Automated Categorization in the International Patent Classification

C. J. Fall[*]
ELCA Informatique SA, Avenue de la Harpe 22-24, CH-1000 Lausanne 13, Switzerland

A. Törcsvári
Arcanum Development, Baranyai utca 10, H-1117 Budapest, Hungary

K. Benzineb
Metaread SA, 9 rue Boissonnas, CH-1227 Genève-Acacias, Switzerland

G. Karetka
World Intellectual Property Organization, 34 Chemin des Colombettes, CH-1211 Genève 20, Switzerland

A new reference collection of patent documents for training and testing automated categorization systems is established and described in detail. This collection is tailored for automating the attribution of international patent classification codes to patent applications and is made publicly available for future research work. We report the results of applying a variety of machine learning algorithms to the automated categorization of English-language patent documents. This procedure involves a complex hierarchical taxonomy, within which we classify documents into 114 classes and 451 subclasses. Several measures of categorization success are described and evaluated. We investigate how best to resolve the training problems related to the attribution of multiple classification codes to each patent document.

Keywords: automated categorization, patent, IPC taxonomy, support vector machines

1 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 crucial to patent-issuing authorities, potential inventors, research and development units, and others concerned with the application or development of technology. The number of patent applications is currently rising rapidly worldwide, creating the need for an automated categorization system (Smith, 2002; Hull et al, 2001; Calvert & Makarov, 2001). In industry, patents are a major source for gathering intelligence about competitors' activities, but this source necessitates sophisticated tools for meaningful data mining (Vachon, 2001).

A number of authors have recently reported on procedures for automating patent classification using machine learning techniques. Chakrabarti et al (1997, 1998) developed a Bayesian hierarchical patent classification system into 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 much improved (Chakrabarti, Dom, & Indyk, 1998). Larkey (1998, 1999) has created a tool for attributing US patent codes based on a k-Nearest Neighbors (k-NN) approach. The inclusion of phrases during indexing is reported to have increased the system's precision for patent searching but not for categorization

(Larkey, 1998). The overall system precision is however not described. Kohnen et al (2000) achieved a precision of 60.6% when classifying patents into 21 categories in the course of creating a two-dimensional map of patent documents. Gey et al (2001) have created a web-based solution for attributing patent codes in the US and international systems, but have not performed detailed tests of precision. A comprehensive set of patent categorization tests is reported in Krier & Zaccà (2002). 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 (Koster, 2001). 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 automated patent categorization that is currently used in a production environment is the OWAKE system for routing patent applications to human classifiers within the Japanese Intellectual Property Cooperation Center (IPCC). Little, if anything, has been published in English about this customized system. It is primarily designed for handling the Japanese F-term patent codes and reportedly achieves a precision of 90% when classifying Japanese patents in 38 different technical groups. It uses a hierarchical classification scheme that combines a Rocchio-like algorithm for rough classification with a k-NN approach for refining 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 (Kakimoto, 2003).

Overall, published results on patent categorization often suffer from a lack of transparency or are focused on an industry-specific application. Instead, we seek here to describe a new open collection of documents made available to the research community for benchmark tests of automated categorization and research into patent classification across a wide variety of topics.

The International Patent Classification (IPC) is a standard taxonomy developed and administered by the World Intellectual Property Organization (WIPO) for classifying patents and patent applications (WIPO, 1999). The IPC covers all areas of technology and is currently used by the industrial property offices of more than 90 countries. The use of patent documents and the IPC for research into automated categorization is interesting for several reasons. The IPC covers a range of topics that spans all human inventions and uses a diverse technical and scientific vocabulary. A large part of it is concerned with chemistry, mechanics, and electronics. Necessarily, the IPC is thus a complex, hierarchical taxonomy, which has been refined for 30 years. Over 40 million documents have been classified in it worldwide. Furthermore, all domain experts in national and regional patent offices currently classify patent documents manually. These experts have an intimate knowledge of the IPC, and aim to provide excellent and consistent classifications. Finally, patent documents are often available in several languages. Professional translators have already performed large numbers of translations of patent documents.

In this paper, we provide an overview of the complex taxonomy used for classifying patents internationally. We describe the collection of documents we have created to train and test automated categorization tools and report results obtained with a variety of algorithms and parameters. Our conclusions highlight the lessons learned.

2 IPC Taxonomy

The International Patent Classification (IPC) is a complex hierarchical classification system comprising sections, classes, subclasses and groups. 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 A01. In turn, each class is divided into several subclasses, whose symbols consist of the class symbol followed by a capital letter, for example, A01B. Table 1 shows a portion of the IPC specification at the start of Section A. An introduction to the taxonomy may be found in Adams (2000). In this work we do not make use of IPC subdivisions below subclass level, which consist of main groups and of a hierarchy of subgroups, as the number of categories becomes very large in comparison with the number of documents in the collection we exploit.

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. IPC classes and subclasses are essentially stable. In the future, the IPC will be divided in a fixed stable core and more dynamic advanced level that will be updated more frequently than at present (Calvert & Makarov, 2001).

Section	SECTION A — HUMAN NECESSITIES
Class	A01 AGRICULTURE, FORESTRY, ANIMAL HUSBANDRY, HUNTING, TRAPPING, FISHING
Subclass	A01B SOIL WORKING IN AGRICULTURE OR FORESTRY; PARTS, DETAILS, OR ACCESSORIES OF AGRICULTURAL MACHINES OR IMPLEMENTS, IN GENERAL
References for this subclass	(making or covering furrows or holes for sowing, planting or manuring A01C 5/00; machines for harvesting root crops A01D; mowers convertible to soil working apparatus or capable of soil working A01D 42/04; mowers combined with soil working implements A01D 43/12; soil working for engineering purposes E01, E02, E21)

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

The IPC is specified in extreme detail, covering 9 volumes in printed form. 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 class E01 exists in subclass A01B. 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. 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 codes: A majority of patents do not have a single main IPC code, 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 codes. 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 A61K, a cosmetic preparation with therapeutic properties must also be classified in A61P. A list of such IPC categories is shown in Table 2.

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 (Lewis et al, 2003). 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 (Vachon, 2001). Combinations of general terms may have a special meaning that is important to identify. Patent documents also include acronyms and much new terminology (Kando, 2000). 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”.

A61K → A61P	C07 → A61P	C12P → C07, C08
B29C → B32B	C08 → A61P	C25B → C01, C07
B29D → B32B	C08J → C08L	D06N → B32B
C01 → A61P	C12N → A61P	G01V → G01S

Table 2: List of IPC classes and subclasses included in WIPO-alpha (detailed in section 2.1) where double classifications are required, at least in some parts of the relevant IPC category and for patent applications that meet certain criteria. A → B indicates that if a patent has main symbol A, it may also require additional symbol B. The reverse is not necessarily true.

2.1 Document collection

In order to perform automated patent categorization, we have collected a large database of suitable patent documents that we name WIPO-alpha. This collection is made publicly available on the WIPO website for future work by other researchers (www.wipo.int/ibis/datasets).

The documents in the WIPO-alpha collection consist of patent applications submitted to WIPO under the Patent Cooperation Treaty (PCT). A patent application includes a title, a list of inventors, a list of applicant companies or individuals, an abstract, a claims section, and a long description. Accompanying figures may also be present, but these are not retained in the WIPO-alpha collection. The documents in our collection are in English and were published between 1998 and 2002. This restriction on dates is mainly governed by the availability of full-text PCT documents in electronic form at WIPO.

The patent applications in the WIPO-alpha collection originate from a variety of countries worldwide. They have been classified in the IPC by PCT designated search authorities, which are a small number of large national or regional patent offices (such as the European Patent Office). The applicant can sometimes choose the search authority, but the identity of the search authority for each of our documents is not available. Patent classifiers at these institutions typically hold university degrees and are domain experts responsible for classifying documents in a small subset of the IPC. Because patent classifiers each cover separate areas, it is difficult to measure accurately the inter-judge agreement for PCT applications. While all IPC codes have been allotted with extreme care, it is possible that two patent offices would classify similar documents differently, particularly in categories with overlapping content. Agreement on classifications is primary guided by specifying categorization rules within the text of the IPC and through committee meetings at WIPO where delegates from all member states are invited to participate in the revisions of the IPC. It should be pointed out that the European, Japanese, and US patent offices, which all classify PCT documents, each possess their own separate patent classification system and use the IPC as an additional taxonomy.

The documents have been converted to electronic form by optical character recognition (OCR). Because of extensive automatic and manual checking of the resulting text, few cases of OCR errors are expected to subsist. Because patent applications are sometimes republished several times in slightly different form following modifications by the inventors, the WIPO-alpha collection may contain a residual number of duplicate or very similar records carried over from the source data. Extensive efforts, based on fields in our source database, have however been made to remove such (near) duplication.

The document collection consists of a set of XML documents with a customized set of markup tags. A sample XML document from the WIPO-alpha collection is provided in Table 3. The document reference information in the `<record>` tag contains the country of origin in a `cy` attribute (equal to WO for the PCT documents in WIPO-alpha), a document reference number (`dnum` attribute), the kind of publication type (`kind` attribute), an application number and a publication number (`an` and `pn` respectively). Priority numbers in a `<prs>` tag indicate patent publication numbers and publication dates in various countries.

A patent application always has a main IPC code, which is reported in the `<ipcs>` tag in the `mc` attribute. Additional secondary IPC codes are listed in the `ic` attribute of the `<ipc>` tags. The IPC class code is obtained by retaining the first three characters of the IPC symbol (D01, for the document in Table 3), while the subclass is obtained by retaining the first four characters of the IPC symbol (D01D, here). The IPC edition used for the classification is indicated as an `ed` attribute of the `<ipcs>` tag. We mainly use documents from IPC edition 7, but include some documents from edition 6 if their main IPC still exists in edition 7.

Inventors and applicant companies are listed in `<ins>` and `<pas>` tags respectively. In the WIPO-alpha dataset, the title, abstract, claims, and full description are provided in English, in `<tis>`, `<abs>`, `<cls>`, `<txts>` tags respectively. The claims section is set in a special legalistic language and serves to determine the exact scope of the future patent.

Submitted patent applications typically cite a number of earlier granted patents in related domains, for which the IPC categories are already known. This information may in principle also guide the categorization of the new application. Due to restrictions imposed by the source data exploited here, the list of cited patents are not specifically tagged in our XML data, but references to other documents are often found in the full description field.

The WIPO-alpha collection consists of two randomly split non-overlapping sub-collections of patent applications, which are named training collection and test collection. The training collection consists of documents roughly evenly spread across the IPC main groups, subject to the restriction that each subclass contains between 20 and 2000 documents. The exact distribution is a reflection of the data at hand. The test collection consists of documents approximately distributed according to the frequency of a typical year's patent applications (year 2001 was used for this purpose), subject to the restriction that each subclass contains between 10 and 1000 documents. The test collection is formed in this way, as it is then hoped to reflect the typical use a categorization system would undergo in a business production setting. We disregard categories with very few documents because some parts of the IPC describe technology that is not used frequently any more (or where novelties are increasingly rare) and because we thus hope to achieve a higher absolute number of correct predictions when categorizing a typical year's documents.[ii]

```
<?xml version="1.0" encoding="iso-8859-1"?>
<record cy="WO" an="US0024942" pn="WO012006320010322" dnum="0120063" kind="A1">
<prs> <pr prn="US19990914 60/153,825"/> </prs>
<ipcs ed="7" mc="D01D00106">
       <ipc ic="D01F00110"></ipc>
       <ipc ic="D01F00606"></ipc> </ipcs>
<ins> <in>COOK, Michael, Charles</in>
       <in>MCDOWALL, Debra, Jean</in>
       <in>STANO, Dana, Elizabeth</in> </ins>
<pas> <pa>KIMBERLY-CLARK WORLDWIDE, INC.</pa> </pas>
<tis> <ti xml:lang="EN">METHOD OF FORMING A TREATED FIBER AND A TREATED FIBER FORMED
        THEREFROM</ti> </tis>
<abs> <ab xml:lang="EN">The present disclosure is directed to a method of forming a
        treated fiber. A molten polymer is delivered to a fiber spinning assembly
        adapted to form and distribute polymer streams. At least one treatment is
        applied in a liquid state to at least one region on the surface of at least
        one molten polymer stream within the fiber spinning assembly. A substantial
        portion of the treatment remains on the surface of the resulting fiber
        within the treated region. One or more regions on the surface of the molten
        polymer may be treated with one or multiple treatments. The degree of
        coverage may vary from little coverage to complete coverage of the fiber
        surface. The treated regions may be in contact with one another or may be
        separate and distinct. A nonwoven web may be produced with selectively
        treated fiber regions by designing one or more fiber spinning assemblies to
        treat selected fibers or to apply multiple treatments. The regions of the
        nonwoven web may vary in treatment type, amount, or degree of coverage.</ab>
        </abs>
<cls> <cl xml:lang="EN">We claim:
        1. A method of forming a treated fiber comprising:
        a) providing a molten polymer;
        b) delivering said molten polymer to a fiber spinning assembly adapted to
        form and distribute a stream of said molten polymer; andc) applying a
        treatment in a liquid state to at least one region on the surface of said
        molten polymer stream within said fiber spinning assembly, such that a
        substantial portion of said treatment remains on the surface of the
        resulting fiber within said treated region.2. The method of claim 1, wherein
        said treatment has a boiling point of at least about 300°F / 149°C.
        [… abridged …]
</cl> </cls>
<txts> <txt xml:lang="EN"> METHOD OF FORMING A TREATED FIBER AND A TREATED FIBER
        FORMED THEREFROM
        Field of the Invention
        The present invention relates to a treated fiber and a method of forming a
        treated fiber. Such treated fibers find many applications, for example, in
        nonwoven fabrics, yarns, carpets, and otherwise where fibers having one or
        more modified properties are desired.
        Background of the Invention
        Nonwoven fabrics are finding increasing use in various applications,
        including personal care absorbent articles such as diapers, training pants,
        incontinence garments, mattress pads, wipers, and feminine care products (e.
        g., sanitary napkins), medical applications such as surgical drapes, gowns,
        wound care dressings, and facemasks, articles of clothing or portions
        thereof including industrial workwear and lab coats, household and
        industrial operations including liquid and air filtration, and the like. It
        is often desirable to modify the properties of the nonwoven fabric to
        perform a function or meet a requirement for a particular application.
        [… abridged …]
        Having thus described the invention in detail, it should be apparent that
        various modifications can be made in the present invention without departing
        from the spirit and scope of the following claims.</txt> </txts>
</record>
```

Table 3: Sample patent record in WIPO-alpha, with abridged content

The training and test distributions are shown in Figure 1. Except for a few outlying classes, the training and test collections follow similar distributions, indicating that IPC groups usually contain similar numbers of documents. However classes A61: “Medical or veterinary science, hygiene”, H04: “Electric communication technique”, G06: “computing, calculating, counting”, and C12: “Biochemistry” currently receive a disproportionately large number of applications. Because of the desire to distribute documents evenly across the taxonomy, not all IPC subclasses have been included in WIPO-alpha. If very few documents have a main IPC symbol in a given subclass, this subclass has not been included. The main IPC symbols of the WIPO-alpha collection are distributed in 114 classes and 451 subclasses (out of about 650 possible subclasses).[iii] Statistics of the document distribution at class and subclass levels are indicated in Table 4.

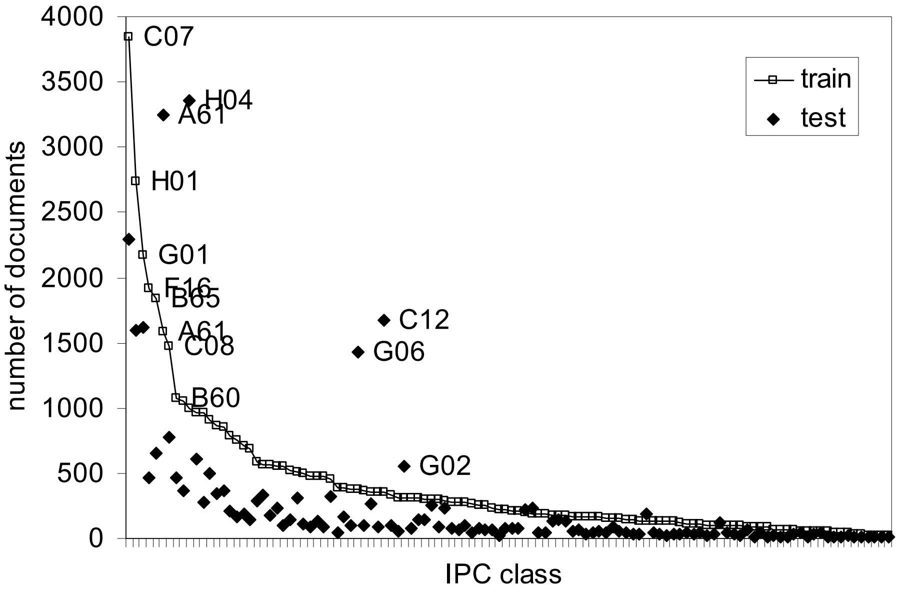


Figure 1: Distribution of training and test documents in WIPO-alpha at IPC class level. The numbers of training and test documents of a given class are aligned vertically. The classes are sorted by decreasing numbers of training documents. Classes with the largest numbers of training documents are labeled, as are those with many more test documents than training documents. Note that it is training subclasses that are restricted to 2000 documents and this figure show the distribution per class.

	Number of documents	Average num. docs per class	Median num. docs per class	Average num. docs per subclass	Median num. docs per subclass
Train collection	46,324	406	213	102	61
Test collection	28,926	253	78	64	19
Total	75,250				

Table 4: WIPO-alpha document distribution statistics

3 Methodology

We make use of two automated categorization tools for a series of multi-classification ranking tasks: the “rainbow” package, part of the bow (bag-of-words) toolkit (McCallum, 1996), and the “SNoW” (sparse network of winnows) learning architecture (Carlson, 1999). The rainbow package implements multinomial Naïve Bayes (NB), k-Nearest Neighbors (k-NN), and Support Vector Machine (SVM) algorithms. Indexing is performed at word level, accounting for word frequencies in each document, 524 common stopwords are removed, word stemming is implemented with the Porter algorithm, and term selection is made on the basis of information gain. In preliminary tests, we have found that word stemming has little effect on the system effectiveness, but lowers runtimes. We thus make use of stemming in all rainbow results reported. The SNoW package is tailored for learning in large feature spaces and implements a sparse network of linear functions where the class labels are represented as linear functions over a common feature space (Carlson, 1999). A variation of the winnow update rule is used for training. SNoW is run with word occurrence indexing (binary weighting), which was found to produce better results than word frequency indexing, the same collection of stopwords as rainbow and without stemming. Word indexing for both packages disregards words with accented characters, which are sometimes present in the patent inventors and the applicant companies.

For the NB and k-NN algorithms, we do not perform any term selection and use the full training vocabulary. Contrary to the experience reported in Lewis et al (2003), term selection performed with the k-NN algorithm did not significantly improve the system precision. The k-NN algorithm samples the 30 closest neighbors throughout all tests. The SVM algorithm uses a linear kernel and for performance reasons we limit the vocabulary to 20,000 words and at most 500 documents per class (for a total of 29,712 training documents) or 100 documents per subclass (for a total of 28,635 training documents). The SNoW algorithm is run with a winnow promotion parameter equal to the inverse of the demotion parameter. These learning update parameters have been optimized from initial tests and are kept fixed throughout this report. SNoW learns by sequentially passing five times through the training documents. Our algorithms are trained with documents pre-classified on the basis of their main IPC code only.[iv] Each document thus appears only once in the training set, despite sometimes being associated with several IPC categories.

We index various fields or collections of XML fields of the patent documents and then perform categorization tests on them. When collections of fields are indexed for each document, they are added to the same feature vector and receive the same weight. The titles (a) or the claims sections (b) are first used to describe each document, yielding vocabulary sizes of around 25,000 and 200,000 words. Next, the first 300 words of the titles, inventors, applicants, abstracts, and descriptions (c) are indexed, covering about 200,000 words. This procedure normalizes the length of the training and test documents and was chosen for simplicity, computational efficiency, and highest precision, as detailed below. Finally, the titles, inventors, applicants, and abstracts only (d) are used for indexing each patent, for a vocabulary of about 150,000 words. The average length of the abstracts is about 130 words (Koster, 2001). One should note that the number of words varies strongly among patents. After stopword removal and without stemming, one 6,200-word document only contained 275 different words, while one 1,400,000-word document contained 4,280 different words. There can thus be a significant amount of word repetition in patent applications.

The vocabulary in the full description of all patents contains over 1,400,000 different words. Such diversity was found in small-scale tests to be detrimental to the effectiveness of the system, and yielded categorization precisions inferior to both the abstracts and the first 300 words. By far the largest contribution to vocabulary diversity is found in class C07: “Organic chemistry”, which notably contains tens of thousands of DNA sequences. Future studies might benefit from a filtering of these easily-identifiable words.

The output from the categorizers consists of a ranked list of categories for each test document. In the following investigations, we consider three different evaluation measures to flag a categorization success, as shown in Figure 2. These evaluations serve different purposes, depending on whether an autonomous classifier is sought or whether a human-assistance tool is required. They have been chosen as they are felt most adapted to the case at hand, where documents are often associated with more than one category but always have a single main classification. The (micro-averaged) precision is defined here as the number 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, chosen on the basis of the confidence levels computed by the various algorithms, with the main IPC class. If a single match is found, the categorization is deemed successful. This measure is 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. In this case, it is tolerable that the correct guess appears second or third in the list of suggestions. Finally, in the all-categories scheme (3), we compare the top prediction of the classifier with all categories associated with the document and present in the main IPC symbol and in additional IPC symbols. If a single match is found, the categorization is deemed successful. This evaluation approach was also used in Krier & Zaccà (2002).

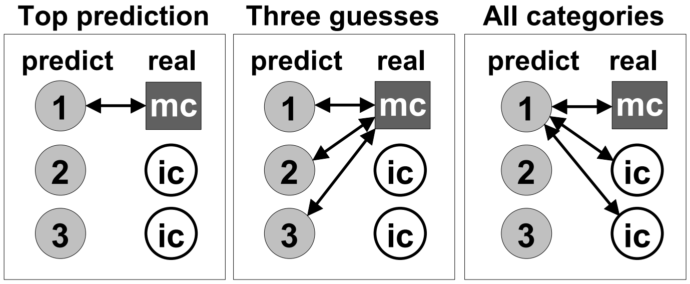


Figure 2: Three measures of categorization success. In the top-prediction scheme, we compare the top predicted category with the main IPC symbol; in the three-guesses scheme, we compare the top three predicted categories with the main IPC symbol; in the all-categories scheme, we compare the top predicted category with the main and additional IPC symbols of each document.

4 Results and analysis

We successively present results at IPC class level, where we investigate the best patent document fields to index, and at IPC sub-class level, where automated categorization would be more useful as the categories are more closely focused on a single topic. In each case, we evaluate the strengths and weaknesses of the algorithms tested.

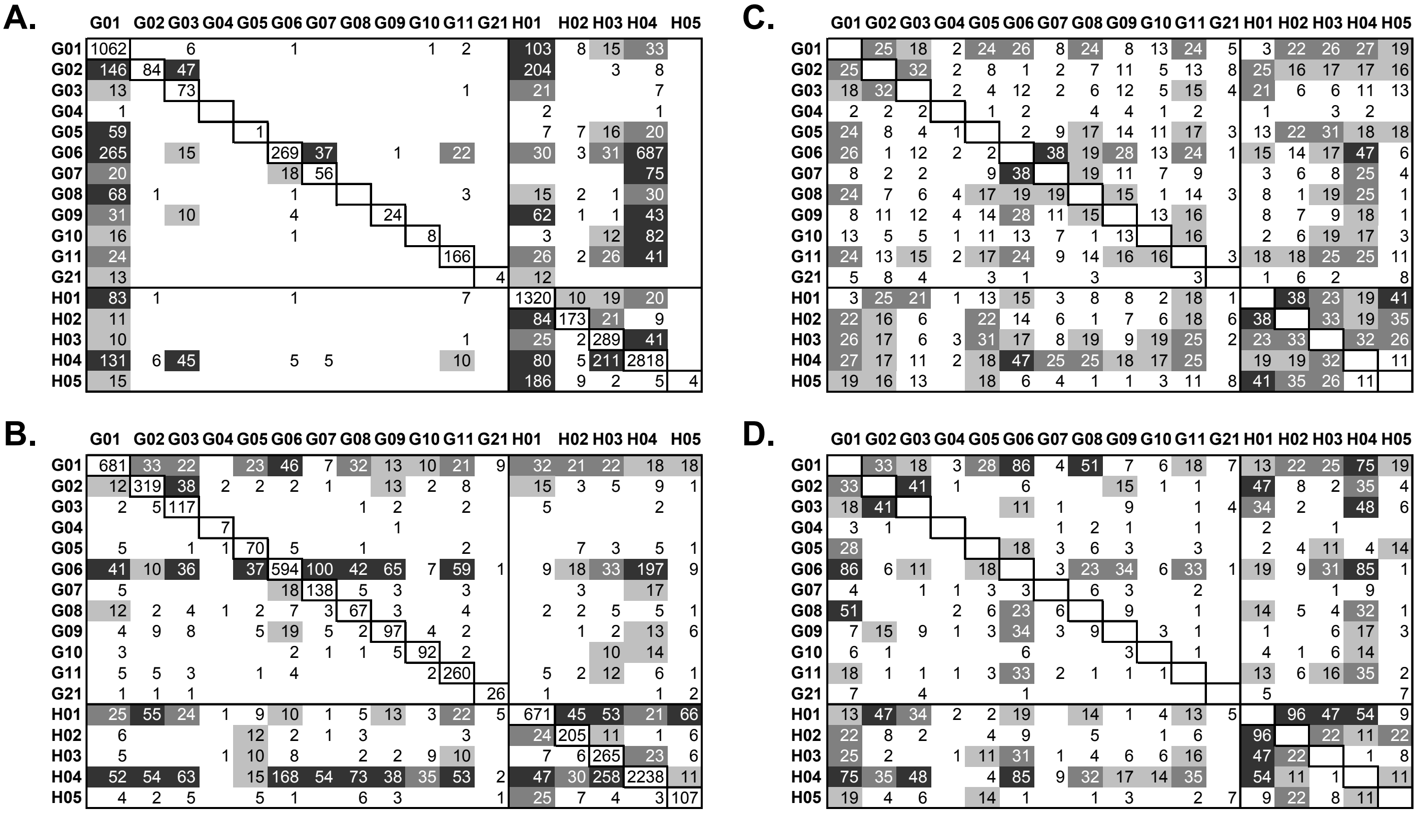
4.1 Class-level categorization

In Table 5, we display the results of class-level categorizations of all test documents on the basis of (a) their titles, (b) the claims sections, and (c) the first 300 words, for the three evaluation measures.[v] We find that best performance is achieved when indexing the first 300 words of each document. This is similar to the result of Larkey (1999), who reports best performance when indexing the title, abstract, first 20 lines of the description, and claims, with the titles receiving three times as much weight as the others. The claims section provides consistently poorer results, despite a vocabulary of similar size. The titles, despite being a single-line in length, provide a remarkably good result, confirming Larkey's high weighting attributed to them.

Indexing fields	Evaluation measure	Naïve Bayes	k-NN	SVM	SNoW
(a) Titles	(1) Top-prediction	45%	33%	-	45%
(b) Claims	(1) Top-prediction	50%	18%	45%	48%
(c) First 300 words	(1) Top-prediction	55%	51%	55%	51%
(a) Titles	(2) Three-guesses	66%	52%	-	64%
(b) Claims	(2) Three-guesses	72%	35%	63%	74%
(c) First 300 words	(2) Three-guesses	79%	77%	73%	73%
(a) Titles	(3) All-categories	52%	38%	-	50%
(b) Claims	(3) All-categories	57%	23%	52%	54%
(c) First 300 words	(3) All-categories	63%	58%	62%	58%

Table 5: Categorization precision at IPC class level, for categorization of all test documents. The SVM algorithm uses a smaller vocabulary and is limited to 500 training examples per class. The SVM runs were unsuccessful on titles.

Although the precisions of the Naïve Bayes and SVM algorithms appear similar when categorizing with the first 300 words, one should not forget that the SVM algorithm is only using two thirds of the total training set. The distribution of errors of these two algorithms is strikingly different, as shown in panels A and B respectively of Figure 3, where we display a fraction of the confusion matrix computed with the top-prediction measure, in a part of the IPC that has a particularly large number of errors. In panel A, we note a structure in columns for the NB algorithm, with a high number of errors in classes that contain large numbers of training documents. The NB algorithm thus tends to erroneously aggregate documents from small classes into the larger ones. However, for the SVM algorithm, the confusion matrix in panel B displays a structure in lines, indicating that it tends to distribute documents from large classes into the smaller ones. This may result partly from the fact that large classes have not been trained with all available documents. We observe in Figure 3 a significant amount of confusion between sections G and H, a feature not seen strongly in others parts of the IPC. We also note, coincidentally, that the instructions in the IPC indicate that section G may be the hardest to classify manually since distinctions between fields rest on intentions of the user, rather than constructional differences or manners of use, and because subjects are often combinations of elements (WIPO, 1999).



A.

	G01	G02	G03	G04	G05	G06	G07	G08	G09	G10	G11	G21	H01	H02	H03	H04	H05
G01	1062		6			1				1	2		103	8	15	33	
G02	146	84	47										204		3	8	
G03	13		73								1		21			7	
G04	1												2			1	
G05	59				1								7	7	16	20	
G06	265		15			269	37		1		22		30	3	31	687	
G07	20					18	56									75	
G08	68	1				1					3		15	2	1	30	
G09	31		10			4			24				62	1	1	43	
G10	16					1				8			3		12	82	
G11	24										166		26	2	26	41	
G21	13											4	12				
H01	83	1				1					7		1320	10	19	20	
H02	11												84	173	21	9	
H03	10										1		25	2	289	41	
H04	131	6	45			5	5				10		80	5	211	2818	
H05	15												186	9	2	5	4

C.

	G01	G02	G03	G04	G05	G06	G07	G08	G09	G10	G11	G21	H01	H02	H03	H04	H05
G01		25	18	2	24	26	8	24	8	13	24	5	3	22	26	27	19
G02	25		32	2	8	1	2	7	11	5	13	8	25	16	17	17	16
G03	18	32		2	4	12	2	6	12	5	15	4	21	6	6	11	13
G04	2	2	2		1	2		4	4	1	2		1		3	2	
G05	24	8	4	1		2	9	17	14	11	17	3	13	22	31	18	18
G06	26	1	12	2	2		38	19	28	13	24	1	15	14	17	47	6
G07	8	2	2		9	38		19	11	7	9		3	6	8	25	4
G08	24	7	6	4	17	19	19		15	1	14	3	8	1	19	25	1
G09	8	11	12	4	14	28	11	15		13	16		8	7	9	18	1
G10	13	5	5	1	11	13	7	1	13		16		2	6	19	17	3
G11	24	13	15	2	17	24	9	14	16	16		3	18	18	25	25	11
G21	5	8	4		3	1		3			3		1	6	2		8
H01	3	25	21	1	13	15	3	8	8	2	18	1		38	23	19	41
H02	22	16	6		22	14	6	1	7	6	18	6	38		33	19	35
H03	26	17	6	3	31	17	8	19	9	19	25	2	23	33		32	26
H04	27	17	11	2	18	47	25	25	18	17	25		19	19	32		11
H05	19	16	13		18	6	4	1	1	3	11	8	41	35	26	11	

B.

	G01	G02	G03	G04	G05	G06	G07	G08	G09	G10	G11	G21	H01	H02	H03	H04	H05
G01	681	33	22		23	46	7	32	13	10	21	9	32	21	22	18	18
G02	12	319	38	2	2	2	1		13	2	8		15	3	5	9	1
G03	2	5	117					1	2		2		5			2	
G04				7					1								
G05	5		1	1	70	5		1			2			7	3	5	1
G06	41	10	36		37	594	100	42	65	7	59	1	9	18	33	197	9
G07	5					18	138	5	3		3			3		17	
G08	12	2	4	1	2	7	3	67	3		4		2	2	5	5	1
G09	4	9	8		5	19	5	2	97	4	2			1	2	13	6
G10	3					2	1	1	5	92	2				10	14	
G11	5	5	3		1	4				2	260		5	2	12	6	1
G21	1	1	1									26	1			1	2
H01	25	55	24	1	9	10	1	5	13	3	22	5	671	45	53	21	66
H02	6				12	2	1	3			3		24	205	11	1	6
H03	5			1	10	8		2	2	9	10		7	6	265	23	6
H04	52	54	63		15	168	54	73	38	35	53	2	47	30	258	2238	11
H05	4	2	5		5	1		6	3			1	25	7	4	3	107

D.

	G01	G02	G03	G04	G05	G06	G07	G08	G09	G10	G11	G21	H01	H02	H03	H04	H05
G01		33	18	3	28	86	4	51	7	6	18	7	13	22	25	75	19
G02	33		41	1		6			15	1	1		47	8	2	35	4
G03	18	41				11	1		9		1	4	34	2		48	6
G04	3	1					1	2	1		1		2		1		
G05	28					18	3	6	3		3		2	4	11	4	14
G06	86	6	11		18		3	23	34	6	33	1	19	9	31	85	1
G07	4		1	1	3	3		6	3		2				1	9	
G08	51			2	6	23	6		9		1		14	5	4	32	1
G09	7	15	9	1	3	34	3	9		3	1		1		6	17	3
G10	6	1				6			3		1		4	1	6	14	
G11	18	1	1	1	3	33	2	1	1	1			13	6	16	35	2
G21	7		4			1							5				7
H01	13	47	34	2	2	19		14	1	4	13	5		96	47	54	9
H02	22	8	2		4	9		5		1	6		96		22	11	22
H03	25	2		1	11	31	1	4	6	6	16		47	22		1	8
H04	75	35	48		4	85	9	32	17	14	35		54	11	1		11
H05	19	4	6		14	1		1	3		2	7	9	22	8	11	

Figure 3: Panels A and B: Extracts of the confusion matrices in sections G and H of the IPC class-level categorization with first 300 words, using the NB and SVM algorithms respectively. The row is the real category and the column is the predicted category; the numbers in the matrices indicate the absolute numbers of documents categorized right (numbers on the diagonal) and wrong (off-diagonal values). Panels C and D: Predicted category similarities based on the vocabulary similarity of the training documents and the distribution of training documents, respectively. See main text for details. In all cases, shading indicates high off-diagonal values.

The identification of IPC areas where automated categorization is difficult depends to a certain extent on the algorithm employed. In Figure 3, we also show two models that nevertheless predict the distribution of errors with some success. The first model is based on the vocabulary of the training documents: for each class, we compute the 100 words with the highest odds ratio. In panel C of Figure 3, we report how many of these discriminating words are in common between the IPC classes. The second model is based on the distribution of documents: by examining both the main and the additional IPC codes of all training documents, we count how many documents are simultaneously associated with two classes, and report these numbers in panel D of Figure 3. Both models necessarily result in symmetric matrices, and highlight similar areas of the IPC where difficulties are expected. The results of the SVM algorithm are particularly well predicted.

In Figure 4, we display how the system precision changes when a fraction of the test documents are classified. Results are reported for the SVM algorithm on the first 300 words. We filter documents for categorization on the basis of the confidence level of the top predicted category. As the fraction of documents classified is reduced from 100%, the precision increases, indicating that the algorithm is successful at predicting its own precision. If 80% of documents are classified, the categorization is successful in 78% of cases when using the three-guesses measure.

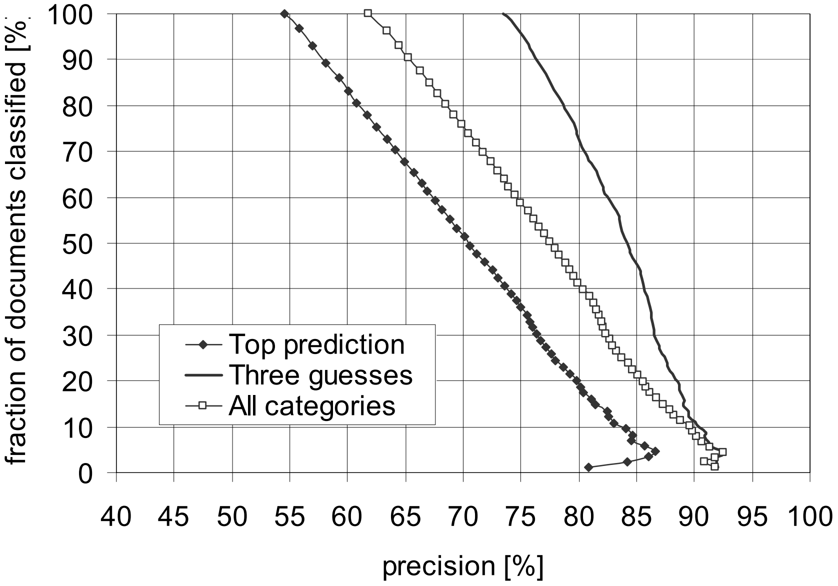


Figure 4: Variation of categorization precision with the fraction of documents classified for SVM class-level categorization using the first 300 words. Results are reported for the three measures of categorization success.

4.2 Subclass categorization

In Table 6, we display the results of various algorithms applied to IPC subclass categorization of all test documents, on the basis of: (c) the first 300 words of each document; (d) their titles, inventors, applicants, and abstracts. Because of computational limitations, the SVM algorithm used the first 150 words, rather than the first 300 words. We note that despite being trained with fewer documents, using a smaller vocabulary, and shorter documents, the SVM algorithm is most successful at predicting the main IPC subclass, or any of the real attributed subclasses. If three guesses are allowed for the main subclass, the k-NN algorithm performs best.

Except for the SNoW algorithm, it is more advantageous to use the first few hundred words of the document than the abstracts for describing each document. We have also performed hybrid tests with the Naive Bayes algorithm, training with the first 300 words and testing with abstracts. In this way, we hoped to build good word statistics in each category, and simultaneously benefit during testing from the human selection implicit in the abstracts. However, the resulting accuracies were intermediate between those using abstracts for training and testing and those using the first 300 words for training and testing. It is thus better to test also with the first 300 words.

Some of our training documents have multiple IPC codes, thus implicitly belonging to several categories (even though they are effectively each used only once for training, according to their main IPC symbol). We attempt to improve on our results by retaining in the training set only those documents that belong to a single subclass,[vi] yielding 30,329 training examples overall. The results are shown in Table 7, in comparison with earlier results. Filtering the training set in this way improves the Naïve Bayes, SVM, and SNoW precision by 1-2 percent. By contrast, the k-

NN algorithm performs slightly worse with this approach. Since the SVM algorithm relies on outlying documents to separate categories, it benefits from the better category separation we impose. By contrast, the k-NN algorithm relies on dense training examples for accurate categorization, and thus performs worse when the training examples are reduced in this way.

Indexing fields	Evaluation measure	Naïve Bayes	k-NN	SVM	SNoW
(c) First 300 words	(1) Top-prediction	33%	39%	41%	36%
(d) Abstracts	(1) Top-prediction	28%	26%	34%	36%
(c) First 300 words	(2) Three-guesses	53%	62%	59%	56%
(d) Abstracts	(2) Three-guesses	47%	45%	52%	58%
(c) First 300 words	(3) All-categories	41%	46%	48%	43%
(d) Abstracts	(3) All-categories	35%	32%	41%	44%

Table 6: Categorization precision at IPC subclass level, for categorization of all test documents. Results for SVM are obtained when indexing the first 150 words of each document, rather than the first 300 words. In addition, the SVM algorithm uses a smaller vocabulary and is limited to 100 training documents per subclass.

Indexing fields	Evaluation measure	Training set	Naïve Bayes	k-NN	SVM	SNoW
(c) First 300 words	(1) Top-prediction	All	33%	39%	41%	36%
(c) First 300 words	(1) Top-prediction	Single subclass	34%	38%	42%	37%
(c) First 300 words	(2) Three-guesses	All	53%	62%	59%	56%
(c) First 300 words	(2) Three-guesses	Single subclass	53%	61%	59%	57%
(c) First 300 words	(3) All-categories	All	41%	46%	48%	43%
(c) First 300 words	(3) All-categories	Single subclass	43%	46%	50%	45%

Table 7: Comparison between categorization precision at sub-class level when training with all documents and when training with single-subclass documents. Results for SVM are obtained when indexing the first 150 words of each document, rather than the first 300 words. In addition, the SVM algorithm uses a smaller vocabulary and is limited to 100 training documents per subclass.

As the SVM algorithm performs better with single-subclass documents, we now train an SVM categorizer with all such documents in WIPO-alpha. We obtain the following accuracies: 46% (top-prediction), 65% (three-guesses), and 54% (all-categories), when all test documents are classified, as shown in Figure 5. These numbers improve on those shown in Table 7 by making use of all single-subclass training examples and represent our best results at IPC subclass level. We can compare them with those published by Krier and Zaccà (2002). They obtained an accuracy of 61% at subclass level with the all-categories measure when using a commercial product implementing a k-NN algorithm. This result is obtained with a training set containing 68,416 different documents (Koster, 2001), 47% more than available in WIPO-alpha. This comparison indicates that our training set should be extended for better results.

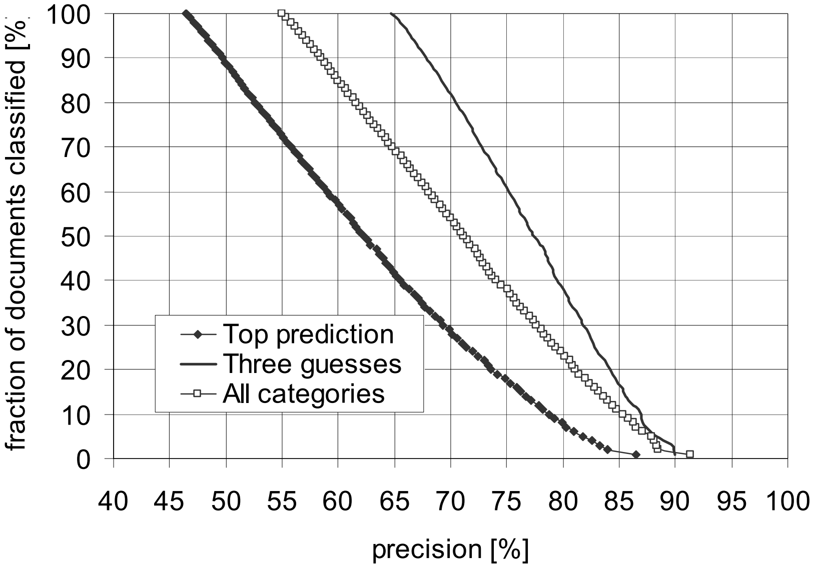


Figure 5: Variation of categorization precision with the fraction of documents classified for SVM subclass-level categorization using the first 150 words following training with all single subclass documents. Results are reported for the three measures of categorization success.

5 Conclusions and outlook

Patent categorization provides a demanding test scenario for machine learning algorithms because of the nature of the taxonomy and of the documents within. Nevertheless, our results show that automation can help users unfamiliar with all the complexities of the International Patent Classification system.

Overall, we find the support vector machine algorithm outperforms the Naïve Bayes, k-NN, and SNoW algorithms under similar conditions, particularly for categorization at IPC subclass level. However, because it is computationally expensive to train, it may be necessary to reduce the training complexity, by limiting the number of training documents, by term selection, and/or by limiting the length of the documents.

Categorization precisions observed at IPC subclass level are below those found for class-level categorization and are best performed by indexing the first 300 words of each document. Training with single-subclass documents improves results for the Naïve Bayes, SNoW, and SVM algorithms, but does not do so for the k-NN approach. Training with the first 300 words of each document and testing with abstracts provides no advantage over testing with 300 words per document.

Because patent classification at patent-issuing authorities must be performed with excellent precision, the combinations of algorithms and training collections reported here do not appear sufficient for building a fully-automated system for the categorization of all patent applications. Nevertheless, the results reported bear well for the development of a patent categorization assistance system, which would suggest a small number of IPC codes to a user. In view of the accuracies reported above, a production system should probably be able to suggest more than three subclasses for each document. This is expected to provide a large increase in accuracy, at the expense of user inconvenience. A real system should also evaluate the confidence levels of the categories suggested and withhold any predictions if these prove too low. It is WIPO's ambition to provide a semi-automated tool for patent classification in the very near future and work in this direction is ongoing.

We are now working to establish a larger multilingual collection of patent documents to improve on results presented here. This collection should allow comparisons of automated categorization difficulty when classifying documents in different languages into the same taxonomy. It is also expected to cover a larger proportion of the IPC, including less frequently used classes and subclasses, thereby creating more difficult test scenarios. By making our datasets publicly available, we hope to promote further research in this area.

Acknowledgements

The authors would like to thank Antonios Farassopoulos, Patrick Fiévet, and Mikhail Makarov for helpful discussions.

References

S. Adams. Using the International Patent Classification in an online environment, *World Patent Information* 22, 291-300, 2000.

J. Calvert and M. Makarov. The reform of the IPC, *World Patent Information* 23, 133-136, 2001.

A. J. Carlson, C. M. Cumby, J. L. Rosen and D. Roth. SNoW User's Guide, UIUC Tech. Report UIUC-DCS-R-99-210, 1999.

S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Using taxonomy, discriminants, and signatures for navigating in text databases, *proceedings of 23rd VLDB conference*, 1997.

S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies, *VLDB Journal* 7,163-178, 1998.

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.

F. C. Gey, M. Buckland, C. Chen, and R. Larson. Entry Vocabulary--a Technology to Enhance Digital Search, in *Proceedings of the First International Conference on Human Language Technology*, San Diego, pp 91-95, 2001.

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.

K. Kakimoto. Intellectual Property Cooperation Center, personal communication, 2003.

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.

C. H. A. Koster, M. Seutter, and J. Beney. Classifying Patent Applications with Winnow, *Proc. Benelearn 2001 conf.*, Antwerpen, 2001.

T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J., Honkela, V. Paatero, and A. Saarela. Self organization of a massive document collection, *IEEE transactions on neural networks* 11 (3), 574-585, 2000.

M. Krier and F. Zaccà. Automatic categorization applications at the European patent office, *World Patent Information* 24, 187-196, 2002.

L. S. Larkey. Some Issues in the Automatic Classification of U.S. patents, *Working Notes for the Workshop on Learning for Text Categorization, 15th Nat. Conf. on Artif. Intell. (AAAi-98)*, Madison, Wisconsin, 1998.

L. S. Larkey. A Patent Search and Classification System, *Proc. DL-99, 4th ACM Conference on Digital Libraries*, 179-187, 1999.

D. D. Lewis, Y. Yang, T. Rose, F. Li. RCV1: A New Benchmark Collection for Text Categorization Research, to appear in *J. Machine Learning Research*, 2003.

A. K. McCallum (1996) Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering, www.cs.cmu.edu/~mccallum/bow.

H. Smith. Automation of patent classification, *World Patent Information* 24, 269-271, 2002.

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.

WIPO. *International Patent Classification: Guide, Survey of Classes and Summary of Main Groups*, Seventh Edition, Volume 9, World Intellectual Property Organization, Geneva, 1999.

[*] Corresponding author. Email: CJF(at)ELCA.CH

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

[ii] The number of categories that should be retained to maximize the absolute number of correct predictions is a non-trivial quantity that would need repeated categorization experiments to determine exactly.

[iii] Secondary IPC symbols may contain additional classes and subclasses.

[iv] Tests performed with the rainbow NB algorithm showed that the precision was lowered when training was made on the basis of both the main IPC code and all additional IPC codes of each patent document.

[v] For tests at class level, the first 300 words do not include the inventors.

[vi] These documents form a superset of the documents possessing a single IPC code.