Computer-Assisted Categorization of Patent Documents in the International Patent Classification

C. J. Fall
ELCA, Avenue de la Harpe 22-24, CH-1000 Lausanne 13, Switzerland, caspar.fall@elca.ch, www.elca.ch

K. Benzineb
Metaread SA, Rue Eugène-Marziano 15, CH-1227 Genève-Acacias, Switzerland, kbenzineb@metaread.com, www.metaread.com

J. Guyot
Centre Universitaire d'Informatique, Université de Genève, Rue Général Dufour 24, CH-1211 Genève 4, Switzerland, guyot.inge@free.fr, cui.unige.ch

A. Törcsvári
Arcanum Development, Baranyai utca 10, H-1117 Budapest, Hungary, attila@arcanum.com, www.arcanum.com

P. Fiévet
World Intellectual Property Organization, Chemin des Colombettes 34, CH-1211 Genève 20, Switzerland, patrick.fievet@wipo.int, www.wipo.int

Abstract
The World Intellectual Property Organization is currently developing a system for assisting users in categorizing patent documents in the International Patent Classification (IPC). The system should support the classification of documents in several languages and aims to assist users in locating relevant IPC symbols by providing them with a convenient web-based service. The approach taken for developing such a system relies on powerful machine learning algorithms that are trained on manually classified documents to recognize IPC topics. We detail in-house results of applying a custom-built state-of-the-art computer-assisted categorizer to English, French, Russian, and German-language patent documents. We find that reliable computer-assisted categorization at IPC subclass level is an achievable goal for the statistical methods employed here. A categorization system suggesting three IPC symbols for each document can predict the main IPC class correctly for around 90% of documents, and the main IPC subclass for about 85% of documents. The accuracy of the system at main group level is enhanced if the user first validates the correct IPC class.

Introduction

When a patent application is considered or submitted, the search for previous inventions in the field—known as prior art—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 documents more easily, and provide extremely focused searches, the World Intellectual Property Organization (WIPO) has developed a standard taxonomy for classifying patents and patent applications [1]. The International Patent Classification (IPC) covers all areas of technology, including large sections for chemistry, mechanics, and electronics. WIPO is now aiming to provide computer assistance in both IPC classification and IPC-based searching to facilitate accurate patent search activities.

Domain experts in national and regional patent offices need an intimate knowledge of the IPC and currently classify all patent applications manually. Necessarily, the IPC is a complex, hierarchical taxonomy, in which about 30 out of over 50 million published patent documents have been classified worldwide. The industrial property offices of more than 90 countries currently use the IPC. The increase in adoption of this system enhances the need for consistent classification of patents and support for IPC-based categorization in small and medium-sized patent offices with limited resources. The number of patent applications is currently rising rapidly worldwide, creating the need for computer-assisted categorization systems to help streamline time-consuming and labor-intensive manual categorizations [2-4].

In industry, patents are a major source for gathering intelligence about competitors' activities, but this source necessitates sophisticated tools for meaningful data mining [5]. In the chemical field, search services such as the Chemical Abstracts Service (stneasy.cas.org) or WIPO's Intellectual Property Digital Library (ipdl.wipo.int) allow patents to be searched on the basis of their IPC symbols, but there is no online service for finding the correct IPC categories of a new patent application. This crucial initial step must be performed accurately for any subsequent searches to return relevant documents. The IPC covers a range of topics that spans all human inventions and uses a diverse technical and scientific vocabulary. Therefore, it is difficult for non-experts to locate relevant IPC categories. A system supporting the categorization and search of patent applications in the IPC would thus be of interest not only in patent offices, but also across industry.

The CLAIMS (Classification Automated Information System) project at WIPO aims at providing information technology support for the IPC revision process and the IPC reform [4]. A computer-assisted classification tool for categorizing patent applications in the IPC and a natural-language-processing system for the retrieval of patents based on keyword searches are being developed. CLAIMS also encompasses enhancing IPC-based searches in the chemical field by providing chemical formulae and illustrative molecular diagrams in electronic form. These systems should facilitate the attribution of IPC symbols to patent applications and provide for accurate searches in the IPC. CLAIMS focuses particularly on supporting small and medium-sized patent offices of WIPO member states in assigning IPC symbols to patents for their classification and easy retrieval. As patent documents are often available in several languages, CLAIMS aims at a system supporting IPC categorization in four languages and at a language-independent solution for concept-searching the IPC. In this paper, we focus on the first of these needs: computer-assisted patent categorization.

A number of authors have reported on procedures for automating patent classification using machine learning techniques. Chakrabarti et al [6] developed a hierarchical patent classification system using 12 subclasses organized in three levels. In these small-scale tests, the authors found that by accounting for the already-known classifications of cited patents, the effectiveness of the categorization could be improved. Larkey [7] has created a tool for attributing US patent classifications. The inclusion of phrases during document indexing is reported to have increased the system's precision for patent searching but not for categorization [7]. The overall system precision is however not reported. A comprehensive set of patent categorization tests is described by Krier and Zaccà [8]. These authors organized a comparative study of various academic and commercial categorizers, but do not disclose detailed results. The participant with the best results has published his findings separately [9]. Categorization is performed at the level of 44 or 549 categories specific to the internal administration of the European Patent Office, with around 78% and 68% precision respectively when measured with a customized success criterion.

To the best of our knowledge, the only system for computer-assisted patent categorization that has been used for several years in a production environment is the OWAKE system for routing patent applications to human classifiers within the Japanese Intellectual Property Cooperation Center (IPCC). It is primarily designed for handling the Japanese F-term patent symbols and achieves a precision of 90% when pre-classifying Japanese patents into 38 different technical groups. It uses a hierarchical classification scheme that combines an initial rough classification with an automated refining operation to attain the predicted category. It makes use of the full text of the patent but extracts keywords from the document based on an extensive customized Japanese dictionary [10].

In this paper, we provide a quick overview of the IPC taxonomy used for classifying patents internationally. We describe our approach to develop a multilingual categorization system; starting with the collections of documents we have created to train our tool. We report the precision obtained with our prototype implementation. Screenshots of the prototype and an outlook to the future are also provided.

IPC Taxonomy

The International Patent Classification (IPC) is a complex hierarchical classification system comprising sections, classes, subclasses, and groups [11]. The latest edition of the IPC contains eight sections, about 120 classes, about 630 subclasses, and approximately 69,000 groups.[i] The IPC divides all technological fields into sections designated by one of the capital letters A to H, according to: A: "Human necessities"; B: "Performing operations, transporting"; C: "Chemistry, metallurgy"; D: "Textiles, paper"; E: "Fixed constructions"; F: "Mechanical engineering, lighting, heating, weapons, blasting"; G: "Physics"; H: "Electricity". Each section is subdivided into classes, whose symbols consist of the section symbol followed by a two-digit number, such as C01. In turn, each class is divided into several subclasses, whose symbols consist of the class symbol followed by a capital letter, for example, C01B. IPC subclasses are in turn divided into main groups, and then into a hierarchy of subgroups. Table 1 shows a portion of the IPC specification at the start of Section C.

Category	Symbol	Title
Section	C	CHEMISTRY; METALLURGY
Class	C01	INORGANIC CHEMISTRY
Subclass	C01B	NON-METALLIC ELEMENTS; COMPOUNDS THEREOF
Main group (and references for this main group)	C01B3/00	Hydrogen; Gaseous mixtures containing hydrogen; Separation of hydrogen from mixtures containing it (separation of gases by physical means B01D); Purification of hydrogen (production of water-gas or synthesis gas from solid carbonaceous material C10J; purifying or modifying the chemical compositions of combustible gases containing carbon monoxide C10K)

Table 1: Sample portion of the IPC taxonomy at the start of Section C

The IPC exists in two authentic versions, English and French, which are published online (www.wipo.int/classifications) and in printed form by WIPO. Complete texts of the IPC are also prepared and published in other languages by national industrial property offices, which have published versions of the IPC in German, Spanish, Czech, Hungarian, Polish, Russian, Japanese, Korean, and Chinese. In the past, the IPC has been updated every 5 years, and is currently is its 7th edition. Updates are currently mostly made at group and subgroup level. With the future IPC reform, the IPC will be divided in a fixed stable core and more dynamic advanced level that will be updated frequently [4].

Patent categorization in the IPC is complicated by the following factors:

IPC References: Many IPC categories contain references and notes, which serve to guide the classification procedure, as illustrated in Table 1. There are two main types of references: limitations of scope, which serve to restrict the patents classified in the category and which indicate related categories where some patents should preferably be placed, and guidance references, which list related categories where similar patents are classified. The references may list categories that are distant in the IPC hierarchy. In Table 1, for example, a reference to subclass B01D exists in main group C01B3/00. The IPC can thus be thought to contain a multitude of hyperlinks at all levels of category.

Placement rules: Patent classification is governed by placement rules. In certain parts of the IPC, a last-place rule governs the classification of documents relating to two categories at the same hierarchical level, for example in class C07: "Organic Chemistry". This rule indicates that the second of two categories should always be selected if two are found to concord with the subject of the patent application. In other parts of the IPC, different specific rules hold, for example in subclass B32B where a first-place rule holds.

Secondary symbols: A majority of patents do not have a single main IPC symbol, but are also associated with a set of secondary classifications, relating to other aspects expressed in the patent. Experts classifying patents are usually free to attribute any number of additional symbols. The taxonomy thus contains large overlapping categories. In some parts of the IPC, it is obligatory to assign more than one category to a patent document if the patent document meets certain conditions. For example, in subclass C12N, the therapeutic activity of single-cell proteins or enzymes is also classified in subclass A61P.

Vocabulary: The terms used in patents are quite unlike those in other documents, such as the news articles that have been widely used for past categorization benchmarks. Many vague or general terms are often used in order to avoid narrowing the scope of the invention. For example, in the pharmaceutical industry, patent applications tend to recite all possible therapeutic uses for a given compound [5]. Combinations of general terms may have a special meaning that is important to identify. Patent documents also include acronyms and much new terminology [12]. Furthermore, unlike news stories, patents are necessarily all different at a semantic level, as each must describe a new invention. This rule may complicate categorizer training. In addition, the IPC categories often cover a vast and sometimes disparate area, creating thereby a large vocabulary size in some categories. For example, class G09 refers to "Educating, Cryptography, Display, Advertising, Seals". By far the largest contribution to vocabulary diversity is found in class C07: "Organic chemistry", which notably contains tens of thousands of long DNA sequences.

Document collections

In order to develop systems for computer-assisted patent categorization, it is necessary to have access to large collections of manually classified documents on which to train the algorithms to recognize word distributions typical of the categories of interest. We have thus exploited or established several large databases of suitable patent documents. In our prototype implementation of patent categorization, we have made use of four different collections of documents to cover as many languages:

English-language documents: We employ Espace Access abstracts, which consist of European and WIPO Patent Cooperation Treaty (PCT) patent application abstracts, dating from 1978 to 2002.

French-language documents: We use French-language European patents that date from 1978 to 2003 provided by the French Institut National de la Propriété Industrielle (INPI).

Russian-language documents: Rospatent, the Russian Federal Institute of Industrial Property has provided an extensive collection of full-text Russian patents dating from 1994 to 2002.

German-language documents: The German Deutsches Patent- und Markenamt (DPMA) has provided us with a collection of German patents from the DEPAROM collection (www.deparom.de). The application dates of these documents range from 1987 to 2002.

A patent document includes bibliographic information (a title, a list of inventors, a list of applicant companies or individuals), an abstract, a claims section, and a long full-text description. Accompanying figures are not retained for the purpose of categorization. In future tests, and for retraining the categorizer in all languages, we intend to additionally exploit the collection of international patents applications published under the PCT. These full-text documents have been converted to electronic form through partial optical character recognition and are available at WIPO. These PCT documents all have titles and abstracts in English and French. They may also have the abstract in one of the other PCT publication languages, as well as a full-text description and claims in the publication language.

The IPC classification of PCT documents is of the responsibility of various search authorities, which are regional or large national patent offices, such as the Swedish Patent Office. Patent classifiers at these institutions typically hold university degrees and are domain experts responsible for classifying documents in a small subset of the IPC. Nevertheless, the relative subjectivity of human categorization makes IPC classification consistency difficult to achieve across the entire set of training documents. We foresee the possibility of enhancing this consistency—and thus the quality of the training set—by using the more-detailed ECLA classification made by the European Patent Office. This is in anticipation of the next edition of the IPC, which will use the ECLA classification as one of the sources to create a single master classification database with 50 million patent documents classified according to the IPC [4].

All the document collections have been converted to single standard XML format with a customized set of markup tags. In Table 2, we display a sample Russian-language chemistry patent, which also includes a title and abstract in English. The document reference information is provided in the `<record>` tag, which contains the country of origin, the application date, a document reference, the kind of publication, an application number, and a publication number. The main IPC symbol, to main group level, is reported in the `<ipcs>` tag in the `mc` attribute. Additional IPC symbols, when present, are listed in the `ic` attribute of the `<ipc>` tags. Inventors and applicant companies are listed in `<ins>` and `<pas>` tags respectively. The titles, abstracts, claims, and full descriptions are provided in `<tis>`, `<abs>`, `<cls>`, `<txts>` tags respectively.

Part of our document collections are made publicly available on the WIPO website for future work by other researchers (www.wipo.int/ibis/datasets). In this way, WIPO hopes to promote the use of the IPC in research into computer-assisted categorization, both in the academic community and for commercial partners. To date, 25 academic and commercial groups have requested access to our data.

```
<record cy="RU" an="2000125640/12 20001011" date="20011110" pn="RU02175638 C1 20011110"
dnum="02175638" kind="C1">
<ipcs mc="C01B021" ed="7"/>
<ins>
 <in>Иванов Е.Г.</in> <in>Ильин В.А.</in> <in>Селин Е.Н.</in> <in>Гордиенко А.И.</in>
 <in>Аншелес В.Р.</in> <in>Дмитриев Д.А.</in> <in>Ласковец П.В.</in>
</ins>
<pas> <pa>ОАО "Череповецкий "Азот"</pa>
</pas>
<tis>
 <ti xml:lang="EN"> METHOD OF PREPARING NITROUS OXIDE </ti>
 <ti xml:lang="RU"> СПОСОБ ПОЛУЧЕНИЯ ЗАКИСИ АЗОТА </ti>
</tis>
<abs>
 <ab xml:lang="EN"> FIELD: nitrous oxide used for industrial and medical purposes.
SUBSTANCE: in order to prepare nitrous oxide from ammonium nitrate, reaction mixture
containing nitrous oxide and uncondeconsable impurities is washed, prior to being
purified and filled into high-pressure bottles, with solution of nitric acid or
ammonium nitrate or mixture thereof at temperatures from 5 to 35 C, pressures from 20
to 200 mm water column, gas to absorbent ratio of 3-6: 1. In order to prepare nitrous
oxide which meets requirements of pharmocopeia article, low- temperature over-
evaporation of decomposition products is carried out. EFFECT: reduced losses of nitrous
oxide in preparing nitrous oxide of high purity and greater process reliability. 2 cl,
1 dwg, 1 ex </ab>
 <ab xml:lang="RU"> Изобретение относится к получению закиси азота, используемой в
технических и медицинских целях. Сущность способа получения закиси азота разложением
аммиачной селитры состоит в том, что реакционную смесь, содержащую закись азота и
неконденсирующиеся примеси перед осушкой, очисткой и заполнением в баллоны высокого
давления подвергают промывке абсорбентом - водой, раствором азотной кислоты или
аммиачной селитры или их смесью при температуре 5-35°С, давлении 20-200 мм, вод. ст.,
соотношении газ : абсорбент, равном 3-6 : 1. Для получения закиси азота,
соответствующей требованиям фармакопейной статьи, осуществляют низкотемпературное
переиспарение продуктов разложения. Технический результат состоит в сокращении потерь
закиси азота в процессе получения закиси азота высокой чистоты, повышении надежности
процесса. 1 з.п. ф-лы, 1 ил.
</ab>
</abs>
<cls>
 <cl xml:lang="RU"> 1. Способ получения закиси азота термическим разложением аммиачной
селитры, [...abridged...]
</cl>
<txts>
 <txt xml:lang="RU">  [...abridged...] </txt>
</cl>
</record>
```

Table 2: Sample bilingual Russian-English chemical patent document, shown with abridged content, in CLAIMS XML format

Methodology

To develop a categorization system that uses machine-learning techniques, one seeks to implement an algorithm that will learn to recognize topics relevant to each category, based on automatically studying a collection of pre-classified documents in the taxonomy. Our document collections are split into non-overlapping sub-collections of patent documents, which are named training collections and test collections. We use approximately two thirds of the documents in each language to train the patent categorizer, and the remaining third to test the precision of the system.

Our approach to train a computer-assisted patent categorizer relies completely on language-independent processing and makes use only of collections of symbols forming words. The first step, known as indexing, consists of identifying all different words appearing in the training documents. Document indexing is performed at word level, without accounting for phrases or word frequencies in each document. We do not use any word stemming, as this step would be language-dependent. After all the different words are identified, those occurring too often or too infrequently are rejected. A second pass is then made through the patent documents, indexing various XML fields. When collections of fields are indexed for each document, they are added to the same feature vector and receive the same weight. In the tests reported below, we have indexed the titles, the inventors, the applicant companies, and the abstracts. For German language documents, where no abstracts are available, we have used the first 300 different words of the description instead. The vocabulary in the full description of all patents contains a vast number of different words (over a million). Such diversity was found in small-scale tests to be detrimental to the effectiveness of the system, and yielded categorization precisions inferior to the abstracts.

To learn word distributions, we make use of state-of-the-art neural network techniques, and apply in particular a variant of the winnow algorithm [9]. Because of the extremely large number of categories, training documents, and vocabularies, it is necessary to make use of an extremely fast and well-optimized algorithm. Our implementation is entirely custom-built for this purpose. The algorithm learns by sequentially passing through the training documents several times, and thereby updating internal weights related to the frequency of appearance of each word in each category. In this way, the words most characteristic of each category are identified automatically. The learning rates have been optimized from small-scale initial tests. Our system is trained with documents classified on the basis of their main IPC symbol only.[ii] Each document thus appears only once in the training set, despite sometimes being associated with several IPC categories.

The system we develop does not make any explicit use of the text of the IPC. Instead, we rely only on patent documents classified in it manually. In this way, we avoid the difficulties related to category references, to differing placement rules in the IPC, and to the presence of secondary IPC symbols. All such features are automatically accounted for by our system through the symbols attributed to the training documents. However, we must ensure that the training documents use a similar vocabulary to future test documents, which may require retraining the categorizer every few months to adapt to neologisms.

The IPC is a hierarchical taxonomy. We thus build automated categorizers to perform at several level of detail within the IPC: we separately train the system to recognize IPC classes, subclasses, and main groups. We do not make use of subgroups as the number of categories then becomes too large in comparison with the number of documents available for training. We also train additional categorization engines to recognize subclasses within each class, and to distinguish between main groups within each subclass. At the end of training, we thus obtain a powerful hierarchy of categorizers that can be used to refine document categorization, as explained below. Such category refining is unavailable in the vast majority of commercial categorization packages. Implementing it on a standard PC requires careful memory management since we build and store weights for independent categorizers in all classes and subclasses. Thanks to extensive code optimizations, training the full set of neural networks with a corpus of over 800'000 documents takes under 8 hours on a desktop PC. Such convenience means that retraining the system when new documents become available can be easily performed every few months.

Because of our desire to distribute documents evenly across the taxonomy for training, not all IPC subclasses and main groups have been included in our categorization tests. If very few documents have a main IPC symbol in a given category (i.e. less than 30 documents are available in our document collections in any one language), this category has not been included. For our initial tests, we disregard categories with very few documents because the category predictions cannot then be made on a statistical basis. However, as the removal of categories with few training documents necessarily limits the categorization of new patent documents, we intend to study the effect of this removal on system precision. Indeed, the possibility of predicting IPC symbols in new areas of technology is important. Furthermore, small patent offices also need

to be able to predict IPC symbols in areas that describe technology that is not used frequently any more or where novelties are increasingly rare.

Statistics of the document distributions are indicated in Table 3. We note that despite having access to several hundred thousand documents, we do not have a sufficiently uniform distribution of documents to be able to make reliable predictions across all the IPC categories, particularly for Russian and German at main group level.

Language	Total documents	Classes	Sub-classes	Main groups
English	450'000	118	538	3'162
French	830'000	119	569	3'333
Russian	210'000	118	494	1'587
German	248'000	118	500	1'790

Table 3: Total number of documents in our English, French, Russian, and German collections. The numbers of classes, subclasses, and main groups covered is also indicated in each case. Categories with insufficient numbers of training documents are not retained for categorization tests.

The output from the categorization algorithm consists of a ranked list of categories for each test document. We consider three different evaluation measures to flag a categorization success, as shown in Figure 1. The system precision is defined here as the proportion of test documents that fulfill one of three conditions: In the top-prediction scheme (1), we compare the top predicted category with the main IPC category. In the three-guesses approach (2), we compare the top three categories predicted by the classifier with the main IPC class. If a single match is found, the categorization is deemed successful. This measure is most adapted to evaluating categorization assistance—where a user ultimately makes the decision of the correct category. In light of the categorization precisions presented below, this scenario seems most realistic in a business scenario. Finally, in the all-categories scheme (3), we compare the top prediction of the classifier with all categories associated with the document, either in the main IPC symbol or in additional IPC symbols. If a single match is found, the categorization is deemed successful. This evaluation approach was also used earlier [8].

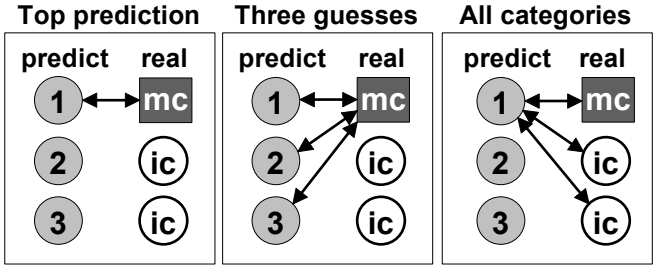

Figure 1: Three measures of categorization success. In the top-prediction scheme, we compare the top predicted category with the main IPC symbol; in the three-guesses scheme, we compare the top three predicted categories with the main IPC symbol; in the all-categories scheme, we compare the top predicted category with the main and additional IPC symbols of each document.

System precision

We train our patent document categorization engine separately in all four languages and the measure the categorization accuracy using the corresponding test documents. In Table 4, we show the categorization precision in English using the three measures of success defined in Figure 1, at class, subclass, and main group levels. We note how the top prediction measure, where the categorizer has a single guess for the correct category, shows a lower precision than when the categorizer has three guesses for the main IPC symbol. It should be pointed out that these results are global precisions across all the IPC. It is expected that the precision within a given area of the IPC will fluctuate according to the number of training documents available therein and the IPC category characteristics. It should be emphasized that the categorization precision only measures the faithfulness of the algorithm to its training set. If a categorizer is trained with patents that not well classified according to the IPC, an ideal algorithm would perform faithfully, i.e. with high precision, and but not lead to practical results of interest. It is therefore also essential to consider the quality of the IPC classification present in the training set when attempting to understand the global quality of predictions and when comparing results obtained with different collections of documents.

Language	Measure	Class	Sub-class	Main group
English	Top prediction	71%	64%	51%
English	Three guesses	90%	85%	72%
English	All categories	80%	74%	64%

Table 4: Categorization precision in English using the three measures of success defined in Figure 1, at class, subclass, and main group levels.

We can compare our results with those published by Krier and Zaccà [8]. They obtained an accuracy of 61% at subclass level with the all-categories measure when using a commercial product implementing a k-NN algorithm. Their result is obtained with an English-language training set containing 68,416 different documents [9]. We see that our precision is 13% better (in absolute terms) than theirs at subclass level. This difference results mainly from the higher number of training documents we employ and from our more sophisticated training algorithm. By using customized developments, we have improved on the precisions we reported earlier with existing categorization products [13].

In Figure 2, we display the precisions achieved by direct categorization at class, subclass, and main group levels using the three guesses measure of precision. We note that the precisions obtained are extremely uniform across languages, thus validating *a posteriori* the language-independent approach employed. The precision obtained for class-level categorization of German documents is slightly below that in the other languages. This is caused by the much larger German vocabulary due to compound word formation. The precisions obtained in Russian and German at main group level are abnormally higher than those in the other languages, due to the smaller number of main groups available with sufficient training documents.

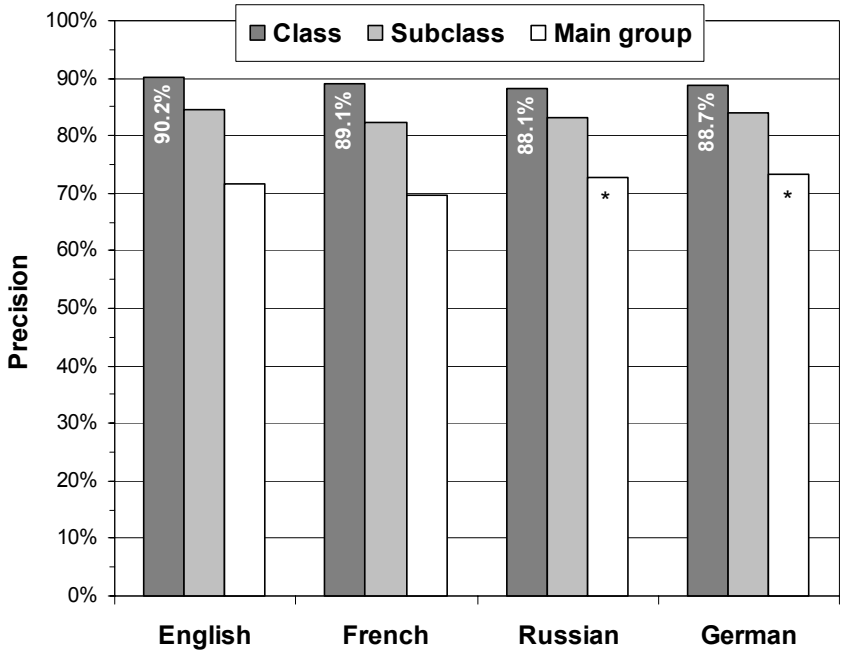

Figure 2: Categorization precision for English, French, Russian, and German-language patent documents at IPC class, subclass, and main-group levels. Results indicated with asterisks correspond to cases where only a small number of main groups are included, thus overestimating slightly the resulting accuracies. Precisions obtained at subclass and main group levels can be improved using category refining.

Our categorization solution is designed to support category refining. In this scenario, instead of directly asking for a predicted main group, the user would request a prediction of the correct class or subclass. When the user has validated the correct (sub)class, the categorization system will predict a main group within it. When refining, the system only has to distinguish between a small number of main groups within the (sub)class rather than the full collection of main groups. Since it has been trained specifically for this task using only the documents within the (sub)class, the resulting accuracy in main group prediction is much improved. Since it relies on human intervention, it is however difficult to make extensive precision statistics on the results.

System interface prototype

In Figure 3, we display a screenshot of the input screen of a prototypical user interface. The user is provided with an upload functionality to input whole documents. The user can alternatively paste an abstract to categorize into the system. The document language is selected at the bottom of the screen, although it would be possible to detect this automatically in the final solution. Such language detection is an additional example of document categorization. The number of predictions requested and the level of categorization are selected at the bottom of the screen. On the left, links to relevant sources of additional information are shown.

Relying primarily on document abstracts for training means that documents abstracts should also be used for querying the system, as the test and training documents must have similar statistical word distributions. In our prototype, it is possible to submit more than a document abstract, but only the first few hundred words will then be retained. Removing the rest of the text avoids introducing noise that may confuse the categorizer. Relying on abstracts has the additional advantage that if no electronic version of the patent

document is available, which is sometimes the case in smaller patent offices, the user will only need to type in the abstract rather than the full document text.

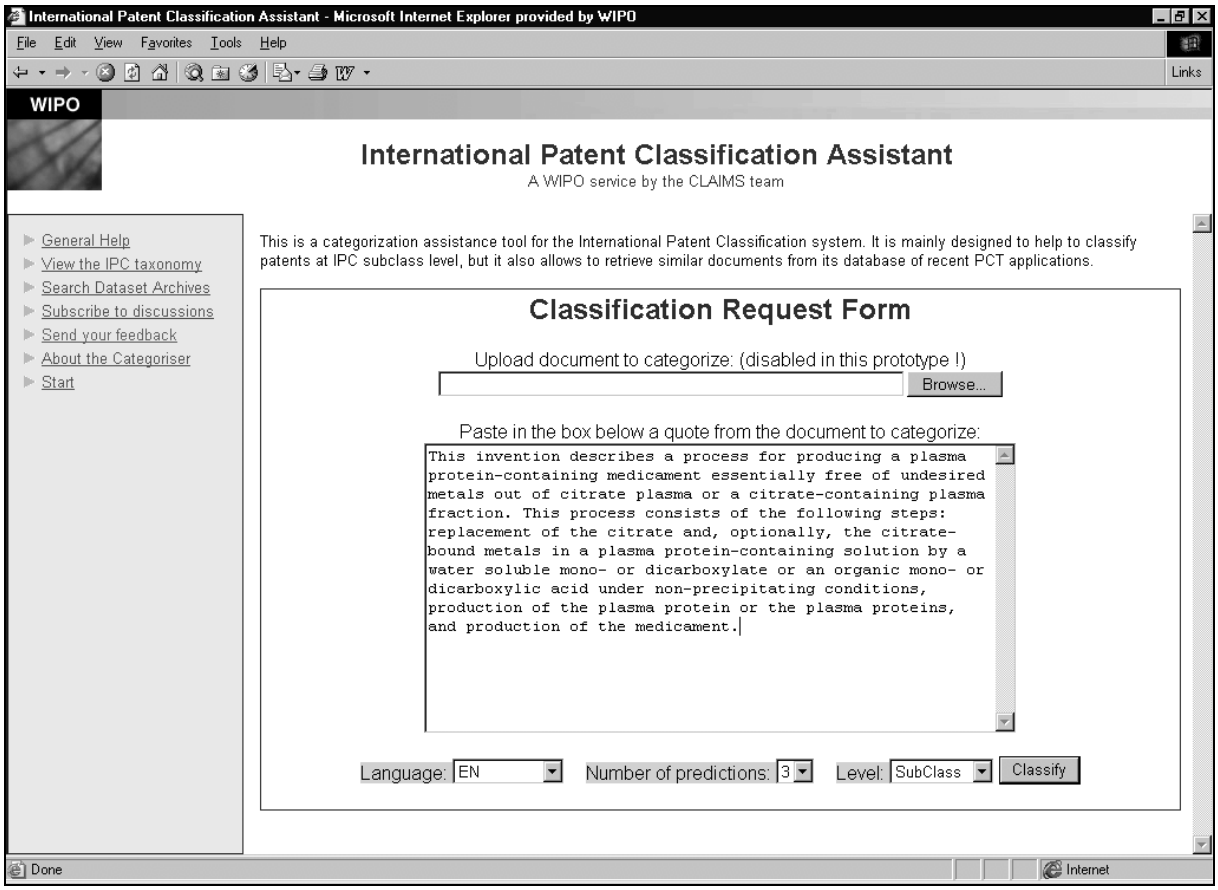

Figure 3: Prototypical user interface of the categorizer: input screen. The user is provided with an upload functionality to input whole documents, and with a box in which to paste an abstract to categorize. The language is selected below, and well as the number of predictions required and the level of categorization. On the left, links to relevant sources of additional information are provided.

When the "Classify" button is pressed, the system responds in a few seconds with a screen similar to that shown in Figure 4, where the result of categorization is displayed. Several suggested categories are listed in order of decreasing confidence, indicated by the number of stars in the first column. Buttons are provided for changing the categorization depth and for refining within categories. Figure 4 displays the result of a refining operation: the user has requested that subclass A23J be refined into main group predictions. The system has then suggested two main groups (shown with a darker background shading), with A23J001 the most probable.

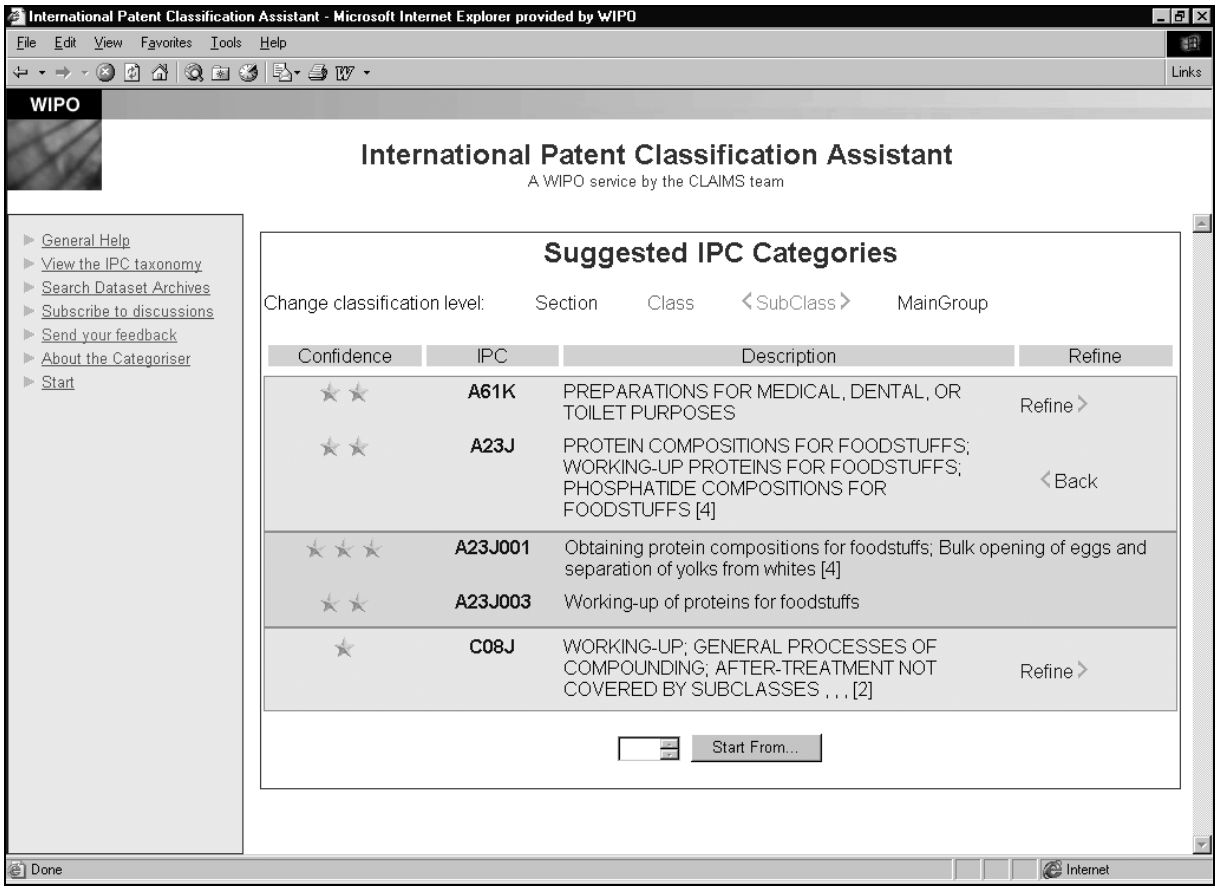

Figure 4: Prototypical user interface of the categorizer: predictions after refining. When the user asks for a refining operation, the categorizer predicts relevant subcategories within the chosen IPC category (here, the subclass prediction A23J has been refined into two main group predictions, A23J001 and A23J003, shown with a slightly darker background).

Conclusions and outlook

Patent categorization provides a demanding test scenario for computer-assisted categorization because of the nature of the taxonomy and the variety of patent documents. Because of this complexity, a custom-built computer-assisted categorized is needed to provide high levels of functionality, fast responses, and to integrate well with existing training documents and online taxonomies.

Our results show that patent classification automation provides high precision and can help users unfamiliar with all the complexities of the International Patent Classification system. Allowing users to refine patent document categorizations provides a convenient, computer-assisted, and interactive approach to classifying patents. If a user can validate the correct IPC class or subclass, the predictions our system provides for main group categorization are much improved. By making use of sophisticated training algorithms in combination with a language-independent approach, we find that categorization can be performed with similar precision in all European languages tested so far. Adapting the system to categorize documents in a new language is a simple procedure once the collection of training documents is available in a standard format. One only needs to run a fully-automatic training procedure on the new documents. If training documents aren't available in the new language, it may nevertheless be possible to categorize documents in it, by making use of translation dictionaries and exploiting the existing document collections. Research in this direction is ongoing. The system described herein is applied to patent classification, but the

categorization method employed is versatile enough to adapt easily to any other hierarchical taxonomy, such as the Engineering Index thesaurus [14], which is also used for chemical indexing.

It is WIPO's ambition to provide a full computer-assisted tool for patent classification in the very near future and work in this direction is ongoing. The full patent categorization solution described herein is expected to be available both as a server-based system from WIPO's offices in Geneva and as a distributed version on a CD for patent offices of WIPO member states. As the same computer-assisted categorization software would be provided to all patent offices, the global IPC classification could also benefit from more consistent predictions. The categorization system will support several European languages, possibly including Spanish. It is also expected that the tool will display documents from its training collection that are most similar to the new document being categorized. The IPC symbols of these earlier documents will provide additional guidance for the user in choosing the correct category.

Acknowledgements

The authors would like to thank Mikhail Makarov, Antonios Farassopoulos, and Gabor Karetka for helpful discussions, Axel Okelmann at DPMA for access to the German dataset, Serge Schambaud at INPI for access to the French dataset, and Alexei Gvinepadze at Rospatent for access to the Russian dataset.

References

1. *International Patent Classification: Guide, Survey of Classes and Summary of Main Groups*, 7th edition, Volume 9, World Intellectual Property Organization, Geneva, 1999.
2. H. Smith. Automation of patent classification, *World Patent Information* 24, 269-271, 2002.
3. D. Hull, S. Aït-Mokhtar, M. Chuat, A. Eisele, E. Gaussier, G. Grefenstette, P. Isabelle, C. Samuelsson, and F. Segond. Language technologies and patent search and classification, *World Patent Information* 23, 265-268, 2001.
4. J. Calvert and M. Makarov. The reform of the IPC, *World Patent Information* 23, 133-136, 2001.
5. T. Vachon, N. Grandjean, and P. Parisot. Interactive Exploration of Patent Data for Competitive Intelligence: Applications in Ulix (Novartis Knowledge Miner), *Proc. Int. Chem. Inform. Conf.*, Nîmes, France, October 2001.
6. S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks, *Proc. SIGMOD98, ACM International Conference on Management of Data*, ACM Press, New York, 307-318, 1998.
7. L. S. Larkey. A Patent Search and Classification System, *Proc. DL-99, 4th ACM Conference on Digital Libraries*, 179-187, 1999.
8. M. Krier and F. Zaccà. Automatic categorization applications at the European Patent Office, *World Patent Information* 24, 187-196, 2002.
9. C. H. A. Koster, M. Seutter, and J. Beney. Classifying Patent Applications with Winnow, *Proc. Benelearn 2001 conf.*, Antwerpen, 2001.
10. K. Kakimoto. Industrial Property Cooperation Center, Tokyo, personal communication, 2003.
11. S. Adams. Using the International Patent Classification in an online environment, *World Patent Information* 22, 291-300, 2000.
12. N. Kando. What shall we evaluate? Preliminary discussion for the NTCIR Patent IR Challenge based on the brainstorming with the specialized intermediaries in patent searching and patent attorneys, *Proc. ACM-SIGIR Workshop on Patent Retrieval*, (pp.37-42). Athens, Greece, July 2000.
13. C. J. Fall, A. Törcsvári, K. Benzineb, G. Karetka. Automated Categorization in the International Patent Classification, *SIGIR Forum* 37 (1), 2003.
14. *Engineering Information Thesaurus*, 4th edition, Elsevier Engineering Information Inc., Hoboken, New Jersey, U.S.A, 2001.

[i] In this paper, the terms "section", "class", "subclass" and "group" refer to these specific IPC subdivisions. When referring to a generic IPC subdivision of any type, we use the word "category".

[ii] Tests showed that the precision was lowered when training was made on the basis of both the main IPC symbol and all additional IPC symbols of each patent document.