Multilingual natural language access to patents

The ePatent project

Bernard Normier
Lingway, 33-35 rue Ledru Rollin 94200 Ivry s/ Seine France
bernard.normier@lingway.com

Abstract

The ePatent main result will be a European wide Internet portal to access information on patents offering multilingual services (English, French, German and Spanish) to non professional users.
Simplification brought by natural language queries, multilingual processing and Internet access will dramatically reduce the cost of accessing IPR information. This will offer a strong market opportunity at European level and will impact the competitively of European organisations looking for IPR protection.

General framework

1.1 The consortium

The ePatent project is conducted in the framework of the eContent program of the EC. It is conducted by a consortium of four national patent institutes:

- INPI in France
- OEPM in Spain
- UKPO in UK
- OEPA in Austria,
-

and two privates companies

- JOUVE as the system integrator
- LINGWAY as the linguistic technology provider

This consortium is open and could include new participants in the future. The project started on january 2002 and will be completed in 2004. Intermediate results will be available and used by the end of 2002.

1.2 Objectives

The ePatent main result will be a European wide Internet portal to access information on patents offering multilingual services (F, UK, D, E) to non professional users.

Simplification brought by natural language queries, multilingual processing and Internet access will dramatically reduce the cost of accessing IPR information. This will offer a strong market opportunity at European level and will impact the competitively of European organisations looking for IPR protection.

The project will aim at providing a European wide cross lingual repository of patent information based on the International Patent Classification
developing a multilingual natural language interface and search engine to patent information databases in 4 languages, in intelligent ranking facilities and reading support
developing a corresponding Internet based service to distribute and exploit patent information at European level

1.3 Rationale

Everyday life shows us the challenges facing industrial property with examples such as the patentability of living organisms and software, patented drugs and generic drugs, piracy lawsuits, etc. Protection of industrial property is consequently a fundamental asset for any organisation. However, identifying or applying for a patent (or to a less extend utility model, industrial design or trademark which are slightly simpler to identify) is a challenging and complex task in the European context where

- Access to information is not easy if you are not a professional information specialist or not able to translate patent material not written in your own language
- Understanding a patent in your non native language requires translation

The difficulty of accessing patent information also impacts their economical exploitation: hard to find means reduced opportunities to licence them, useless re-inventions and risks to be sued.
On-line access to patent databases (national patents, European patents and international applications) and simplified procedures must be offered to European organisations, and in particular SMEs, to guarantee their competitivity through patenting or exploitation of available patents at reasonable cost.
From a technological perspective, there are currently several opportunities that would match this demand:

- The deployment of Internet which is now widely spread and available for almost all organisations including SMEs. Internet is becoming the major information retrieval tool (at least on a first level) for almost everyone
- Multilingual tools for searching and retrieving information from natural language queries are now effective, especially when underlying ontology is well structured
- Patent descriptions use world-wide a formal representation (International Patent Classification - IPC) which is a language independent classification addressing the just mentioned requirement of natural language processing tools
- Patent Offices have made available their patent database on the Internet and they all offer access to the IPC indexing on the Internet

- European patent organisations agreed to reduce the number of languages needed in the first place to translate patent descriptions at European level

The ePatent project is an answer to address this patenting problem met by European organisations through the exploitation of European patent databases using the Internet and multilingual facilities

Current difficulties in accessing patent information

There are today essentially two type of systems: taxonomy based systems and full-text based systems. These two approaches are not fully satisfactory today.

Taxonomy based systems

World-wide, the IPC is a recognised standard used by all countries although some also use in parallel other systems, but it is a difficult to use professional tool not adapted to casual users, in particular SMEs, for several resaons :

- it is a very large taxonomy
- the vocabulary is very technical, professional, and often obscure
- as every taxonomy, navigation trough it is not easy, as the same object could often be placed at different places, and the right one is not obvious to find.
- Even if it is a large classification plan, there are still places not detailed enough, identifying large number of patents.

.

Full-text based systems

Full-text based systems do not use pre-existing taxonomies, and search directly in text, comparing words in the query with words in the texts. If this approach seems to be easy to implement and to use, it appears to be largely ineffective, for several reasons:

- the same concept can be expressed by a lot of different words. To be efficient, a query has to list all these different words (synonyms, related words, etc.) making the construction of a query particularly difficult.
- Words in the query have generally several meanings, and documents using the wrong meaning are found
- Classical full-text systems do not manage compound words, which are very frequent in patents.

These difficulties using full-text systems generate a lot of silence (the relevant patents are not found) and noise (non-relevant patents are found), which can eventually managed by a professional information scientist, but hardly by a non-professional user.

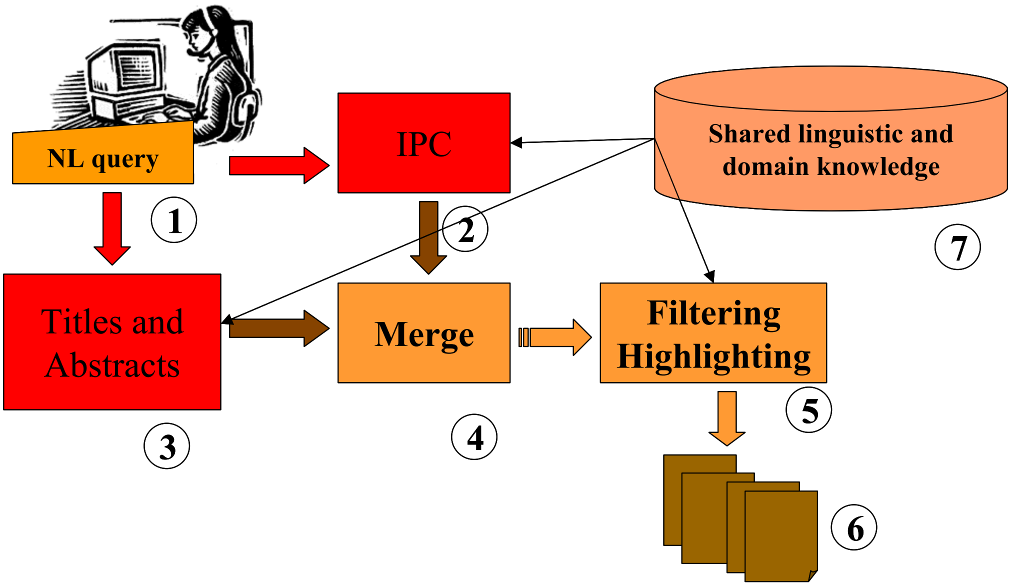
2 ePatent approach

To solve theses issues , the project approach will be as follows:

- Develop a cross lingual retrieval engine exploiting both
  - the International Patent Classification (IPC) formalism as a "pivot" language, as there already exist translations of IPC in most of the European languages
  - filtering techniques applied on the texts of summaries and, when available, on the full text of the patent
- Develop the interfaces and infrastructure to interact with European Patent databases, offering
  - Natural language interface in French, English, German and Spanish (for the project) and other languages afterwards.
  - Intelligent searching facilities to retrieve relevant information
  - Query results delivered in the language of the query and based upon IPC representation and summaries of patent, ranked according to the query analysis. In case a summary is not available in the language of the query, translation facilities will be provided
- Set-up a European distributed patent information service through an European Internet portal that will provide access to distributed patent data

3 EPatent architecture

The following diagram illustrates the ePatent architecture:



1- A natural language query is typed by the user. This query can be in one of the 4 languages currently handled by the project. This query is full natural language, without any vocabulary restrictions.

2- The query in translated in the relevant IPC code. Such a system has already done in France by INPI, based on LINGWAY linguistic technology, and is currently being implement by JOUVE on the PLUTARQUE portal.

3- In parallel, and only in those situations where a underlying boolean search engine in used, the same initial natural language query is translated in a expanded boolean query. This means that the query will be saturated with a lot, hopefully all, possible synonyms and related words.

4- These two parallel processes will find different sets of documents. They will be combined and ranked, acoording to their relative ranking obtained by each way.

5- At that step, a reasonably small number of relevant patents is identified. The system will then use filtering techniques. The text of the patent (or eventually only the abstract) is analysed on-the-flight (which means having very efficient natural language analysis technique), to eliminates residual noise, achieve better ranking and highlight the most relevant parts of the document.

6- Finally, additional linguistic tools as translation or summarization could be applied on the results

7- An important point to notice is that the same linguistic knowledge base is used for all the steps in that process

Conclusion

The ePatent project is a good example of using state of the art technology, in natural language processing and Internet fields, to improve access and usage of existing information, for the benefit of non-professional users. The focus of the project on multilinguality, which is a key issue in Europe, will be reinforced in the future, as the consortium intends to develop the system to other European languages.