CLEF-IP 2009: retrieval experiments in the Intellectual Property domain

Giovanna Roda[1], John Tait[2], Florina Piroi[2], Veronika Zenz[1]
[1]Matrixware Information Services GmbH
[2]The Information Retrieval Facility (IRF)
Vienna, Austria
{g.roda,v.zenz}@matrixware.com
{j.tait,f.piroi}@ir-facility.org

Abstract

The Clef–Ip track ran for the first time within Clef 2009. 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 track focused on the task of prior art search. The 15 European teams who participated in the track deployed a rich range of Information Retrieval techniques adapting them to this new specific domain and task. A large-scale test collection for evaluation purposes was created by exploiting patent citations.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.4 Systems and Software - *Performance evaluation*

General Terms

Algorithms, Experimentation, Measurement

Keywords

Patent retrieval, Prior art search, Intellectual Property, Test collection, Evaluation track, Benchmarking

1 Introduction

The Cross Language Evaluation Forum Clef[1] originally arose from a work on Cross Lingual Information Retrieval in the US Federal National Institute of Standards and Technology Text Retrieval Conference Trec[2] but has been run separately since 2000. Each year since then a number of tasks on both cross–lingual information retrieval (Clir) and monolingual information retrieval in non–English languages have been run. In 2008 the Information Retrieval Facility (Irf) and Matrixware Information Services GmbH obtained the agreement to run a track which allowed groups to assess their systems on a large collection of patent documents containing a mixture of English, French and German documents derived from European Patent Office data. This became

[1] http://www.clef-campaign.org
[2] http://trec.nist.gov

known as the Clef–Ip track, which investigates IR techniques in the Intellectual Property domain of patents.

One main requirement for a patent to be granted is that the invention it describes should be novel: that is there should be no earlier patent or other publication describing the invention. The novelty breaking document can be published anywhere in any language. Hence when a person undertakes a search, for example to determine whether an idea is potentially patentable, or to try to prove a patent should not have been granted (a so-called opposition search), the search is inherently cross–lingual, especially if it is exhaustive.

The patent system allows inventors a monopoly on the use of their invention for a fixed period of time in return for public disclosure of the invention. Furthermore, the patent system is a major underpinning of the company value in a number of industries, which makes patent retrieval an important economic activity.

Although there is important previous academic research work on patent retrieval (see for example the Acm Sigir 2000 Workshop [9] or more recently the Ntcir workshop series [5], there was little work involving non–English European Languages and participation by European groups was low. Clef–Ip grew out of desire to promote such European research work and also to encourage academic use of a large clean collection of patents being made available to researchers by Matrixware (through the Information Retrieval Facility).

Clef–Ip has been a major success. For the first time a large number of European groups (15) have been working on a patent corpus of significant size within an integrated and single IR evaluation collection. Although it would be unreasonable to pretend the work is beyond criticism it does represent a significant step forward for both IR community and patent searchers.

2 The CLEF-IP Patent Test Collection

2.1 Document Collection

The Clef–Ip track had at its disposal a collection of patent documents published between 1978 and 2006 at the European Patent Office (Epo). The whole collection consists of approximately 1.6 million individual patents. As suggested in [6], we split the available 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)

Patents published prior to 1985 were excluded from the outset, as before this year many documents were not filed in electronic form and the optical character recognition software that was used to digitize the documents produced noisy data. The upper limit, 2006, was induced by our data provider—a commercial institution—which, at the time the track was agreed on, had not made more recent documents available.

The documents are provided in Xml format and correspond to the Alexandria Xml Dtd[3]. Patent documents are structured documents consisting of four major sections: bibliographic data, abstract, description and claims. Non-linguistic parts of patents like technical drawings, tables of formulas were left out which put the focus of this years track on the (multi)lingual aspect of patent retrieval: Epo patents are written in one of the three official languages English, German and French. 69% of the documents in the Clef–Ip collection have English as their main language, 23% German and 7% French. The claims of a granted patent are available in all 3 languages and also other sections, especially the title are given in several languages. That means the document collection itself is multilingual, with the different text sections being labeled with a language code.

[3] http://www.ir-facility.org/pdf/clef/patent-document.dtd

Patent documents and kind codes

In general, to one patent are associated several patent documents published at different stages of the patent's life–cycle. Each document is marked with a *kind code* that specifies the stage it was published in. The kind code is denoted by a letter possibly followed by a one–digit numerical code that gives additional information on the nature of the document. In the case of the Epo, "A" stands for a patent's application stage and "B" for a patent's granted stage, "B1" denotes a patent specification and "B2" a later, amended version of the patent specification[4].

Characteristic to our patent document collection is that files corresponding to patent documents published at various stages need not contain the whole data pertinent to a patent. For example, a "B1" document of a patent granted by the Epo contains, among other, the title, the description, and the claims in three languages (English, German, French), but it usually does not contain an abstract, while an "A2" document contains the original patent application (in one language) but no citation information except the one provided by the applicant.[5]

The Clef–Ip collection was delivered to the participants "as is", without joining the documents related to the same patent into one document. Since the objective of a search are patents (identified by patent numbers, without kind code), it is up to the participants to collate multiple retrieved documents for a single patent into one result.

2.2 Tasks and Topics

The goal of the Clef–Ip tasks consisted in finding prior art for a patent. The tasks mimic an important real–life scenario of an IP search professional. Performed at various stages of the patent life-cycle, prior art search is one of the most common search types and a critical activity in the patent domain. Before applying for a patent, inventors perform a such a search to determine whether the invention fulfills the requirement of novelty and to formulate the claims as to not conflict with existing prior art. During the application procedure, a prior art search is executed by patent examiners at the respective patent office, in order to determine the patentability of an application by uncovering relevant material published prior to the filing date of the application. Finally parties that try to oppose a granted patent use this kind of search to unveil prior art that invalidates patents claims of originality.

For detailed information on information sources in patents and patent searching see [3] and [8].

Tasks

Participants were provided with sets of patents from the topic pool and asked to return all patents in the collection which constituted prior art for the given topic patents. Participants could choose among different topic sets of sizes ranging from 500 to 10000.

The general goal in Clef–Ip was to find prior art for a given topic patent. We proposed one main task and three optional language subtasks. For the language subtasks a different topic representation was adopted that allowed to focus on the impact of the language used for query formulation.

The main task of the track did not restrict the language used for retrieving documents. Participants were allowed to exploit the multilinguality of the patent topics. The three optional subtasks were dedicated to cross–lingual search. 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 is 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, in the respective language, extracted from the same "B1" patent document. Participants were presented the same patents as in the main task, but

[4] For a complete list of kind codes used by various patent offices see http://tinyurl.com/EPO-kindcodes

[5] It is not in the scope of this paper to discuss the origins of the content in the Epo patent documents. We only note that applications to the Epo may originate from patents granted by other patent offices, in which case the Epo may publish patent documents with incomplete content, referring to the original patent.

with textual parts (title, claims) only in one language. The usage of bibliographic data, e.g. IPC classes was allowed.

Topic representation

In CLEF–IP a topic is itself a patent. Since patents come in several version corresponding to the different stages of the patent's life-cycle, we were faced with the problem of how to best represent a patent topic.

A patent examiner initiates a prior art search with a full patent application, hence one could think about taking highest version of the patent application's file would be best for simulating a real search task. However such a choice would have led to a large number of topics with missing fields. For instance, for EuroPCTs patents (currently about 70% of EP 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).

In order to overcome these shortcomings of the data, we decided to assemble a virtual "patent application file" to be used as a topic by starting from the "B1" document. If the abstract was missing in the B1 document we added it from the most current document where the abstract was included. Finally we removed citation information from the bibliographical content of the patent document.

Topic selection

Since relevance assessments were generated by exploiting existing manually created information (see section 3.1) CLEF–IP had a topic pool of hundreds of thousands of patents at hand. Evaluation platforms usually strive to evaluate against large numbers of topics, as robustness and reliability of the evaluation results increase with the number of topics [15] [16]. This is especially true when relevance judgments are not complete and the number of relevant documents per topic is very small as is the case in CLEF–IP where each topic has on average only 6 relevant documents. In order to maximize the number of topics while still allowing also groups with less computational resources to participate, four different topic bundles were assembled that differed in the number of topics. For each task participants could chose between the topics set S (500 topics), M (1,000 topics), L (5,000 topics), and XL (10,000 topics) with the smaller sets being subsets of the larger ones. Participants were asked to submit results for the largest of the 4 sets they were able to process.

From the initial pool of 500, 000 potential topics, candidate topics were selected according to the following criteria:

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. Starting from a topics pool of approximately 500,000 patents, we were left with almost 16,000 patents fulfilling the above requirements. From these patents, we randomly selected 10,000 topics, which bundled in four subsets constitute the final topic sets. In the same manner 500 topics were chosen which together with relevance assessments were provided to the participants as training set.

For an in-depth discussion of topic selection for CLEF–IP see [13].

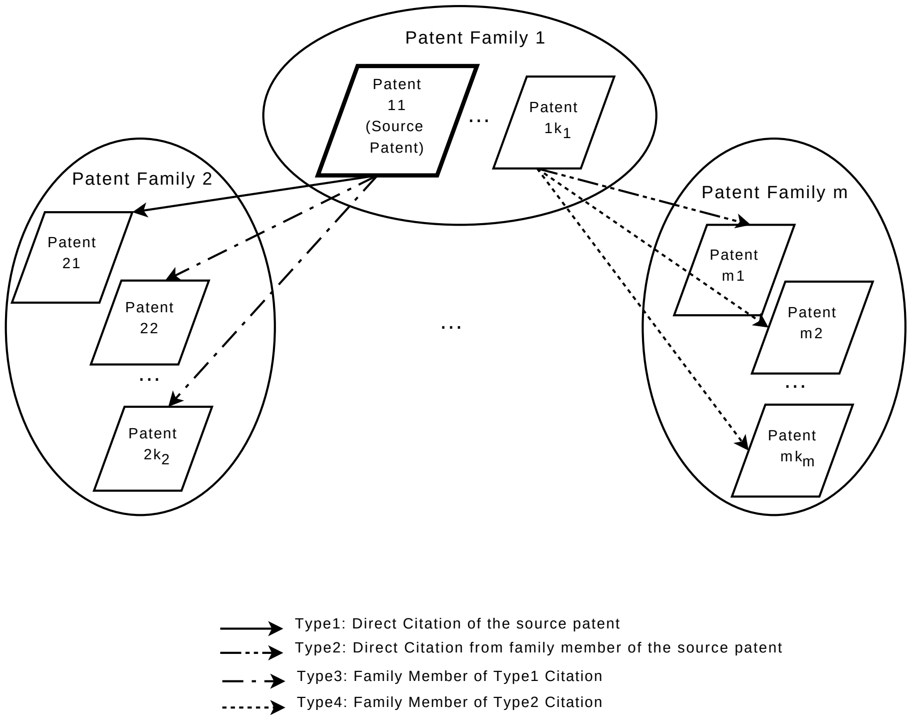
Patent Family 1
Patent
11
(Source
Patent)
...
Patent
$1k_1$
Patent Family 2
Patent
21
Patent
22
...
Patent
$2k_2$
...
Patent Family m
Patent
m1
Patent
m2
...
Patent
mk_m
Type1: Direct Citation of the source patent
Type2: Direct Citation from family member of the source patent
Type3: Family Member of Type1 Citation
Type4: Family Member of Type2 Citation

Figure 1: Patent citation extension used in CLEF-IP09

3 Relevance Assessment Methodology

This section describes the two types of relevance assessments used in Clef–Ip2009: (1) assessments automatically extracted from patent citations as well as (2) manual assessments by patent experts.

3.1 Automatic Relevance Assessment

A common challenge in IR evaluation is the creation of ground truth data against which to evaluate retrieval systems. The common procedure of pooling and manual assessment is very labor-intensive. Voluntary assessors are difficult to find, especially when expert knowledge is required as is the case of the patent field. Researchers in the field of patents and prior art search however are in the lucky position of already having partial ground truth at hand: patent citations.

Citations are extracted from several sources:

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).

There are two major advantages of extracting ground truth from citations. First citations are established by members of the patent offices, applicants and patent attorneys, in short by highly qualified people. Second, search reports are publicly available and are made for any patent

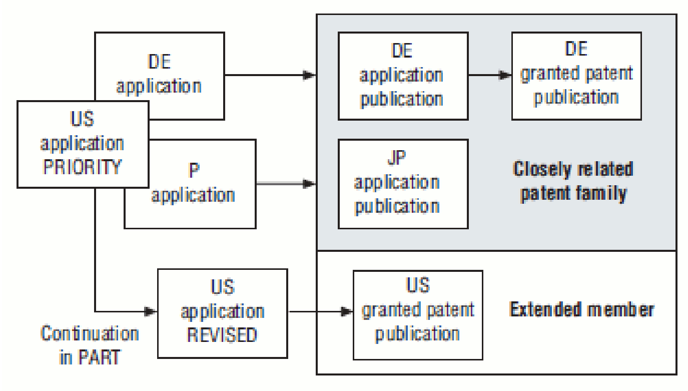
DE application
US application PRIORITY
P application
US application REVISED
Continuation in PART
DE application publication
DE granted patent publication
JP application publication
Closely related patent family
US granted patent publication
Extended member

Figure 2: Example for close and extended patent families. Source: OECD ([10])

application, which leads to a huge set of assessment material that allows the track organizers to scale the set of topics easily and automatically.

Methodology

The general method for generating relevance assessments from patent citations is described in [6]. This idea had already been exploited at the NTCIR workshop series[6]. Further discussions within the 1st IRF Symposium in 2007 [7] led to a clearer formalization of the method.

For CLEF–IP 2009 we used an extended list of citations that includes not only patents cited directly by the patent topic, but also patents cited by patent family members and family members of cited patents. By means of patent families we were able to increase the number of citations by a factor of seven. Figure 1 illustrates the process of gathering direct and extended citations.

A *patent family* consists of patents granted by different patent authorities but related to the same invention (one also says that all patents in a family share the same *priority* data). For CLEF–IP this close (also called *simple*) patent family definition was applied, as opposed to the extended patent family definition which also includes patents related via a split of one patent application into two or more patents. Figure 1 (from [10]) illustrates an example of extended families.

In the process of gathering citations, patents from ∼ 70 different patent offices (including USPTO, SIPO, JPO, etc.) were considered. Out of the resulting lists of citations all non–EPO patents were discarded as they were not present in the target data set and thus not relevant to our track.

Characteristics of patent citations as relevance judgments

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] prescribe that if the search results in several documents of equal relevance, the search report should normally contain no more than one of

[6] http://research.nii.ac.jp/ntcir/
[7] http://www.ir-facility.org/symposium/irf-symposium-2007/the-working-groups

them. This means that we have incomplete recall bases which must be taken into account when interpreting the evaluation results presented here.

Further automatic methods

To conclude this section we describe further possibilities of extending the set of relevance judgements. These sources have not been used in the current evaluation procedure as they seem to be less reliable indicators of relevancy. Nevertheless they are interesting avenues to consider in the future, which is why they are mentioned here:

A list of citations can be expanded by looking at patents cited in cited patents, if we assume some level of transitivity of this relation. It is however arguable how relevant a patent C is to patent A if we have something like *A cites B* and *B cites C*. Moreover, such a judgment cannot be done automatically.

In addition, 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.

Recently, a new approach for extracting prior art items from citations has been presented in [14].

3.2 Manual Relevance Assessment by Patent Experts

A number of patent experts were contacted for the manual assessment of a small part of the track's experimental results. Communicating the project's goals and procedures was not an easy task, nor was it motivating them to invest some time for this assessment activity. Nevertheless, a total of 7 experts agreed to assess the relevance of retrieved patents for one or more topics. Topics were chosen by the experts out of our collection according to their area of expertise. A limit of around 200 retrieved patents to assess seemed to provide an acceptable amount of work. This limit allowed us to pool experimental data up to depth 20.

The engagement of patent experts resulted in 12 topics assessed up to rank 20 for all runs. A total of 3140 retrieval results were assessed with an average of 264 results per topic.

The results were submitted too late to be included in the track's evaluation report. In the section on evaluation activities we are going to report on the results obtained by using this additional small set of data for evaluation even though this collection is too small a sample to draw any general conclusions.

4 Submissions

4.1 Submission format

For all tasks, a submission consisted of a single ASCII text file containing at most 1,000 lines per topic, in the standard format used for most TREC submissions: white space is used to separate columns, the width of the columns is not important, but it is important to have exactly five columns per line with at least one space between the columns.

```
EP1133908    Q0    EP1107664    1    3020
EP1133908    Q0    EP0826302    2    3019
EP1133908    Q0    EP0383071    3    2995
```

where:

- the first column is the topic number (a patent number);

Id	Institution		Tasks	Sizes	Runs
Tud	*Tech. Univ. Darmstadt, Dept. of CS, Ubiquitous Knowledge Processing Lab*	DE	Main, En, De, Fr	s(4), m(4), l(4), xl(4)	16
UniNe	*Univ. Neuchatel - Computer Science*	CH	Main	s(7), xl(1)	8
uscom	*Santiago de Compostela Univ. - Dept. Electronica y Computacion*	ES	Main	s(8)	8
Utasics	*University of Tampere - Info Studies & Interactive Media and Swedish Institute of Computer Science*	FI SE	Main	xl(8)	8
clefip-ug	*Glasgow Univ. - IR Group Keith*	UK	Main	m(4), xl(1)	5
clefip-unige	*Geneva Univ. - Centre Universitaire d'Informatique*	CH	Main	xl(5)	5
cwi	*Centrum Wiskunde & Informatica - Interactive Information Access*	NL	Main	m(1), xl(4)	4
hcuge	*Geneva Univ. Hospitals - Service of Medical Informatics*	CH	Main, En, De, Fr	m(3), xl(1)	4
humb	*Humboldt Univ. - Dept. of German Language and Linguistics*	DE	Main, En, De, Fr	xl(4)	4
clefip-dcu	*Dublin City Univ. - School of Computing*	IR	Main	xl(3)	4
clefip-run	*Radboud Univ. Nijmegen - Centre for Language Studies & Speech Technologies*	NL	Main, En	s(2)	1
Hildesheim	*Hildesheim Univ. - Information Systems & Machine Learning Lab*	DE	Main	s(1)	1
Nlel	*Technical Univ. Valencia - Natural Language Engineering*	ES	Main	s(1)	1
Uaic	*Al. I. Cuza University of Iasi - Natural Language Processing*	RO	EN	s(1)	1

Table 1: List of active participants and runs submitted

- the second column is the query number within that topic. This is currently unused and should always be Q0;
- the third column is the official document number of the retrieved document;
- the fourth column is the rank of the document retrieved;
- the fifth column shows the score (integer or floating point) that generated the ranking. This score must be in decreasing order.

4.2 Submitted runs

A total of 70 experiments from 14 different teams and 15 participating institutions (the University of Tampere and Sics joined forces) was submitted to Clef–Ip 2009. Table 1 contains a list of all submitted runs.

Experiments ranged over all proposed tasks (one main task and three language tasks) and over three (S, M, XL) of the proposed task sizes.

Submission System

Clear and detailed guidelines together with automated format checks are critical in managing large-scale experimentations.

Group-ID	MT	qterm selection	indexes	ranking model
cwi	-	tf-idf	?	boolean, bm25
clefip-dcu	-	none	one english only	
hcuge	x	none	?	bm25
Hildesheim	-	none	one german only	?
humb	-	?	one per language, one additional phrase index for english, crosslingual concept index	kl, bm25
Nlel	-	random walks	mixed language passage index	passage similarity
clefip-run	-	none	one english only	tf-idf
Tud	-	none	one per language, one for Ipc	tf-idf
Uaic	-	none	one mixed language index (split in 4 indexes for performance reasons)	
clefip-ug	-	tf-idf	one mixed language	bm25, cosine retrieval model
clefip-unige	-	?	one engilsh only	tf-idf, bm25, fast
UniNE	-	tf-idf	one mixed language index	tf-idf, bm25, dfr
uscom	-	tf-idf	one mixed language index	bm25
Utasics	x	ratf, tf-idf	1 per language, 1 for Ipc	?

Table 2: Index, Query formulation

For the upload and verification of runs a track management system was developed based on the open source document management system Alfresco[8] and the web interface Docasu[9]. The system provides an easy-to-use Web-frontend that allows participants to upload and download runs and any other type of file (e.g. descriptions of the runs). The system offers version control as well as a number of syntactical correctness tests. The validation process that is triggered on submission of a run returns a detailed description of the problematic content. This is added as an annotation to the run and is displayed in the user interface. Most format errors were therefore detected automatically and corrected by the participants themselves. Still one error type passed the validation and made the postprocessing of some runs necessary: patents listed as relevant on several different ranks for the same topic patent. Such duplicate entries were filtered out by us before evaluation.

4.3 Description of Submitted Runs

A comparison of the retrieval systems used in the Clef–Ip Task is given in Table 2. The usage of Machine Translation (MT) is displayed in the second column, showing that MT was applied only by two groups, both using Google Translate. Methods used for selecting query terms are listed in the third column. As Clef–Ip topics are whole patent documents many participants found it necessary to apply some kind of term selection in order to limit the number of terms in the query. Methods for term selection based on term weighting are shown here while preselection based on patent-fields is shown separately in Table 3. Given that each patent document could contain fields in up to three languages many participants chose to build separate indexes per language, while others just generated one mixed-language index or used text fields only in

[8] http://www.alfresco.com/

[9] http://docasu.sourceforge.net/

		fields used in index				fields used in query				
Group-ID	Ipc	title	claims	abs	desc	title	claims	abs	desc	Other
cwi	x	?	?	?	?	?	?	?	?	?
clefip-dcu	x	x	x	x	x	x	x	x	x	-
hcuge	x	x	-	x	x	x	x	x	x	citations
Hildesheim	-	x	x	-	-	x	x	-	-	-
humb	x	x	x	x	x	x	x	x	x	citations, priority, applicant, ecla
Nlel	-	x	-	-	x	x	-	-	x	-
clefip-run	-	-	x	-	-	-	x	-	-	-
Tud	x	-	x	x	x	x	x	-	-	-
Uaic	-	x	x	x	x	x	x	x	x	-
clefip-ug	x	x	x	x	x	x	x	x	x	*
clefip-unige	x	x	x	x	-	x	x	x	x	applicant, inventor
UniNE	x	x	x	x	x	x	x	x	x	-
uscom	-	?	?	?	?	x	x	x	x	-
Utasics	x	x	x	x	x	x	x	x	x	-

Table 3: Fields used in indexing and query formulation

one languages discarding information given in the other languages. The granularity of the index varied too, as some participants chose to concatenate all text fields into one index fields, while others indexed different fields separately. In addition several special indexes like phrase or passage indexes, concept indexes and Ipc indexes were used. A summary on which indexes were built and which ranking models were applied is given in Table 2.

Table 3 gives an overview over the patent fields used in query formulation and indexing. The text fields title, claims, abstract and description were used most often. Among the bibliographic fields Ipc was the field exploited most, it was used either as post-processing filter or as part of the query. Only two groups used the citation information that was present in the document set. Other very patent-specific information like priority, applicant, inventor information was only rarely used.

Some further remarks on the runs that were submitted to the track:

- As this was the first year for Clef–Ip many participants were absorbed with understanding the data and task and getting the system running. The Clef–Ip track presented several major challenges
  - A new retrieval domain (patents) and task (prior art).
  - The large size of the collection.
  - The special language used in patents. Participants had not only to deal with German, English and French text but also with the specialities of patent-specific language ("Patentese").
  - The large size of topics. In most Clef tracks a topic consists of few selected query words while for Clef–Ip a topic consists of a whole patent. The prior art task might thus also be tackled from the viewpoint of a document similarity or as proposed by **Nlel** as a plagiarism detection task.
- Cross-linguality: participants approached the multilingual nature of the Clef–Ip document collection in different ways: Some groups like **clefip-ug** or **Uaic** did not focus on the multilingual nature of the data. Other participants like **Hildesheim** and **clefip-dcu** chose to use only

data in one specific language while many others used several monolingual retrieval systems to retrieve relevant documents and merged their results. Two groups made use of machine translation: **Utasics** used Google translate in the Main task to make patent-fields available in all three languages. They report that using the Google translation engine actually deteriorated their results. **hcuge** used Google translate to generate the fields in the missing languages in the monolingual tasks. **humb** applied cross-lingual concept tagging.

- Several teams integrated patent-specific know-how in their retrieval systems by:
  - Using classification information (Ipc, Ecla) was mostly found helpful. Several participants used the Ipc class in their query formulation as a post-ranking filter criterium. While using Ipc classes to filter out generally improves the retrieval results, it also makes it impossible to retrieve relevant patents that don't share an Ipc class with the topic.
  - **hcuge** and **humb** exploited citation information given in the corpus.
  - Apart from patent classification information and citations further bibliographic data (e.g. inventor, applicant, priority information) was used only by **humb**.
  - Only few groups had patent expertise at the beginning of the track. Aware of this problem some groups started cooperation with patent experts, like for example **Utasics** who are currently analysing patent experts' query formulation strategies.
- Even though query and indexing time were not evaluation criteria, participants had to start thinking about performance due to the large amount of data.
- Different strategies were applied for indexing/ranking on patent level. Several teams applied the concept of virtual patent documents introduced by the organizers in the presentation of topics for indexing a set of patent documents as a single entity.
- Some teams combined several different strategies in their systems: this was done on a large scale by the **humb** team. **cwi** proposes a graphical user interface for combining search strategies.
- The training set, consisting of 500 patents with relevance assessments, was used by almost all of the participants, mostly for tuning and checking their strategies. **humb** used the training set also for Machine Learning. For this aim, it showed to be too small and they generated a larger one from the the citations available in the corpus.
- Having made the evaluation data available allowed many participants (among them **Tud**, **Utasics**, **Hildesheim**, **clefip-ug**) to run additional experiments after the official evaluation. They report on new insights obtained (e.g. further tuning and comparisons of approaches) in their working notes papers.

5 Results

We evaluated the experiments by some of the most commonly used metrics for IR effectiveness evaluation. A correlation analysis shows that the rankings of the systems obtained with different topic sizes can be considered equivalent. The manual assessments obtained from patent experts allowed us to perform some preliminary analysis on the completeness of the automatically generated set of relevance assessments.

The complete collection of measured values for all evaluation bundles is provided in the Clef–Ip 2009 Evaluation Summary ([11]). Detailed tables for the manually assessed patents will be provided in a separate report ([12]).

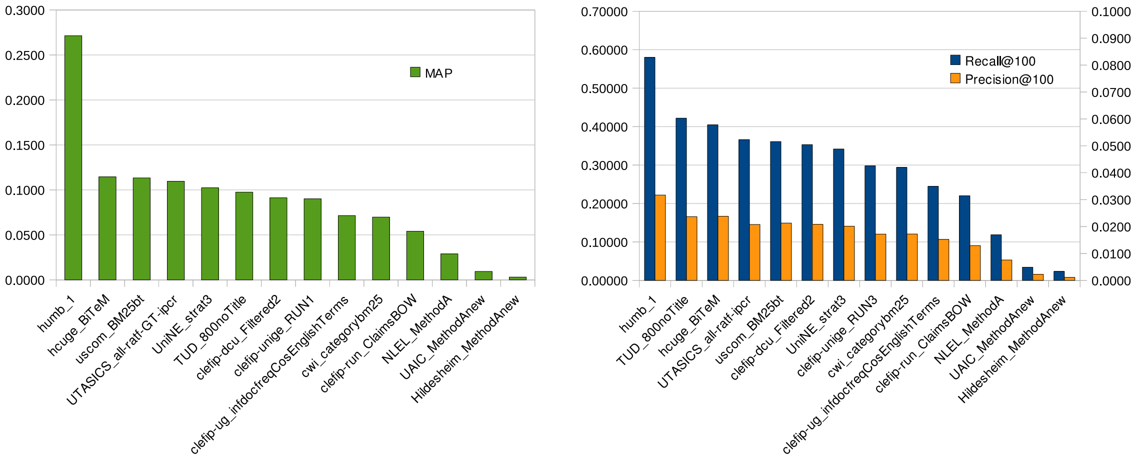

Figure 3: MAP, Precision@100 and Recall@100 of best run/participant (S)

Group-ID	Run-ID	MAP	Recall@100	Precision@100
humb	1	0.2714	0.57996	0.0317
hcuge	BiTeM	0.1145	0.40479	0.0238
uscom	BM25bt	0.1133	0.36100	0.0213
UTASICS	all-ratf-ipcr	0.1096	0.36626	0.0208
UniNE	strat3	0.1024	0.34182	0.0201
TUD	800noTitle	0.0975	0.42202	0.0237
clefip-dcu	Filtered2	0.0913	0.35309	0.0208
clefip-unige	RUN3	0.0900	0.29790	0.0172
clefip-ug	infdocfreqCosEnglishTerms	0.0715	0.24470	0.0152
cwi	categorybm25	0.0697	0.29386	0.0172
clefip-run	ClaimsBOW	0.0540	0.22015	0.0129
NLEL	MethodA	0.0289	0.11866	0.0076
UAIC	MethodAnew	0.0094	0.03420	0.0023
Hildesheim	MethodAnew	0.0031	0.02340	0.0011

Table 4: MAP, Precision@100, Recall@100 of best run/participant (S)

5.1 Measurements

After some corrections of data formats, we created experiment bundles based on size and task. For each experiment we computed 10 standard IR measures:

- Precision, Precision@5, Precision@10, Precision@100
- Recall, Recall@5, Recall@10, Recall@100
- MAP
- nDCG (with reduction factor given by a logarithm in base 10)

All computations were done with SOIRE[10], a software for IR evaluation based on a service–oriented architecture. Results were double–checked against `trec_eval`[11], the standard program for evaluation used in the TREC evaluation campaign, except for nDCG for which, at the time of the evaluation, we were not aware of a publicly available implementation.

[10] http://soire.matrixware.com
[11] http://trec.nist.gov/trec_eval

MAP, recall@100 and precision@100 of the best run for each participant are listed in Table 4 and illustrated in Figure 3. The values were calculated on the small topic set. The MAP values range from 0.0031 to 0.27 and are quite low in comparison with other CLEF tracks. The Precision values are generally low, but it must be noted that the average topic had 6 relevant documents, meaning that the upper boundary for precision@100 was at 0.06. Recall@100, a highly important measure in prior art search, ranges from 0.02 to 0.57. It must be noted these low values might be due to the incompleteness of the automatically generated set of relevance assessments.

5.1.1 Correlation analysis

In order to see whether the evaluations obtained with the three different bundle sizes (S, M, XL) could be considered equivalent we did a correlation analysis comparing the vectors of MAPs computed for each of the bundles.

In addition to that, we also evaluated the results obtained by the track's participants for the 12 patents that were manually assessed by patent experts. We evaluated the runs from three bundles extracting only the 12 patents (when present) from each runfile. We called these three extra-small evaluation bundles and named them ManS, ManM, ManXL. Table 5 lists Kendall's τ and Spearman's ρ for all compared rankings.

Figures 4 5 illustrates the correlation between pairs of bundles together with the best least-squares linear fit.

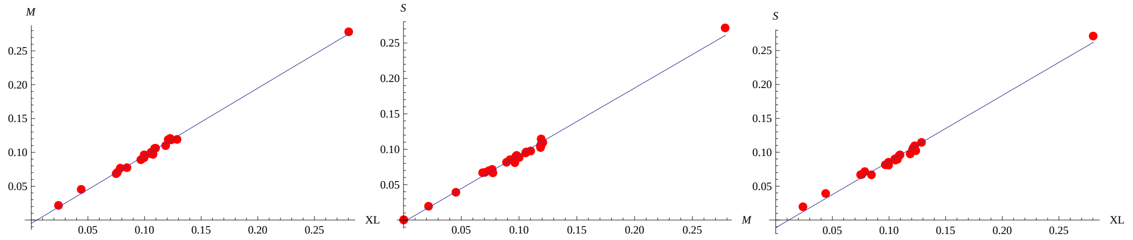

Figure 4: Correlation of rankings by MAP: comparison of XL, M, S bundles

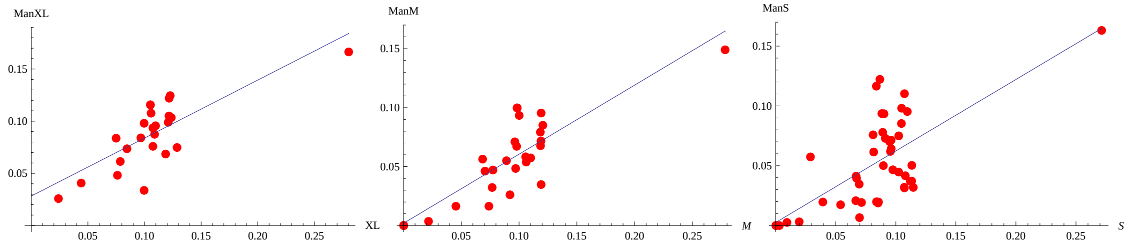

Figure 5: Correlation of rankings by MAP: bundles XL, M, S versus manual relevance assessments (≤ 12 topics)

The rankings obtained with topic sets S, M, and L are highly correlated, suggesting that the three bundles an be considered equivalent for evaluation purposes. As expected, the correlation between S, M, XL and the respective ManS, ManM, ManXL rankings by MAP drops drastically.

It must however be noted that the limited number of patents in the manual assessment bundle (12) is not sufficient for drawing any conclusion. We hope to be able to collect more data in the future in order to assess the quality of our automatically generated test collection.

Correlation	#runs	τ	ρ
M vs XL	24	0.9203	0.9977
S vs M	29	0.9160	0.9970
S vs XL	24	0.9058	0.9947
XL vs ManXL	24	0.5	0.7760
M vs ManM	29	0.6228	0.8622
S vs ManS	48	0.4066	0.7031

Table 5: Correlations of systems rankings for MAP

5.1.2 Some remarks on the manually assessed patents

Patent experts marked in average 8 of the proposed patents as relevant to the seed patent. For a comparison:

- 5.4 is the average number of citations for the 12 seed patents that were assessed manually
- for the whole collection, there are in average 6 citations per patent

Furthermore, some of the automatically extracted citations (13 out of 34) were marked as not relevant by patent experts. Again, in order to have some meaningful results a larger set of data is needed.

6 Lessons Learned and Plans for 2010

In the 2009 collection only patent documents with data in French, English and German were included. One area in which to extend the track for 2010 is provide additional patent data in more European languages.

Patents are organized in what are known as “patent families”. A patent might be originally filed in France in French, and then subsequently to ease enforcement of that patent in the United States a related patent might be filed in English with the US Patents and Trademarks Office. Although the full text of the patent will not be a direct translation of the French (for example because of different formulaic legal wordings) the two documents may be comparable, in the sense of a Comparable Corpus in Machine Translation). It might be that such comparable data will be useful to participants to mine for technical and other terms. The 2009 collection does not lend itself to this use and we will seek to make the collection more suitable for that purpose.

For the first year we measured the overall effectiveness of systems. A more realistic evaluation should be layered in order to measure the contribution of each single component to the overall effectiveness results as proposed in the GRID@CLEF track ([4]) and also by [7]. Analysis of the data should be statistical.

The 2009 task was also somewhat unrealistic in terms of a model of the work of patent professionals. Real patent searching often involves many cycles of query reformulation and results review, rather than one off queries and results set. In 2010 we would like to move to a more realistic model.

7 Epilogue

CLEF IP has to be regarded as a major success: looking at previous CLEF tracks we regarded four to six groups as a satisfactory first year participation rate. Fifteen is a very satisfactory number of participants - a tribute to those who did the work and to the timeliness of the task and data. In terms of retrieval effectiveness the results have proved hard to evaluate: if there is an over all conclusion the effective combination of a wide range of indexing methods is best, rather than a single silver bullet or wooden cross. However some of the results from groups other than

Humboldt University indicate that specific techniques may work well: we look forward to more results next year. Also it is unclear how well the 2009 task and methodology maps to what makes a good (or better) system from the point of view of patent searchers - this is an area where we clearly need to improve. Finally we need to be clear that a degree of caution is needed for what is inevitably an initial analysis of a very complex set of results.

Acknowledgements

We would like to thank Judy Hickey, Henk Tomas and all the other patent experts who helped us with manual assessments and who shared their know-how on prior art searches with us. Thanks to Evangelos Kanoulas and Emine Yilmaz for interesting discussions on creating large test collections.

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] Stephen R Adams. *Information sources in patents*. K.G. Saur, 2006.

[4] N. Ferro and D. Harman. Dealing with multilingual information access: Grid experiments at trebleclef. In Esposito F. In Agosti, M. and C. Thanos, editors, *Post-proceedings of the Fourth Italian Research Conference on Digital Library Systems (IRCDL 2008)*, 2008.

[5] 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.

[6] 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.

[7] A. Hanbury and H. Müller. Toward automated component-level evaluation. In *SIGIR Workshop on the Future of IR Evaluation, Boston, USA*, pages pages 29–30., 2009.

[8] David Hunt, Long Nguyen, and Matthew Rodgers. *Patent searching : tools and techniques*. Wiley, 2007.

[9] Noriko Kando and Mun-Kew Leong. Workshop on Patent Retrieval (SIGIR 2000 Workshop Report). *SIGIR Forum*, 34(1):28–30, 2000.

[10] Organisation for Economic Co-operation and Development (OECD). *OECD Patent Statistics Manual*, Feb. 2009.

[11] Florina Piroi, Giovanna Roda, and Veronika Zenz. CLEF-IP 2009 Evaluation Summary. July 2009.

[12] Florina Piroi, Giovanna Roda, and Veronika Zenz. CLEF-IP 2009 Evaluation Summary part II (in preparation). September 2009.

[13] Giovanna Roda, Veronika Zenz, Mihai Lupu, Kalervo Järvelin, Mark Sanderson, and Christa Womser-Hacker. So Many Topics, So Little Time. *SIGIR Forum*, 43(1):16–21, 2009.

[14] Shahzad Tiwana and Ellis Horowitz. Findcite – automatically finding prior art patents. In *PaIR '09: Proceeding of the 1st ACM workshop on Patent information retrieval*. ACM, to appear.

[15] Ellen M. Voorhees. Topic set size redux. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 806–807, New York, NY, USA, 2009. ACM.

[16] Ellen M. Voorhees and Chris Buckley. The effect of topic set size on retrieval experiment error. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 316–323, New York, NY, USA, 2002. ACM.