



TECHNOLOGY CONSULTING INNOVATION

ELCA

Computer-Assisted Categorization of Patent Documents

How can patents be classified accurately into 69,000 categories? How can this process be supported all over the world in a convenient fashion? Can a computer-assisted solution make a difference to patent classifiers? These are the starting questions that an international organization active in intellectual property wanted to answer.

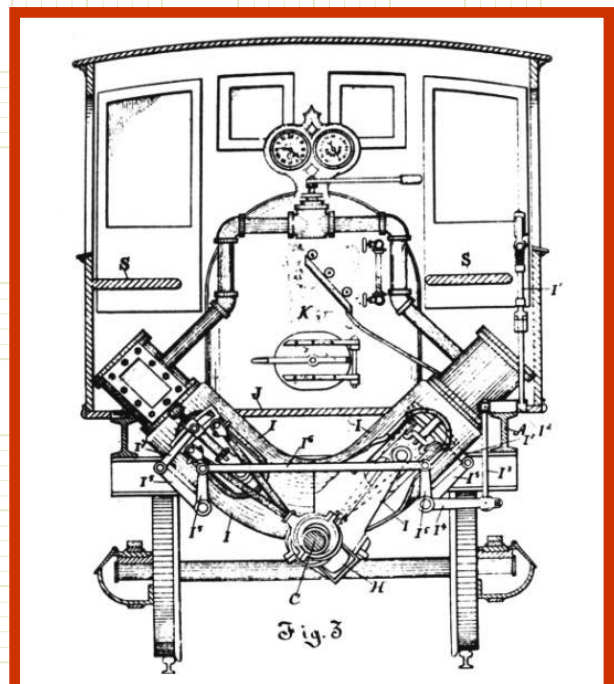
When a patent application is considered or submitted, the search for previous inventions in the same field relies crucially on accurate patent classification. The retrieval of patent documents is vital for patent-issuing authorities, potential inventors, research and development units, and others concerned with the application or development of technology. In order to locate relevant earlier patent documents more easily, and provide extremely focused searches, a standard taxonomy exists for classifying patents and patent applications. This taxonomy is known as the International Patent Classification (IPC), and covers all areas of technology, including large sections for chemistry, mechanics, and electronics.

The Problem

Patent experts in national and regional patent offices currently classify all patent applications manually. To date, about 30 out of over 50 million published patent documents have been classified worldwide, in more than 90 countries. To perform this important task, patent experts need an intimate knowledge of the classification system. The International Patent Classification is a complex hierarchical taxonomy comprising sections, classes, subclasses, and groups. The latest edition of the IPC contains about 120 classes, about 630 subclasses, and approximately 69,000 groups. With such a range of categories, it is difficult for human classifiers to classify patent applications consistently worldwide.

The number of patent applications is currently rising rapidly worldwide, creating the need for computer-assisted categorization systems to help streamline time-consuming and labour-intensive manual categorizations. One of the objectives of the project was to provide the information technology support for this task.

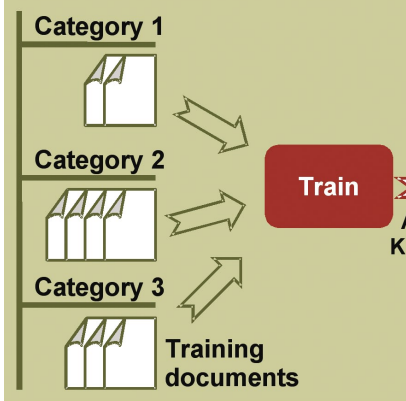
The computer-assisted categorization system should facilitate the attribution of IPC categories to patent applications and provide for accurate searches of patent documents. The project focused particularly



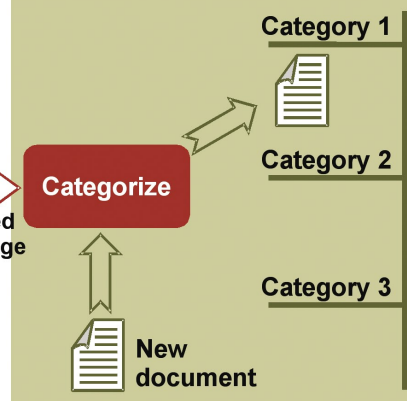
Where should locomotive patents be classified?

on supporting small and medium-sized patent offices in classifying patents for their future easy retrieval. As patent documents are often available in several languages, the project aimed at a system supporting the categorization of patents in several different European languages.

Phase 1: Learning



Phase 2: Sorting



Categorization proceeds in two phases: First the system automatically learns to recognize topics by reading a set of training documents in each category. Then, by powerful generalization, it can be used to sort new documents into the correct categories. Human validation of the predicted categories provides a computer-assisted solution.

Classifying patents is a difficult task to perform manually, for several reasons:

- The classification system contains a large number of cross-references and also has a few overlapping categories,
- There are a number of placement rules, which indicate where documents should be categorized, but which depend strongly of the content or focus of the patents,
- Patent documents can—and in some cases must—be classified in several different categories,
- The vocabulary used in patents is quite unlike that in other documents. Many vague or general terms are often used in order to avoid narrowing the scope of the invention.

The Solution

ELCA provided consulting services with a view to determining the state-of-the-art in patent categorization research and realizations within both academia and industrial environments. Contacts were established with over 30 academic partners and personal visits to the European Patent Office and Japanese Patent Office provided vital experience from similar efforts.

Patent categorization provides a demanding test scenario for computer-assisted categorization because of the nature of the taxonomy and the variety of patent documents. ELCA's task was to

determine the possibilities offered by a variety of commercial and freeware categorization products, evaluate their respective advantages, and test them in real-life demanding situations. Following these activities, a custom-built computer-assisted categorizer was specified to provide high levels of functionality, fast responses, and to integrate well with existing document collections.

Using sophisticated machine-learning techniques, the idea is to train a program to associate patent topics with categories in the taxonomy. Training is performed completely automatically, on the basis of large collections of documents that have already been classified manually. In order to achieve highest accuracy, several hundred thousand patent documents are employed to train the system. By using recent pre-classified documents, the system can learn which topics are associated with which category. In this way, all classification rules are automatically taken into account since the training is based on previous manual classifications by human experts. Training can be performed overnight, and thus updated regularly.

End users are provided with a web interface to upload patent documents and request category predictions from the system. By generalizing from the documents it has been trained on, the system can predict the correct categories, and also evaluate its own confidence in its predictions.

Solution Highlights

The technical implementation of the solution was developed by Metaread (www.metaread.ch), a company that builds on technological innovations from the University of Geneva to design and implement break-through solutions for knowledge-intensive companies. The solution developed has a number of advantages.

- By making predictions at different levels of detail within the taxonomy, the tool provides users with a convenient, computer-assisted, and interactive approach to classifying patents.
- By comparing the accuracy of the categories predicted with those of other academic and commercial document classifiers, we have found that in nearly all cases our customized solution performs with higher accuracy.
- The tool is designed to be completely language-independent, and can therefore just as easily be trained with English-language documents and with Russian patent applications. Adapting the system to categorize documents in a new language is therefore an extremely simple procedure. Extensive tests of the resulting classification accuracy have shown negligible differences between languages.
- Training the system is extremely quick. Several hundred thousand documents can be read and processed overnight.

- The tool is not specific to patent documents, and could be employed for any complex categorization task.

The full patent categorization solution is expected to be available both as a server-based system and as a distributed version on a CD-ROM for patent offices around the world. As the same computer-assisted categorization software will be provided to all patent offices, the patent documents will be classified more consistently in the future.

Technical Details

The system was developed purely in Java, for maximum portability. A web interface accesses a server that runs over a freeware infrastructure. The training module reads XML patent document collections written in any language, and computes a small digested set of weights associated with each category, using a sophisticated neural network algorithm. Each category has a certain number of words and weights associated to it, which are determined automatically by the system during training. The categorization module reads this collection of weights and uses it to predict categories for new documents from the words within them. Since the computationally intensive task is performed during training, categorizing a new document is near instantaneous.

ELCA

ELCA is a leading Swiss supplier in software development, systems integration, and business consulting (www.elca.ch). The company, headquartered in Lausanne, employs over 300 people, with offices in Lausanne, Bern, Zürich, Geneva, Paris, and Ho Chi Minh City. Solutions have been deployed for various companies active in the medical and pharmaceutical fields, as well as in the financial or insurance markets. ELCA positions itself as an implementation partner with a tight project and cost management and a large knowledge of software technology and integration techniques. Partnerships with leading IT suppliers such as Documentum, Microsoft, IBM, and others complete the offer.

As an independent consulting firm and system integrator ELCA has successfully completed a variety of projects in the document management area. It has thereby acquired extensive know-how in the field, as well as practical experience with various leading products in the world market. ELCA is one of the few IT companies in Switzerland that are able to competently and reliably handle complex document management projects with demanding integration requirements.

ELCA, Av. de la Harpe 22-24, 1001 Lausanne, Switzerland
Tel. +41 (0)21 613 21 11, Fax +41 (0)21 613 21 00, info@elca.ch, www.elca.ch