ADWOW	Ni.6
NIJMEGEN.RUN_ADWTCIT	Ni.7
NIJMEGEN.RUN_ADWT	Ni.8
WISENUT.WISENUT_R1_BASE	Wi.1
WISENUT.WISENUT_R2_BASE_10	Wi.2
WISENUT.WISENUT_R3_BASE_20	Wi.3
WISENUT.WISENUT_R4_BASE_30	Wi.4
WISENUT.WISENUT_R5_CO	Wi.5
WISENUT.WISENUT_R6_CO_10	Wi.6
WISENUT.WISENUT_R7_CO_20	Wi.7
WISENUT.WISENUT_R8_CO_30	Wi.8

Table 3. Original and short run names for the Cls1 task

Original id	Short id
NIJMEGEN.RUN_WINNOW_WORDS	Ni.1
WISENUT.WISENUT_R1_BASE	Wi.1
WISENUT.WISENUT_R2_BASE_10	Wi.2
WISENUT.WISENUT_R3_BASE_20	Wi.3
WISENUT.WISENUT_R4_BASE_30	Wi.4
WISENUT.WISENUT_R5_CO	Wi.5
WISENUT.WISENUT_R6_CO_10	Wi.6
WISENUT.WISENUT_R7_CO_20	Wi.7
WISENUT.WISENUT_R8_CO_30	Wi.8

Table 4. Original and short run names for the Cls2 task

Original id	Short id
XEROX-SAS.3MAX3MEAN	Xe.1
XEROX-SAS.3MAX3MEAN_LATEMONO	Xe.2
XEROX-SAS.3MAX3MEAN_MT_CIT	Xe.3
XEROX-SAS.3MAX_LATEMONO	Xe.4
XEROX-SAS.3MAX_MT	Xe.5
XEROX-SAS.FVORH_3MAX	Xe.6
XEROX-SAS.FVORH_3MAX3MEAN	Xe.7
XEROX-SAS.MAXMEANMODAD	Xe.8
XEROX-SAS.MAXMEANMODAD_MT	Xe.9
XEROX-SAS.MAX_MT_CIT	Xe.10

Table 5. Original and short run names for the Img–Pac task

Original id	Short id
CHEMNITZ.CUT_UHI_CLEFIP_BOW	Ch.1
CHEMNITZ.CUT_UHI_CLEFIP_BOW_DESC	Ch.2
CHEMNITZ.CUT_UHI_CLEFIP_BOW_DESC_IPCR	Ch.3
CHEMNITZ.CUT_UHI_CLEFIP_BOW_EN_ABSTRACT	Ch.4
CHEMNITZ.CUT_UHI_CLEFIP_BOW_EN_P	Ch.5
CHEMNITZ.CUT_UHI_CLEFIP_BOW_EN_P_ABSTRACT	Ch.6
CHEMNITZ.CUT_UHI_CLEFIP_BOW_IPCR	Ch.7
HPRUSSIA.1	Hp.1
HYDERABAD.CATVECTORSIMILARITY	Hy.1
HYDERABAD.CATVECTORTEXT CITATIONALL	Hy.2
HYDERABAD.CATVECTORTEXT RETREIVAL	Hy.3
HYDERABAD.CATVECTORTFIDF TEXTCITATIONALL	Hy.4
HYDERABAD.TEXTRANK_IDF CITATIONALL	Hy.5
HYDERABAD.TEXTRANK_IDF_20	Hy.6
LUGANO.CIT_NOTEXTIL_LLQM	Lu.1
LUGANO.CIT_PATTEXTIL_LLQM	Lu.2
LUGANO.NOTEXTIL_LLQM	Lu.3
LUGANO.PATTEXTIL_LLQM	Lu.4
NIJMEGEN.RUN_COMBINED1	Ni.1
NIJMEGEN.RUN_COMBINED2	Ni.2
NIJMEGEN.RUN_KEYWORDS	Ni.3
NIJMEGEN.RUN_TRIPLES	Ni.4
WISENUT.WISENUT_R1_BASE	Wi.1
WISENUT.WISENUT_R2_BASE_10	Wi.2
WISENUT.WISENUT_R3_BASE_30	Wi.3
WISENUT.WISENUT_R4_BASE_30	Wi.4
WISENUT.WISENUT_R5_CO	Wi.5
WISENUT.WISENUT_R6_CO_10	Wi.6
WISENUT.WISENUT_R7_CO_20	Wi.7
WISENUT.WISENUT_R8_CO_30	Wi.8

Table 6. Original and short run names for the Pac task