# UAIC: Participation in CLEF-IP Track

Adrian Iftene, Ovidiu Ionescu, George-Răzvan Oancea, Andrei-Dumitru Balmoş

UAIC: Faculty of Computer Science, "Alexandru Ioan Cuza" University, Romania
{adiftene, ovidiu.ionescu, george.oancea, andrei.balmos}@info.uaic.ro

**Abstract.** The CLEF-IP track was launched in 2009 to investigate IR techniques for patent retrieval. It is part of the CLEF 2009 evaluation campaign. Also, in 2009 we built a system in order to participate in the CLEF-IP competition. Our system has three main components: filtering module, indexing module, and searching module. Because the process of indexing of all 75 G of input documents with patents took too much time we decided to work in a peer-to-peer environment with four computers. In this paper we will see how we managed to built this system.

## 1 Introduction

The CLEF-IP[1] (Intellectual Property) was a new track in CLEF 2009. Like corpora in IP track was utilized a collection of more than 1M patent documents mainly derived from EPO sources. The collection covered English, French and German languages with at least 100,000 documents in each language.

There were two kinds of tasks in the track:

- The main task was to find patent documents that constitute *prior art* to a given patent.
- *Three facultative subtasks* that use parallel monolingual queries in English, German, and French. The goal of these subtasks is to evaluate the impact of language on retrieval effectiveness.

Queries and relevance judgments were produced by two methods. The first used queries produced by Intellectual Property Experts and reviewed by them in a fairly conventional way. The second was an automatic method using patent citations from seed patents. Search results reviewed ensured that the majority of test and training queries produce results in more than one language. The first results reported retrieving across all three languages.

The track was coordinated by: Information Retrieval Facility & Matrixware (AT).

The way in which we built the system for CLEF-IP track is presented in Section 2, while Section 3 presents the run submitted. Last Section presents conclusions regarding our participation in CLEF-IP 2009.

[1] CLEF-IP track: www.ir-facility.org/the_irf/current-projects/clef-ip09-track/

## 2 UAIC System

Our system has three main modules: module one responsible with extracting of relevant fields from XML files, module two that indexes the relevant fields, and the third module that does the searching. The Figure 1 presents the system architecture.

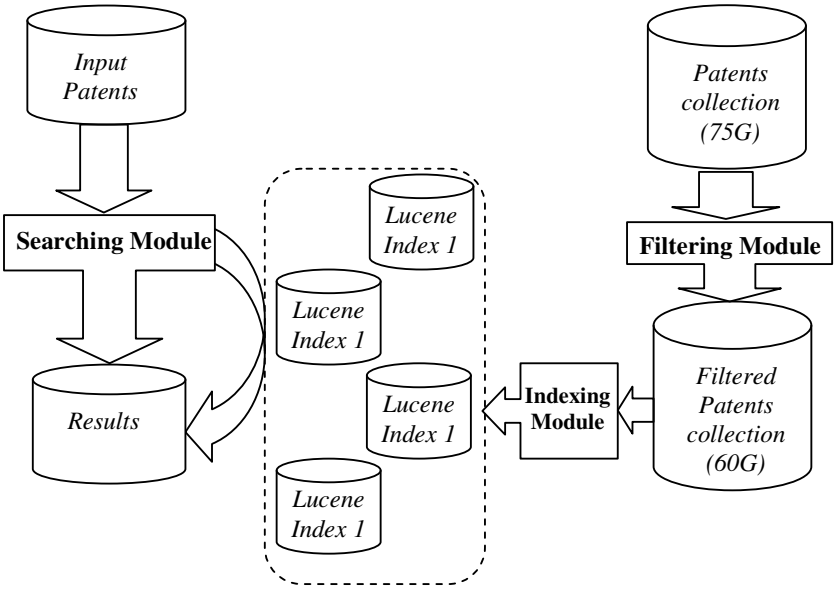


**Figure 1**: UAIC system used in INFILE 2009

In what follows we will see details about these modules.

### 2.1 Extracting of Relevant Fields

The aim of this step is to reduce the amount of XML files that must be indexed and to work only with relevant tags from XML files. *Why*? Our initial tests demonstrate that in order to index ~150Mb we need around 94 seconds. Removing the irrelevant fields the time need for indexing was reduced to 82 seconds.

We agreed on the tags we will keep (<*invention-title*>, <*claim text*> and <*abstract*> tags. If <*abstract*> tag would not be found we will search and keep the <*description*> tag). These fields were determined after a review of initial pre-processed documents.

### 2.2 Index Creation

In order to index the corpora we use Lucene (Hatcher and Gospodnetic, 2005) a suite of free libraries used for indexing and searching in large corpora. For each XML document filtered at previous step, we add useful fields to the Index. To do this we use Lucene Indexer class adapted for our purpose.

Because the process of indexing of all 75 G of documents took too much (for our initial indexing we need around 80 minutes) we decided to work in a peer-to-peer environment. Thus, we split the initial corpora on four machines and we create separate indexes on each of them. In this way the time necessary for indexing was reduced to 20 minutes.

### 2.3 Searching Component

This component allows performing searches in index created at previous step. Starting from patent files with size of 1 M, we extract the same tags used in indexing part and built a Lucene query in order to search in Lucene index.

When we create the Lucene query from input patent we use different boost factors for used tags. Thus, we have the following cases:

i. The greater boost value is used when we find one tag from patent in the same corresponding field from Lucene index. For example, this is the case when we search words from *invention-title* tag of current patent in the *invention-title* field of Lucene index.
ii. The boost values depend by current tag name. Thus, boost values are in descending order starting from *invention-title*, *claim text*, *abstract* and *description* tags.
iii. The lower boost values are used when we have cross-searches between tag from patent and field from Lucene index. For example, if we search words from *invention-title* tag from patent in *abstract* field from Lucene index.

## 3 Submitted Run

Fourteen groups submitted 70 runs at this track. We submitted one run. Details from official evaluation are presented below:

**Table 1**: Official results for UAIC run

| Run ID | P | R | MAP | nDCG |
|---|---|---|---|---|
| UAIC_MethodA | 0.0004 | 0.0670 | 0.0094 | 0.1877 |

## 4 Conclusions

The CLEF-IP was a new track in CLEF 2009, which utilized a collection of more than 1M patent documents from English, French and German languages with at least 100,000 documents in each language.

The UAIC system, which took part in the CLEF-IP 2009 competition, has three main components: filtering module, indexing module, and searching module. The *filtering module* has the aim to reduce the amount of XML files that must be indexed and to work only with relevant tags from XML files. The *indexing module* used Lucene and because the process of indexing of all 75 G of documents took too much we worked in a peer-to-peer environment. The *searching module* component allows performing searches in index created at previous step. When we create the Lucene query from input patent we use different boost factors for used tags.

## Acknowledgements

We want to give a "thank you" to those who helped from the beginning of the project: students from second year group 5A.

## References

1. Hatcher, E. and Gospodnetic, O.: Lucene in action. *Manning Publications Co*. (2005)