Chapter 12
Automated Patent Classification

Karim Benzineb and Jacques Guyot

Abstract Patent classifications are built to set up some order in the growing number and diversity of inventions, and to facilitate patent information searches. The need to automate classification tasks appeared when the growth in the number of patent applications and of classification categories accelerated in a dramatic way. Automated patent classification systems use various elements of patents' content, which they sort out to find the information most typical of each category. Several algorithms from the field of Artificial Intelligence may be used to perform this task, each of them having its own strengths and weaknesses. Their accuracy is generally evaluated by statistical means. Automated patent classification systems may be used for various purposes, from keeping a classification well organized and consistent, to facilitating some specialized tasks such as prior art search. However, many challenges remain in the years to come to build systems which are more accurate and allow classifying documents in more languages.

Abbreviations

AI	Artificial Intelligence
APC	Automated Patent Classification
ECLA	European Patent Classification
EPO	European Patent Office
IPC	International Patent Classification
kNN	*k* Nearest Neighbors
MCD	Master Classification Database
NN	Neural Network
PCT	Patent Cooperation Treaty
SVM	Support Vector Machine
WIPO	World Intellectual Property Organization

K. Benzineb (✉) · J. Guyot
SIMPLE SHIFT, Ruelle du P'tit-Gris 1, 1228 Plan-les-Ouates, Switzerland
e-mail: karim@simple-shift.com

M. Lupu et al. (eds.), *Current Challenges in Patent Information Retrieval*,
The Information Retrieval Series 29,
DOI 10.1007/978-3-642-19231-9_12,

12.1 Introduction

An efficient way to facilitate the retrieval of objects is to arrange them beforehand according to an order, which makes sense for most of the people who will do the searching. A good illustration is the way books are organized on the shelves of a library: they are generally grouped by subject, so that the searcher is easily oriented toward the relevant area. A side advantage of this method is that reaching to any specific book automatically provides you with more information on the same topic.

Such is the purpose of classification. The more information you have to manage, the more structured and detailed your classification should be to allow for easy navigation and precise search.

This chapter is meant to explain how classification supports patent management and retrieval, why it is useful to automate it, and how this may be done. Automated patent classification (hereafter APC) has several objectives, which are defined in Sect. 12.2 below. Two of the most widely used patent classifications, namely IPC and ECLA, are reviewed in Sect. 12.3. The structure and content of patent collections used to build automated classification systems are described in Sect. 12.4. Selected algorithms and tools on which classification software is often developed are explained in Sect. 12.5. Various evaluation approaches of APC are suggested in Sect. 12.6, Use Cases are presented in Sect. 12.7, and the main challenges of APC in the close future are discussed in Sect. 12.8.

12.2 Definition and Objectives of Automated Patent Classification

12.2.1 Definition

Automated patent classification may be defined as the process by which a computer suggests or assigns one or several classification codes to a patent on the basis of the patent's content. This definition implies that several conditions must be satisfied:

- a taxonomy, i.e. a patent classification in which each category is clearly defined and has a unique code, must previously exist;
- a full collection of patents previously classified by humans under that classification must be available to train and test the system;
- the content of the patent to be classified (text and possibly pictures, graphs, etc.) must be in electronic format so as to allow for computer processing.

Although these consequences appear to be trivial, they place a heavy constraint on the very possibility to build an automated classifier because one of the most difficult parts is generally to find or build a training set of patents, which is large and well distributed enough.

12.2.2 Objectives

The overall objective of patent classification is to assign one or several category codes to a patent, which is not categorized yet in a given patent classification system. The objective of automating the process is to make it much faster and more systematic than the human process, thus saving time and costs.

- Faster: Human examiners must read the whole patent text, than browse the classification categories which seem most relevant to them, and finally make one or several choices. The whole process can take up to several hours, while a computer performs the same task in a matter of milliseconds.
- More systematic: Human classification may be subjective because it is based on an individual examiner's judgment (which in turn depends on his/her education, experience and mindset) and because it is entrusted to a large number of examiners (up to several thousands in some Patent Offices).

Automated classification systems tend to make the same choices under the same conditions; this leads to more harmonized results. Beyond this immediate objective, APC has in fact a deeper purpose, which is twofold: it has an organizational mission and it must facilitate search tasks.

12.2.2.1 Organization

The essential process of classification is tagging, i.e. assigning a code to an element, which must be classified. In the case of patent classifications, the number of codes to choose among may be extremely high: The International Patent Classification (IPC) has over 60,000 categories and the European Patent Classification (ECLA) has about 129,000.

Besides, the number of patent applications to classify is also very large and it keeps growing: according to statistics from the World Intellectual Property Organization (WIPO), over 1,850,000 patents applications were filed in 2007, up from about 926,000 in 1985.

APC's organizational mission is to assign to those patent applications a classification code in order to preserve the consistency of an order which was defined by human experts. It puts patents "where they belong". This can be done for new incoming patents, but also backwards on previously categorized patents in order to rearrange them when the classification was modified (this is called "re-classification").

As a side product of this role, APC also allows building so-called "pre-classification" systems: In a large patent organization, an APC system can read an entering patent application and route it to the relevant team of experts who will be in charge of making a decision about its final categorization.

12.2.2.2 Search for Prior Art, Novelty, etc.

A major function of APC is to support patent search. Typical goals of a patent search include prior art (patentability), novelty, validity (legal status), freedom to operate, infringement search, etc.

A major issue in patent search is the size of the search space: There are about 40 million patents in WIPO's Master Classification Database (MCD), and this does not represent all of the world's patents in all languages.

The objective of APC in terms of search support is twofold.

- Reducing the search space: By proposing one or several classification codes for a patent application, APC allows to focus the search on the most relevant patent categories, thus excluding most of the patent search space.
- Allowing for search on the basis of similarity: Since the vast majority of patents were classified manually by human examiners, some patents may not be categorized under the expected codes. In order to extend a search to other categories (e.g. for prior art search), APC allows comparing the content of the patent application with the content of each patent in the training set. It may thus retrieve patents which were not classified in the same category as the patent application, but whose content is very similar to it.

It should be underlined here that APC systems are not only used by patent practitioners such as inventors and patent attorneys. Many organizations use them for other purposes such as technological watch, economic intelligence, etc.

12.2.3 Historical Factors

Classifying patents became necessary because of a fast growth in both the number of patents and the number of patent fields.

Logically, the need to *automate* patent classification resulted from the same factors when they reached a higher scale. In particular, the fast growing number of patent applications mentioned above was a driving factor of research on automated tools.

Additional factors also played a role, in particular the fast-increasing number of human examiners, which was (and still is) leading to classification consistency issues. Besides, the hyper-specialization of patent categories made it impossible to entrust the classification job to "universal experts"; it called for a specialization of the examiners themselves, which in turn provoked a diversification of the classifying methods, criteria—and results.

This situation is further complicated by the fact that some patents are of horizontal nature, i.e. they can or should be classified in several categories. For example, a tobacco humidifier can be linked to industrial processes, to storing processes and to agricultural products. Multi-category classification made it even more complex to categorize and to retrieve similar patents, and called for some mechanical help.

12.3 A Few Words about Patent Classifications

Patent classifications, or taxonomies, generally come in the form of a hierarchy of categories: the top level includes very broad categories of inventions, so the number of top categories is very small. The second level includes more narrow categories, the third level even more precise categories, and so on. Thus as we go down the taxonomy levels, the number of categories grows dramatically.

Two patent taxonomies are briefly considered below: the International Patent Classification (IPC), which is built and maintained by the World Intellectual Property Organization (WIPO), and the European Patent Classification (ECLA), which is built and maintained by the European Patent Office (EPO). Both of them are available online on the respective organization's website.

12.3.1 IPC

The IPC (Edition20090101) is divided into a Core and an Advanced Level; the Core Level goes from Section down to Main Group, with some technical sub-groups. The Advanced Level contains all the sub-groups of the IPC. According to WIPO, "the core level is intended for general information purposes, for example, dissemination of information, and for searching smaller, national patent collections. (...) The advanced level is intended for searching larger, international patent collections."

At the Advanced Level, the IPC has the following tree structure: eight Sections, 129 Classes, 639 Sub-Classes, 7,352 Main Groups and 61,847 Sub-groups.

The sections (top categories) are the following: Section A—Human Necessities; Section B—Performing Operations; Transporting; Section C—Chemistry; Metallurgy; Section D—Textiles; Paper; Section E—Fixed Constructions; Section F—Mechanical Engineering; Lighting; Heating; Weapons; Blasting; Section G—Physics; Section H—Electricity.

This top level illustrates the essential challenge of any patent taxonomy: it has to describe the world, and it must be able to include objects and ideas, which, by definition, were never thought of before. Thus it has to be as general and open as possible. For that reason it does not make much sense to classify patents at the Section level. Even the Class and Sub-Class levels are often considered too wide to be useful for professionals (examiners, patent attorneys, etc.). Therefore automated classification systems are generally required to categorize patents at least at the Main Group level. This means any such system should at least be able to manage over 7,000 categories; it also means the system must support a large number of patent examples for the training phase (see Sect. 12.5), since each category must have at least a few example documents for the system to correctly identify it.

At first glance, this enormous quantity of data makes the field of patent classification particularly fit for computerized statistical processing. However, history, while bringing about the reasons for automating patent classification, also produced complicating factors, which actually hamper the efficiency of computerized processing.

These factors are essentially linked to exceptions to the general classification rules: many patent categories contain one or several notes, which indicate that specific types of inventions should actually be classified somewhere else. For example, the category A23 in the IPC has the following title: “Foods or foodstuffs; their treatment, not covered by other classes”. This initially requires knowing which treatments are “covered by other classes”. But additionally, category A23 includes the following notes:

> “Note(s)
> Attention is drawn to the following places:
> C08B: Polysaccharides, derivatives thereof
> C11: Animal or vegetable oils, fats, fatty substances or waxes
> C12: Biochemistry, beer, spirits, wine, vinegar
> C13: Sugar industry.
> Processes using enzymes or micro-organisms in order to: liberate, separate or purify a pre-existing compound or composition, or to treat textiles or clean solid surfaces of materials are further classified in Sub-Class C12S.”

The very human nature of a patent taxonomy and of its evolution therefore makes it very difficult to define systematic classification rules and build them into a model; categories are best described by the documents they contain. This is why example-based training technologies tend to be favored to build automated patent classifiers.

12.3.2 ECLA

The ECLA taxonomy is worth mentioning in addition to the IPC because it is a kind of extension of the IPC: It is identical to the IPC down to Main Group level, but it is more detailed at Sub-Group level, where it contains 129,200 categories, thus allowing for a finer-grain classification.

While the IPC seems to be more oriented toward the publication of patents, ECLA is rather more focused on supporting patent information search in the context of a patent application. It is extensively used, for example, by the EPO examiners in their daily work. ECLA is also used to classify the PCT’s minimum documentation and other patent-related documents, including utility models.

According to the EPO, 26.2 million documents had an ECLA class in 2005. Combined with the 129,200 categories, this gives a broad idea of the “classification space” which must be managed by any automated classifier.

12.4 Patent Collections

12.4.1 Structure of a Patent

The structure of a patent is important because the precision of an APC system directly depends on the quality of the training data, which in turn means that the content to be provided as training material must be carefully chosen. Although there are many ways of representing the structure of a patent (with more or less information details), the content of most patents is organized in the following way.

- The bibliographic data: the patent ID number, the names of the inventor and the applicant, the title of the patent, and the abstract.
- The claims, in which the applicant explains what the invention is made of and which application fields the patent is sought for.
- The full text, which contains the complete description of the patent.

Other fields may be found, such as the agent's name, priority data, publication and filing languages, etc. It is also frequent, for example in the fields of chemistry or mechanics, to find graphics or other types of illustrations. The fields are generally represented in an XML structure, which may look like this:

```
<record cy="WO" an="SE0001823" pn="WO012189020010329"
dnum="0121890" kind="A1"> [Unique patent number, which can include
the date]

<ipcs ed="7" mc="D21H01120"> [This is the IPC classification number]
<ipc ic="D21H01725"></ipc>
</ipcs>

<ins> [Inventors]
<in>LINDSTRÖM, Tom</in>
<in>GLAD-NORDMARK, Gunborg</in>
<in>RISINGER, Gunnel</in>
<in>LAINE, Janne</in>
</ins>

<pas> [Patent Applicant]
<pa>STFI</pa>
</pas>

<tis> [Title]
<ti xml:lang="EN">METHOD FOR MODIFYING CELLULOSE-BASED
FIBER MATERIAL
</ti>
</tis>
```

```
<abs> [Abstract]
<ab xml:lang="EN">A method for modifying cellulose fibers, which are treated for at least 5 minutes with an aqueous solution of CMC or CMC-derivative (. . . )
</ab>
</abs>

<cls> [Claims]
<cl xml:lang="EN"> Claims
1. Method for modifying cellulose fibers, characterized in that the cellulose fibers
are treated for at least 5 minutes with an aqueous electrolyte-containing solution (. . . )
</cl>
</cls>

<txts> [Full Text Description]
<txt xml:lang="EN">
Method for modifying cellulose-based fiber material
This invention concerns the technical field of paper manufacture, in particular chemical (. . . )
</txt>
</txts>
</record>
```

Our experience showed that most of the time, only a part of this content should be used to feed an automated classifier. First, some data have a higher classifying power: it may be the case, for example, of the inventor's name, because inventors tend to invent in a specific field. The applicant's name is also important because it is often a company with a specific area of expertise (although large companies may apply for inventions in various fields). Second, a field such as the Claims may be of little interest for training purposes because it was deliberately written in a vague style so as to cover the widest possible application area. Words in the Claims section tend to be ambiguous and do not help the classifier to make a decision (our experience showed for example that a classifier often has a higher accuracy when trained on the Abstract than on the Claims section). Third, most automated classifiers (and in any case the classifiers based on the algorithms described in Sect. 12.5 below) are exclusively based on text and cannot make use of any graphic information. This means for example that some categories such as Chemistry or Mechanics, whose descriptions heavily rely on diagrams, are not so well classified by text-only tools. Finally, text-based learning machines can get saturated beyond a given number of words: adding more and more words in each example actually ends up creating noise and confusing the machine, which drives the classification precision down. In fact, we found that it is generally more efficient to train an APC system on a large number of small examples for each category than on a small number of large documents.

Therefore the fields which tend to be preferred (through empirical findings) as training material for APC are essentially the bibliography fields, namely the inventor, applicant, title and abstract. However, it was observed that adding some information from the full text description does improve the classification precision, provided that the full text is truncated (our experience suggested the limit of 350 or 400 different indexed words) so as to avoid the saturation issue.

12.4.2 The Distribution Issues

Creating a classification inherently creates distribution imbalances. In the case of patent classifications, those imbalances are essentially found in the distribution of example documents (patents) across the categories, and in the distribution of words within a patent.

12.4.2.1 Distribution of Example Documents: The Pareto Principle

Building a patent collection with regards to a classification amounts to separating patents according to *external* criteria: patents are not grouped because of intrinsic properties (such as the number of words they would share, for example) but because they address a topic which was defined externally by human experts, with regards to their own "knowledge of the world".

When groups of objects are separated according to external criteria, they tend to show a Pareto-like distribution across the categories, i.e. over a large volume of categories and documents, about 80% of all the documents are classified in about 20% of the categories. This creates a structural issue for any artificial intelligence system, which is based on example-based learning. Most automated patent classifiers belong to this family of tools: they need to be trained on typical examples of each category to be able to correctly identify those categories later on.

If some categories are poorly documented, i.e. they have little or no typical patents to feed the computer with, they are very unlikely to be ever predicted by the system because it will never be able to identify a patent typical of such a category. A distribution reflecting the Pareto Principle means that although 20% of the categories will be well documented, the remaining 80% will share only 20% of the total training set. Inevitably, many of those categories will "disappear" in the training process. Solutions to this issue are considered in Sect. 12.5 below, but this remains one of the core problems of any automated classification system.

12.4.2.2 Distribution of Words in a Patent: Zipf's Law

APC systems depend heavily on the words, which are contained in a patent. Neural network applications, for example, give a weight to each word with regards to each category, depending on whether it appears more frequently in one category than in the other ones. The presence, in the text of a patent, of a large number of words which are heavily weighted in favor of a given category drives the application to assign that category code to the patent.

Since the weighting is roughly performed according to the number of times a given word appears in the examples of a category, the system would work better if most of the words appeared frequently in the training documents: it would be easier for the computer to compare their respective occurrences in each category. The problem is that the number of very frequent words is very small, and most words occur in fact rather rarely. This situation is described by Zipf's law, which is well explained on Wikipedia: "The frequency of any word is inversely proportional to its rank in the frequency table". In other words, "the most frequent word occurs approximately twice as often as the second most frequent word, which occurs twice as often as the fourth most frequent word, etc."[1]

Zipf's law tells us that most of the words encountered by the APC system will actually not occur very frequently, so the system will have to work on rather rare words. Moreover, the combination of Pareto's Law and Zipf's Law suggests keeping in the index words, which only occur a small number of times (e.g. four or five times) because they could be representative of a class, which is poorly documented.

12.4.3 The Language Issues

Automated classification systems are met with two major issues when trained on patents: one is linked to natural language in general, and the other one to the particular tongue used in the patent.

12.4.3.1 Natural Language Issues

The ambiguity of natural languages is a well-documented problem, which has a major impact on information retrieval in general, and on automated classification in particular. We will focus here on the issues specifically linked to classification.

Most APC systems use example-based training, i.e. they "read" the content of the various texts provided as typical examples of a given category in order to correctly identify that category. The system is not able to distinguish the various meanings of ambiguous words (polysemy). As a result, it will consider that it is always the same word, which reduces the quantity of information available to identify specific categories.

Another issue with natural languages is linked to semantics: the fact that a word like "not" or "without", which may reverse the meaning of a sentence, is probably

[1] http://en.wikipedia.org/wiki/Zipf's_law. Accessed 23 Dec 2010.

going to be ignored because it is frequent and thus the invention may be classified in a category of things it does *not* do.

A different kind of problems is also met with the various peculiarities of human languages, such as the so-called "collocations" or "compound words", i.e. expressions which are composed of more than one word and whose collective meaning is different from the added meaning of each individual word. For example, an "electric plug" is very different from a "spark plug". Automated classifiers, when used without any linguistic processing, index each word separately and thus lose this collective meaning.

There are many other difficulties linked to linguistics, such as inflections, agglutination (some languages like German stick together various words to build a new entity), or segmentation (choosing the correct number of ideograms which constitute a word in Asian languages), etc.

The issues described above call for linguistic processing, but this may be a costly improvement because it is different for each language (so a linguistic system must be built for each working language) and it may slow down the program's execution.

There are other language issues, however, which may not be solved by linguistic tools, but more probably by statistic processing. The most important one is probably the growing vocabulary extent in each category: new applicants file new patent applications over time, and each of them uses his/her own vocabulary to describe the invention. The underlying issue here is linked to the compositional nature of human language: by combining different words (synonyms) it is possible to say the same thing in many different ways. Thus the number of words typical to a given category grows over time, and a growing number of those words tend to be found in a growing number of categories.

Overall, it should be stressed that technologies such as neural networks (see Sect. 12.5) and others allow one to represent the global context of a patent, which is an efficient solution to get rid of the ambiguity issue in natural languages. An isolated word may have several meanings, but its frequent association with other non-ambiguous words helps a computer to differentiate the various uses of that word. Additionally, in the specific case of neural networks, if a word is so ambiguous that it may be found often and in very diverse situations, the weight of this word will become so low that the word will eventually be discarded for classification purposes.

12.4.3.2 The Corpus Language Issue

APC systems are trained on previously classified examples to recognize patterns of words, which are typical of a given category. However, a system which is trained on English may only classify new patents in English.

Classification (as well as information retrieval) in foreign languages, and more particularly in Asian languages (above all Chinese, Japanese and Korean), are services which tend to be increasingly required by patent applicants, patent attorneys and many other patent professionals and organizations. However, it is still difficult to find large training sets in these languages, mostly for one or several of the following reasons:

- only a small number of patents were filed in the language considered, both at the national and international levels (some countries have a small number of inventors and they are specialized in a small number of fields);
- a number of patents are available but they did not originally exist in electronic form, so only an image scan is available. In this situation the patents may only be used if a good Optical Character Recognition (OCR) software exists for the language considered;
- large training sets were compiled by a private or public organization, but they are kept private or they are sold at a cost which is prohibitive;
- patents are available in a reasonable quantity and in electronic form but they are not classified under international classifications such as the IPC or ECLA so there is a problem to build the initial training set (this is known as the "bootstrap" problem).

For those countries whose number of filed patent applications is quickly growing, training corpuses will soon be available. For the other ones, should an automated patent classifier be necessary, various solutions may be considered, in particular machine translation.

12.4.3.3 The Time Issue

Patent classifications obviously evolve over time: new categories are added while old ones may become deprecated, and some categories may be merged together or broken down in finer ones. However, time also has a direct effect on the language used in the patents. First, the vocabulary of any given category may change over time; this is in particular the case in the field of computer science, where new technologies and standards frequently drive a terminological evolution. Second, the creation of new categories may increase polysemy, as some existing words (like "cookie") are being re-used in new contexts ("Internet cookie"). Finally, the very definition of a category may evolve over time; this issue is known as the "topic drift". This may happen for example when a traditional field (such as printing) is slowly being changed by the introduction of new technologies (in this case, IT): the application or result of the new patents is still directly linked to printing, but the domain itself becomes much wider.

12.5 State-of-the-Art Technologies

Many algorithms may be used for the purpose of automated classification. Most of them come from the world of artificial intelligence (AI) and have been known for several decades (sometimes more); they were revived over the past decade because the spectacular progression in CPU and RAM capacities allowed one to perform more and more calculations in a decreasing time and at a decreasing cost. A general review of most technologies used in the field of APC can be found in [1].

We will focus here on three algorithms, which are among the most frequently used in the field of automated classification, namely Neural Networks (NN), Support Vector Machine (SVM) and the *k* Nearest Neighbors (*k*NN). Other technologies such as Bayesian algorithms, the Rocchio method or Decision rules are also interesting in specific cases, but a complete review of existing technologies and their merits is out of the scope of this chapter.

12.5.1 Neural Networks

A neural network is a network of individual values, each value representing the weight of a given word with regards to a given category, i.e. it tells how well the word (called a "feature") represents the category. Initially the system reads all the documents provided as "good examples" of each category. It compares all the words of all the training documents for all the categories: words which are found too often are given a low weight, because they have a low "classifying power". Conversely, words which seem to be very typical of a given category are given a higher weight because they are strong discriminators. After the training phase, when the classifier is required to classify a new patent, the patent's content is turned into a set of words and weights, which are then compared to those in the neural network; the system chooses the category for which the weights are maximized.

Neural networks are currently among the best-performing patent classifiers for several reasons:

- they scale up extremely well, i.e. they support a large classification space (defined as the number of features time the number of categories);
- they can be combined, so in a tree hierarchy such as the IPC or ECLA, a large number (over a thousand) of neural networks may be built and connected at each level, thus allowing users to ask for a direct classification at any level of the tree;
- after the training phase, the resulting neural networks can be saved for later querying, so the system can reply extremely quickly to users;
- neural networks are trained on strings of characters, so they can be used to classify any type of symbols (e.g. any language, but they can also be used to classify DNA strings, etc.).

12.5.2 SVM

The Support Vector Machine, like the Neural Networks, is a system which has to be trained beforehand, but unlike the NN, it does not assign weights to words; the words are considered as dimensions of a space, and each example of a category is considered as a point is this space. SVM tries to find the plane surface, which separates two categories with a gap as wide as possible.

SVM is interesting to use because it is very accurate: in fact its capacity to build separations between categories is higher than that of the NN. Besides, and unlike the NN, it has no internal configuration parameters, so it is easier to use. However, it does not seem to be widely used to classify patents at the lowest levels of large classifications such as the IPC or ECLA, probably because it only supports a small combination of words and categories, and it is very slow to train.

12.5.3 kNN

The *k* Nearest Neighbor algorithm compares a document to be classified to a number of other, previously classified documents (*k* stands for the number of documents to be compared). Similarities between documents are computed by comparing word distributions. The category of the new document is calculated with regards to the categories of the neighboring documents by weighting their contributions according to their distance; thus it is also a geometric measure.

Unlike the NN and SVM algorithms, the *k*NN does not have to be trained before being used: all the calculations are performed when a document is submitted to the system. For this reason, it is generally not used to predict categories within large classifications such as the IPC or ECLA, because it is considered too slow to reply.

On the other hand, it is very useful for prior art search because it can systematically compare the patent application with the existing patents, retrieve the closest ones and show them to the user.

12.6 Evaluating Automated Patent Classification

The accuracy of an APC system is calculated over a test set which was held out from the training set, so the classifier has to categorize patents it has never seen before.

In a neural network system, for example, the classifier is initially trained over patents which were previously classified by human experts. Then a test set is submitted to the classifier, which predicts one or several categories for each test patent. Finally those predicted categories are compared to the correct classes (which are also known since the test set was also previously classified by humans) and an accuracy score is automatically calculated.

12.6.1 Standard Evaluation Methods: Precision, Recall, F1

In the general field of information retrieval, accuracy scores are often calculated according to one or several of the following three standard methods: a precision score, a recall score and a so-called “F1” score, which is an average of precision

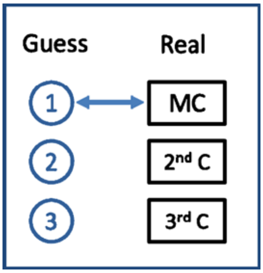

Fig. 12.1 Top prediction

and recall. Those methods are also valid to assess the accuracy of an automated patent classifier.

According to Wikipedia, “Precision can be seen as a measure of exactness or fidelity, whereas Recall is a measure of completeness.”[2] More specifically, in the field of patent classification, for a given patent submitted to the automated classifier:

- Precision is the number of categories correctly predicted by the classifier divided by the total number of predicted categories;
- Recall is the number of categories correctly predicted by the classifier divided by the total number of existing correct categories (i.e. the number of categories which should have been retrieved).

As for the F1 score (also called F-score or F-measure), it is a weighted average (more precisely the harmonic mean) of precision and recall:

$$F1 = 2 \bullet \frac{precision \bullet recall}{precision + recall}.$$

Several variations of the F1 score can be used to place more emphasis on precision or on recall. Choosing the most relevant measure directly depends on the intended use of the search engine or the classifier: sometimes only precision or recall may be looked at. For example, a patent classifier which assigns one or several correct category codes to an incoming patent application may be considered good enough, although it may not have assigned *all* the correct categories. In such a case, only the precision score may be taken into account.

12.6.2 Customized Use of Accuracy Measures

Whether the type of measure is precision, recall or F1, there are still various ways of calculating it. The most common way is to determine whether the top category predicted by the classifier corresponds to the first real category of the patent. This is all the more useful when the classification includes a main category (called “MC” in Fig. 12.1) and several secondary categories:

[2] http://en.wikipedia.org/wiki/Precision_and_recall. Accessed 23 Dec 2010.

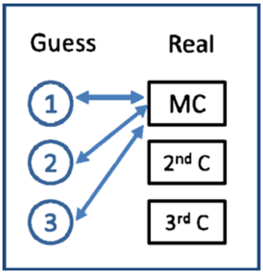

Fig. 12.2 Three Guesses measure

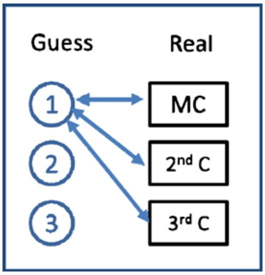

Fig. 12.3 All Categories method

This accuracy assessment makes sense for fully automated solutions, for example when an APC system receives electronic patent application files and routes them to the most relevant team of human examiners (this application is called "pre-classification"). In this case only the best choice is kept because the system must only route the file to a single team, so the accuracy score should reflect this constraint.

However, other methods of calculating the accuracy may be chosen, for example when the APC system is used by a human expert as a classification assistant. In this case, more than one guess may be accepted by the user, since he/she will make the final decision. On the other hand, in the example of the pre-classification task described above, only one good answer is possible, because only one team is the most relevant to classify the patent. In this situation we can use for instance a "Three Guesses" measure as shown in Fig. 12.2.

In this measure, if the most relevant category appears in one of the top three predictions, the patent is considered to be correctly classified.

Other types of applications can be considered (see Sect. 12.7 below); for example, patents can generally be attributed to multiple categories, so a human examiner who would use the APC as a classification assistant may be willing to accept any good suggestion, even if it is not the main category or if there is no main category (which is the case under ECLA). In this situation the accuracy measure would be calculated according to an "All Categories" method as illustrated in Fig. 12.3.

This last method of calculating the accuracy produces of course the highest scores, since it is more flexible as to what constitutes a correct prediction. It is for example the method used by WIPO to assess IPCCAT's precision, because IPCCAT is a classification assistant. Its typical precision scores for English patents are about 90% at Class level, 85% at Sub-Class level and 75% at Main Group level.

It should be mentioned that the methods proposed above address the situation where the APC system provides so-called “unsupervised” predictions, i.e. it has no human help such as limiting the classification space to a given section, or refining to a finer category after validating a coarser one. More interactive systems, where the APC system can be guided by a human user, may provide for more accurate predictions, especially at the more detailed levels.

12.7 Use Cases

There is a wide range of possible applications for APC systems. The ones that are suggested below have been actually implemented (interactive classification and prior art search) or are being considered (pre-classification and re-classification) in particular by international organizations.

12.7.1 Pre-classification

Pre-classification, as mentioned earlier, is the task of automating the distribution of incoming patent applications among the various possible groups of examiners. Large patent offices have organized their teams of examiners according to fields of expertise. When a new patent application reaches the patent office, it must be routed to the most relevant group of experts. This can be automated with an APC system.

Such an application has generally an excellent accuracy performance because the number of groups of experts is generally less than 100 (i.e. much less than the number of patent classification categories, for example). However, since the system operates without any human supervision, it is generally required to compute a confidence score for each decision, and the automated routing is only allowed when the confidence score is above a pre-defined threshold.

12.7.2 Interactive Classification

APC systems can be used as classification assistants for human examiners in an interactive manner: the examiner submits a patent application, the APC system makes one or several predictions at a given level of the classification, and then the examiner can decide to:

- ask for a refined prediction down to a finer-grain level of the classification (for instance at Main Group level after an initial prediction at Sub-Class level);
- ask for another prediction directly at the finer-grain level (e.g. a prediction directly at Main Group level, instead of going first to Sub-Class level—it is important

to understand that in the case of neural network systems, the APC's prediction will not be the same when predicting directly at the lower level and when going through an intermediate level because the APC does not use the same neural networks);
- force a prediction under a given category, for instance by defining that he/she wants only predictions under the A01B Sub-Class.

The World Intellectual Property Organization (WIPO) proposes such an interactive APC tool on its website; the tool is called IPCCAT and it is freely available to the public.

12.7.3 Re-classification

In large patent classifications such as the IPC or ECLA, some categories grow over time up to the point where they contain too many patents of various content. In that case they must be broken down into several more detailed categories (at the same level). Conversely, some categories may end up with very few or no patents, either because they are too narrow or specialized, or because they were defined through a rather theoretical process.

Thus patent classifications must be re-organized from time to time; for instance, the current version of the IPC is the ninth one, and new versions might be published from now on at least on an annual basis.

Re-classification is the process by which patent categories are grouped together in larger ones, or broken down in smaller ones, as well as the subsequent process of re-tagging the patents which were classified under the modified categories. This last process, in particular, may be extremely time-consuming and costly to implement.

APC systems can support the re-classification process by

- suggesting new (larger or smaller) categories, in particular through the use of clustering technologies (algorithms such as K-Mean, etc.);
- automatically re-tagging the patents according to the new patent categories.

The major issue in re-tagging patents is to build a training corpus, since there is no existing set of patents with the correct new categories to train the APC system. When the number of modified categories is not too important, the solution is generally to feed manually the new categories with typical examples; most often, 10 or 20 examples may be sufficient for the system to correctly identify the new category and automatically re-classify the rest of the patent collection.

12.7.4 Prior Art Search

APC systems are extremely useful to assist patent examiners in their prior art search. From a collection of several million patents, an algorithm like the kNN can retrieve, in a matter of a few seconds, the ten patents which are closest to a submitted patent application. As underlined earlier, this tool is all the more interesting because it does not depend on classification codes: it browses the entire patent collection to find similar documents.

The IPCCAT application mentioned above, which is hosted on WIPO's website, provides for such a prior art search tool.

12.7.4.1 Non-Patent Documents

It should be mentioned here that the classification and prior art search tools described above can be applied to all kinds of documents, not only patents. Technical literature, for instance, is a good playground: patent attorneys are eager to find the literature, which relates to a patent application, and a tool such as the *k*NN can be very efficient in this context.

Automated classification systems also have a bright future in the field of web mining: the search for novelty, for example, requires one to browse large volumes of documents on the Web. Classifying them automatically and finding the closest documents to a patent application may save considerable time and work.

12.8 Main Issues

12.8.1 Accuracy

Improving the accuracy of APC systems is the essential challenge today. Although these systems tend to be currently used as classification assistants, the ultimate goal for researchers in the field is to provide fully unsupervised systems, for instance to build pre-classification tools or to classify large volumes of patents in batch mode.

Several research tracks are being considered to improve the accuracy of APC systems.

- *More training data*: Adding training examples is one of the most immediate solutions. This process should essentially target the categories where samples are scarce. If no additional examples are available for a given category, it can be considered to make several copies of the available documents; this technique is called "oversampling". It allows at least a poorly represented category to "exist" in the system's classification space. Another approach is to find non-patent documents (such as technical literature) strictly related to the category's topic in order to add words, which are typical of that topic. The underlying problem here is not that patent collections, classified by human experts, are not available, but that they are not *readily* available, i.e. they are generally privately owned and not available on commercial terms, or at relatively high prices.

- *Better training data:* The patents provided as training examples must be accurately classified under the most recent version of the classification. They must also be recent: old patents may belong to categories, which no longer exist (if they were not re-classified) and thus may blur the information provided by other examples. Two techniques may be used in particular to improve the quality of the training set: the first is using a time-window which will be moved over time so as to keep only patents which were granted over the last, say, 15 years. Thus every year the latest patents are added, the oldest ones are removed, and the system is re-trained. The second one is to use a so-called "validity file": this file defines all the categories, which are valid in the latest version of the classification. Before using a training set, all the categories assigned to its patent examples are compared to the validity file, and the examples whose categories are no longer valid are removed. This eliminates noise in the training set.
- *Building a "Committee of Experts":* It was described above that not all sections of a patent are used to train the classifier. The situation is in fact somewhat more complex: some categories (such as, for instance, chemistry or electronics) may be best described by specific sections (such as the title or the inventor and applicant names), while other categories could be better characterized by other sections (such as the abstract). This is partly because some inventors or applicants are very specialized, and some fields use very specific words while others (e.g. for more general or conceptual inventions) are described with a broader vocabulary and thus need more information to be specified. One possible solution is to build a so-called "Committee of Experts": one technology (for example neural networks) is used to create a large number of classifiers, each of them being built on different patent sections and being tested against all the classification categories. Then another technology (in this example, it would be an SVM machine) is used to assess which classifier is more fit to which category. Let us imagine that the best classifier for a given category of electronic devices is the one based on the inventor's and applicant's names, while for some agricultural devices it is the one using the abstract and first 350 words of the full text description. When a patent application is submitted, all the neural networks will be required to make predictions, but the SVM machine will favor the answer of the one which it found best fit to the specific context of that application.
- *Using linguistic processing:* In order to gain classification accuracy, the general need is to add information to help the APC system to draw clearer separations between each category.

Linguistics can help. An important step may be to disambiguate the vocabulary by using so-called *word sense disambiguation*. This may be done by using the context of the word in combination with a semantic network (a so-called *ontology*) to help the system discriminate between several possible senses of a word.

The use of *collocations* and *compound words* also help to discriminate between concepts. A special program looks at the co-occurrence of all the words in all the training examples, and if some words occur together very frequently they are automatically considered a single entity. In the example given above, "spark plug" would be processed as one term. In recent experiments, using collocations allowed an NN-based APC to improve its precision score by up to 5%.

If there are too many words, an efficient solution to reduce the number of words is to use *stemming*: only the radical part of the word is index, which allows one to group several words with a common beginning but different endings due to plural or feminine forms, conjugated verbs, etc. The efficiency of stemming depends on the algorithm chosen: in the case of neural networks, stemming actually proved counter-productive because it tended to blur the concepts behind the words.

Another technique called *n-gram processing* is useful for some languages with specific issues.

- German, for example, is an agglutinative language, i.e. it can stick together several words when they are used together. This creates a "new word" for the indexer, so it may be important to find back the compounding words in order to limit the size of the index and get more examples of the words considered.
- Chinese and other languages which use ideograms (symbols) instead of letters are difficult to segment into words: 2, 3, 4 or more ideograms can compose a word, so specific rules may be used to solve this issue.

N-gram processing is a technique by which the first n letters (2, 3, 4 or more) which are read by the indexer are built into a word, then the following 2, 3, 4 letters, starting from the second character, are built into another word, and so on. The system stops when a blank space is met. A 2-letter n-gram rule is called a bi-gram, a 3-letter rule is a 3-gram, etc. This technique allows us to keep the words which were found both independently, and within a larger string, thus enriching the information provided by the training examples.

All the techniques proposed above can be used in combination in order to maximize the chances to classify more accurately. For example, it may be a good idea to test various linguistic processing techniques and to add the best-performing ones when building a Committee of Experts.

12.8.2 Scalability

Most APC tools currently classify patents down to Main Group level with a reasonable accuracy level. A common request from users is now to classify at Sub-group level, which means an increase in the number of possible categories by about a factor 10. It also implies to work with much larger training sets since examples are needed for each individual category.

A tool such as the kNN has intrinsic scalability issues, which tend to be more often studied in the field of search engines because its scalability depends on the size of the corpus (not of the classification). As mentioned earlier, from the point of view of APC it is too slow because it has no training phase so it cannot be prepared in advance.

As for the scalability of SVM, as far as we know, this issue is not solved yet because what would be needed is a highly efficient and robust implementation of a parallelized architecture—which does not seem to have been implemented so far. Some interesting research work is being done in this field, in particular by implementing SVM on Graphical Processing Units (GPUs), which have hundreds of processors.

For systems based on neural networks, the size of the neural networks to be built is the product of the number of words and of categories. The problem is not linked to the calculation power because it is possible to use many processors simultaneously for the training phase. However, to be efficient, the neural networks must be stored in RAM memory. Thus the main limit to NN systems currently lies with the RAM capacity. For example, in order to build an APC system over 70,000 categories and 3 million words the system would need about 256 Gb of RAM. This type of architecture is in fact available but it is still costly.

12.9 Conclusion

Artificial intelligence is still largely perceived as a "magical" resource and expectations tend to be excessive with regards to its real capacities and to the services it can provide. It is not possible for AI systems to classify with regards to *any* sort of classification; the classification needs to make some sort of sense to the APC system, i.e. the system must be able to draw clear limits between the categories.

Additionally, classifying on words has intrinsic limits: it is not always possible to define or represent categories with words. For example, it would probably be very difficult to describe with patent examples the category of inventions, which are easy or difficult to implement, or the category of inventions, which are profitable, or not.

The good news is that, so far, most categories of the IPC or ECLA are well represented and recognized, with the notable exception of the chemistry field, for which specific tools making use of graphics have now been developed.

Clearly today, the main challenge is to improve the accuracy of APC systems at the lowest levels of patent classifications. This will essentially be achieved by adding information, but not just *any* information because this can be counterproductive. Another promising approach will be to specialize the APC systems according to the various patent fields, for instance by choosing the most appropriate technologies for each particular family of topics. Besides, in addition to more powerful technologies and equipment, APC systems will also require larger and better data sets to be trained, tested and improved.

When all these conditions are gathered, APC applications can indeed become very effective and useful tools, both as assistants to human experts and as independent tools, and both for pure organizational tasks and for information retrieval purposes. For that reason their use is most probably going to expand fast in the future.

References

1. Sebastiani F (2002) Machine learning in automated text categorization. ACM Comput Surv 34(1):1–47

Further Reading

2. WIPO's website page dedicated to international patent classifications (IPC, Nice, Locarno, Vienna): http://www.wipo.int/classifications/en/. Accessed 23 Dec 2010
3. EPO's website page dedicated to ECLA: http://test.espacenet.com/ep/en/helpv3/ecla.html. Accessed 23 Dec 2010
4. World Patent Information, Elsevier: an International Journal for Industrial Property Documentation, Information, Classification and Statistics (Quarterly)
5. Berry MW, Castellanos M (eds) (2007) Survey of text mining: clustering, classification, and retrieval. Springer, Berlin
6. Fall CJ, Törcsvári A, Benzineb K, Karetka G (2003) Automated categorization in the international patent classification. SIGIR Forum 37(1)
7. Fall CJ, Benzineb K, Guyot J, Törcsvéri A, Fiévet P (2003) Computer-assisted categorization of patent documents in the international patent classification. In: Proceedings of the international chemical information conference (ICIC'03), Nîmes, France, Oct 2003
8. Proceedings of the CLEF-IP 2010 (classification task), to be published in 2011. The related web site is here: http://www.ir-facility.org/research/evaluation/clef-ip-10. Accessed 23 Dec 2010