International Journal of Computational Intelligence Systems, Vol. 8, No. 3 (2015) 502-516

A Supervised Requirement-oriented Patent Classification Scheme Based on the Combination of Metadata and Citation Information

Fujin Zhu
School of Management and Economics, Beijing Institute of Technology，Beijing, 100081, China
E-mail: fujinzhu.bit@gmail.com

Xuefeng Wang*
School of Management and Economics, Beijing Institute of Technology，Beijing, 100081, China
E-mail: wxf5122@bit.edu.cn

Donghua Zhu
School of Management and Economics, Beijing Institute of Technology，Beijing, 100081, China
E-mail: zhudh111@bit.edu.cn

Yuqin Liu
Beijing Academy of Printing and Packaging Industrial Technology, Beijing Institute of Graphic Communication, Beijing, 102600, China
E-mail: liuyuqin2004@126.com

Received 9 October 2014

Accepted 1 February 2015

Abstract

Patent classification systems are applied extensively in innovative analysis. Existing patent classification schemes are either technology-dependent or TRIZ-based. The former ones, such as the IPC and UPC, are normally developed by different patent offices in the world mainly for the purpose of patentability examination and patent retrieval, while the latter is for TRIZ users and analysts with no more than 40 categories. These static classifications are too complex and general to meet the in-depth patent classification requirements of a specific technology area or organization. To tackle these drawbacks, in this paper, we propose an automatic requirement-oriented patent classification scheme as a complementary method using supervised machine learning techniques to classify patent dataset into a user-defined taxonomy. The requirement-oriented patent taxonomy can be technology-dependent, application-dependent or a mixture of both tailored to specific business objectives. It is more comprehensible and adaptable to various patent management requirements. Through a set of experiments on a collection of 14,414 patents in a case study in the technology area of system on a chip (SoC), we recommend using the combination of the metadata and citation information as the document representation for the new method since it can obtain relatively high classification accuracy with a dramatically simplified document preprocessing process.

Keywords: Patent classification, Requirement-oriented taxonomy, Document representation, Machine learning

* Corresponding author. E-mail: wxf5122@bit.edu.cn

1. Introduction

Patents are a strategic source for knowledge management because the huge number of technical information contained and their uniqueness can be used for many different research and development (R&D) activities: new product development, technology transfer, problem solving [1], technology evolutionary pathways identification [2], technology forecasting, technological opportunity analysis [3] and mergers and technology acquisitions (M&A) analysis [4], etc. With the rapid developments in various technology areas, the number of patents has increased dramatically in recent years. How to manage and make use of the constantly growing volume of patents is becoming an important issue. In order to obtain high quality patent information to support science and technology management, such a great volume of patent documents need to be classified into some predefined taxonomies. A well-established patent classification system can not only improve the efficiency of patent searching, but also help to analyze the patent technology distribution, identify core patents, monitor competitors, and identify potential rivals and new technologies, assisting to make competition strategy for companies and concerned government departments[5].

For the sake of patentability examination and retrieval, different patent offices in the world have developed a variety of patent classification code systems. Among these patent classification systems, the International Patent Classification (IPC) system maintained by the World Intellectual Property Organization (WIPO) is the most popular. However, these classification systems have three main drawbacks for a domain or organization's specific patent analysis task: First, they are too complicated and difficult to comprehend. Take IPC as an example, the complex hierarchical classification system comprises 8 sections, 128 classes, 648 sub-classes, about 7200 main groups, and approximately 72,000 subgroups. A typical category might be "D05C1/06", which refers to Section D (Textiles; Paper), Class D05 (Sewing; Embroidering; Tufting), Subclass D05C (Embroidering; Tufting), Group D05C 1/00 (Apparatus, devices, or tools for hand embroidering) and Subgroup D05C 1/06 (Needles specially adapted for hand embroidering; Holders for needles or threads). In the IPC classification system, each patent is labeled by at least one IPC code. To figure out the technical category of a patent, patent examiners or analysts need to search the technical explanations corresponding to its IPC codes. Due to the complexity of the IPC system, this process is really difficult and time consuming. In addition, many fuzzy terminologies exist in these classification systems, making it more difficult to understand. Second, they are too general to adequately represent detailed technology information for patent intelligence analysis and patent management for a focused technology area or a single organization. Take the patent analysis on computer operating system as an example: no existing classification system contains a category of "computer operating system". What is more, analysts hold different opinions on the definition of the technical scope of computer operating system. The result of the analysis by these classifications is insufficient to reflect the technological niche of a company and mis-categorization results in further difficulties for patent management. Third, these patent classification systems are static in a short-term period, which means they do not evolve as quickly as the development of technology[6]. Therefore, a requirement-oriented patent classification or a personalized knowledge organization system (KOS)[7] on a focused domain need to be explored to represent the technology information in an agile manner with more details according to the specific patent management and analysis requirement. In this paper, we propose a supervised patent classification scheme tailored to the business requirements of a specific technology area or organization to conduct patent analysis and management. The requirement-oriented taxonomies are grouped into technical taxonomy, application taxonomy, and application-technical mixed taxonomy. The requirement-oriented taxonomy is user-defined, aiming to a specific technology area or organization, which means it's more comprehensible and adaptable to various patent analysis requirements. In the case study and experiments, comparisons of classification performances using different features are conducted to figure out the most effective document representation for the new method. Compared with classification performances based on narrative text, metadata, citation information or their different combinations, we got a conclusion that using the combination of patents' metadata and citation information can obtain relatively high classification accuracy with a dramatically simplified document preprocessing process.

The main contributions of this paper are: (i) it proposes a patent classification scheme aiming to a specific technology area or organization, which overcomes the flaws of conventional patent classification systems and is more comprehensible and adaptable to various patent management requirements; (ii) it illustrates the detail process of the scheme using supervised machine learning techniques; (iii) it concludes the most effective document representation for this scheme through a set of experiments.

The paper is organized as follows: In Section 2, we review previous researches on patent classification and address the significance of our study. In Section 3, we describe the requirement-oriented patent classification scheme proposed in the paper and the automatic classification process in detail. The issue of multi-label classification is also discussed in this part. In Section 4, a case study on the SoC industry is conducted according to the automatic classification process, and classification performances using different machine learning techniques based on various document representations are compared. In Section 5, we draw our conclusions and present our future study directions.

2. Literature Review

In recent decades, a number of researchers have reported their works on automatic patent classification. In this section, we give an overview of existing classification schemes and explain why it is necessary to propose an automatic requirement-oriented patent classification scheme. Researches on automatic patent classification leveraging machine learning techniques based on different document representations will then be discussed.

2.1. *Patent classification scheme*

Existing patent classification schemes can be grouped into two categories: (i) technology-dependent schemes such as the IPC code or the UPC code for patent application, examination and retrieval; (ii) TRIZ-based scheme that classifies patents according to the Contradictions and the Inventive Principles for the demand of TRIZ users.

The first category of patent classification schemes are mostly developed by different patent offices in the world. The main purpose of these classifications is to briefly describe the invention granted in a patent for the sake of patent application, examination and retrieval. Besides the IPC, they include the European Classification system (ECLA) by European Patent Office (EPO), the United States Patent Classification (UPC) by United States Patent and Trademark Office (USPTO), the F-Index and F-term (FI/F-term) by Japanese Patent Office (JPO). Also, some commercial patent database providers design their own special patent classification systems. In the Derwent Patent Classification system as an example, the system divides technology areas into three classes: Chemical, Engineering, Electronic and Electrical. Classes are subdivided into sections; sections are further subdivided into subclasses. Among these patent classification systems, IPC is the most popular. While in order to standardize the classification systems of all major patent offices, the new classification system Cooperative Patent Classification (CPC)[1] is introduced since January 2013.

In contrast to official patent classifications for patentability examination and retrieval, the inventors using TRIZ[8], which is the Russian acronym for the Theory of Inventive Problem Solving developed by Genrich Altshuller in Russia in 1965, are not only interested in searching for inventions in related fields (or prior art), but also analogous problems in other fields that have previously solved the same Contradiction. By referring to how analogous patents have used the inventive principles summarized by Altshuller to solve the same contradiction, the inventors could be oriented towards the most effective solutions directly[9]. Therefore, to facilitate searching patents for TRIZ users, patents are required to be classified according to the Contradictions and the Inventive Principles. Loh et al[9] , He and Loh[10] proposed a TRIZ-based patent classification system for the TRIZ users who are in need of patents identified and clustered by the Contradiction addressed and the solutions (Inventive Principles) to the Contradiction. In their system, the original 40 Inventive Principles were grouped into 22 new classes by text and meaning similarity. Automatic patent classification is then performed and evaluated according to the new classes.

In general, in these classification schemes, patents are classified into a static taxonomy, either a general technology hierarchy of all technological fields or the 40 TRIZ inventive principles. However, as discussed in the introduction, they are no longer effective when an organization want to classify a set of patent documents to a user-defined taxonomy specific to various business objectives of a specific technology area or organization,

such as in-depth patent classification and patent landscape analysis[11].

2.2. *Automatic patent classification*

Automatic patent classification may be defined as the process by which a computer suggests or assigns one or several classification codes to patent on the basis of the patent's content. Benzineb and Guyot [12] conducted a comprehensive introduction of the objectives, some widely used patent classifications, the structure and content of patent collections used, selected algorithms and tools, evaluation approaches, use cases and future challenges of automatic patent classification.

Automatic patent classification based on supervised and unsupervised machine learning algorithms has been studied for over a decade[12-16]. A range of machine learning algorithms for patent classification have also been proposed for different languages such as English, Japanese and German[14, 17]. In particular, Larkey[12] created a web-based system for the retrieval and classification of patent text system to classify US patents into UPC. Krier and Zacca[13] reported their research on automatic categorization applications at the EPO. Fall et al[14] published their results of automatic classification in the IPC. Moreover, a series of tracks and workshops have focused on patent retrieval and automatic classification, mainly including the CLEF-IP track [a], TREC-CHEM track [b] and NTCIR workshop [c]. Some well-known international academic conferences such as CIKM and ACM SIGKDD have also workshops on automatic patent classification. A number of large clean datasets of patent documents and benchmarking methods are provided for researchers around the world. Different methods and algorithms on the issue of automatic patent classification are proposed in these platforms.

However, a common drawback of these previous works is that they try to classify patents into an official patent classification code system such as IPC, UPC or ECLA, which are too complex and general to meet the in-depth patent classification requirements of a specific technology area or organization and thus do not necessarily provide direct insight into the technical or market niche.

As all machine learning algorithms need a formal representation of the document, a feature selection phase is essential before classification. In this phase, some terms of a document are selected to build a feature space on which the classification algorithms can work. This stage is an important initial step which can affect not only the learning process but also the efficiency of classification[18]. A better performance is often achieved using features derived from the original input. Building a feature representation is an opportunity to incorporate domain knowledge into the data and can be very application specific[19].

Unlike other text documents, a patent is a relatively structural document consisting of several parts. Take a US invention patent as an example, some parts provide metadata information such as patent number, date of application and UPC code. Some parts contain citation information, such as citing patents and cited patent number. Other parts are narrative text providing information regarding the patent and are given under the headings: Title, Abstract, Field, Background, Detailed description, and Claims.

There are many ways to represent the whole patent document in previous patent automatic classification and retrieval applications[15, 20, 21]. Some researchers believe that the human generated abstracts of patent documents are very precise and the most important section for patent classification. Instead of using the full text of a patent document as the basis for classification, Larkey[12] indexed only the title, abstract and the first 20 lines of the description, and the claims; Koster et al[16] used only abstracts instead of the full-texts of the patents as features; Fall et al[14] separately used (a) the title, (b) the claims, and (c) the first 300 words of full text as the document representation, and found that the best performance is achieved using the last representation. Some have employed patent metadata (e.g., the inventor's name) to achieve improvements in the classification performance[22]. Some take advantage of semantic structural information of patent documents to classify patents. Kim J H and Choi K S[15] conducted patent document categorization based on semantic structural information of patent documents. Other approaches have utilized citation relationships to improve the performance of patent classification.[6, 23, 24, 25]. Li X et al[25] referred to related studies on webpage classification and used the citation network structure information to address the patent classification.

[a] http://www.clef-campaign.org
[b] http://trec.nist.gov/tracks.html
[c] http://research.nii.ac.jp/ntcir/index-en.html

In this paper, we group different fields of a patent document into narrative text, metadata information and citation information. For the sake of patent protection, most descriptions in the narrative text of patents are intentionally obscurely written by their applicants. In our new patent classification scheme, instead of using only the narrative text, such as the title and abstract, as the document representation, the combination of metadata and citation information of the patent is indexed. The result of our experiments demonstrates that this kind of document presentation can obtain relatively high classification accuracy with a dramatically simplified document preprocessing process.

3. A Supervised Requirement-oriented Patent Classification Scheme

In a specific technology area or organization, patents are often needed to be classified according to different requirements of patent analysis and management. To fulfill the practical requirement for classifying patent documents into a predefined requirement-oriented taxonomy, we propose an automatic requirement-oriented patent classification scheme as a complementary method using supervised machine learning techniques.

In this section, we describe different types of requirement-oriented taxonomies firstly. Automatic classification process of patent documents of a specific technology area or organization using supervised machine learning techniques will be presented then. Multi-label classification issue will be discussed as well.

3.1. *Requirement-oriented taxonomy*

When carrying out R&D planning for a specific technology area or organization, patents may need to be categorized by their technical, application, functional attributes, or even possible combinations of them. As a result of various business requirements, requirement-oriented patent taxonomies can be grouped into three categories: technical topic, application topic and application-technical mixed topic patent taxonomy.

3.1.1. *Technical taxonomy*

Technical topic is the most important classification criterion for patent classification. Technical topics of a patent highlight technology innovations or breakthroughs achieved. For example, “de-mapping based on irregular LDPC codes” is a technique for the technical topic of “de-mapping”. Current patent classification schemes, including the first two discussed in section 2, are generally based on technical topic. Especially, the IPC is a typical technical topic classification code system covering all technology areas. When classifying patents to this type of taxonomy, technical topics of the area should be firstly defined by domain experts. A technical taxonomy of navigation technology is defined in Figure 1.

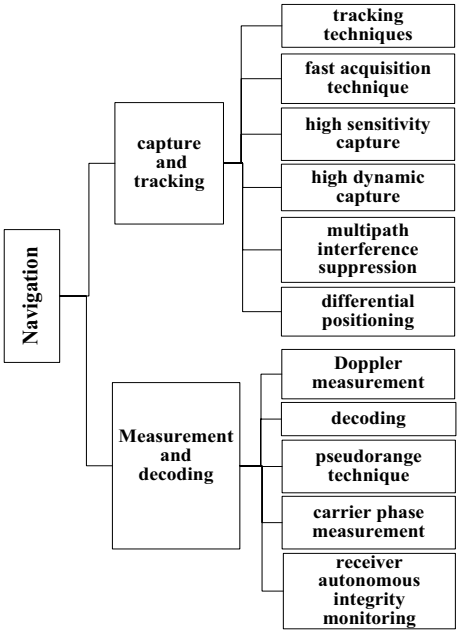

Fig.1. An example of technical taxonomy: a technical taxonomy of the specific technology area of navigation.

3.1.2. *Application taxonomy*

Application topic is another important criterion to conduct patent topic classification. In contrast to technical topic, application topic of a patent focuses on its application areas. Although patent documents, especially patents for invention, are focused on new technologies, patent analysts may be curious about what application areas they can be applied to when doing R&D planning in an organization. For example, a machine learning technique can be used in a set of areas: text classification, image recognition, and so on. Before we conduct patent classification by application topic in a specific technology area, related applications of the technology area should be concluded by domain experts to build a tree-like hierarchy of application topics. An application taxonomy example of computer numerical control (CNC) is shown as Figure 2.

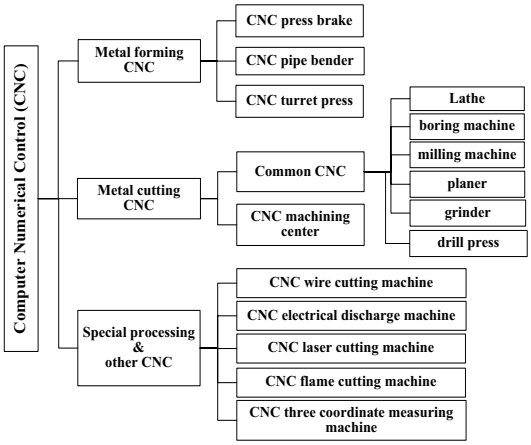

Fig.2. An example of application taxonomy: an application taxonomy of the specific field of computer numerical control (CNC).

3.1.3. *Application-technical mixed taxonomy*

In the above sections, patent classification topics are grouped into technical and application topics. We made two observations from our past requirement-oriented patent classification practices. First, it is sometimes very difficult to distinguish "technical topic" from "application topic": terminologies like "communication" can refer to both communication technology and communication applications. Second, a great number of patent topic classification requirements focus on both of them, analysts care about not only the technical topic, but also application areas of a patent. Given this, the third kind of patent topic taxonomy is an application-technical mixed hierarchy illustrated as Figure 3.

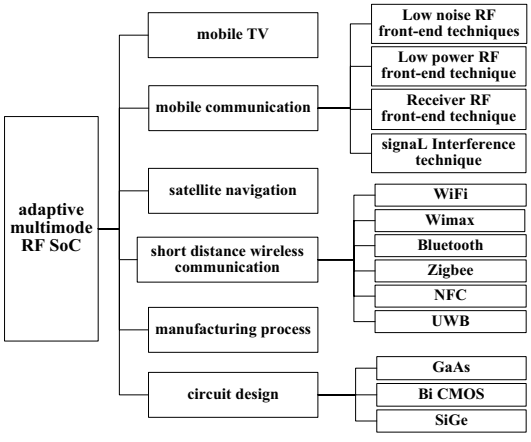

Fig.3. An example of application-technical mixed taxonomy: a classification framework of the specific field of adaptable multimode RF system on chip (SoC).

3.2. *Automatic classification process*

Figure 4 shows the process of the supervised patent classification using machine learning techniques. The process consists of four phases:

- Phase I: Requirement analysis and data collection;
- Phase II: Document preprocessing;
- Phase III: Classifier building;
- Phase IV: Patent classification using the metadata and citation information.

First of all, after the business requirement in a specific technology area or organization is analyzed, relevant patents are then collected and a requirement-oriented patent taxonomy is built. A sample dataset should be classified manually for supervised classifier training and validation later. At the second phase, every patent document are cleaned and processed to construct a feature vector. Supervised machine learning algorithms are introduced to build a satisfactory automatic classifier using the manually classified sample dataset at the following phase. Unlabeled patents are at last classified into the predefined taxonomy by the validated classifier. After the four phases, the task of requirement-oriented patent classification is finished, and patents in the specific technology area or organization are finally classified into the requirement-oriented taxonomy.

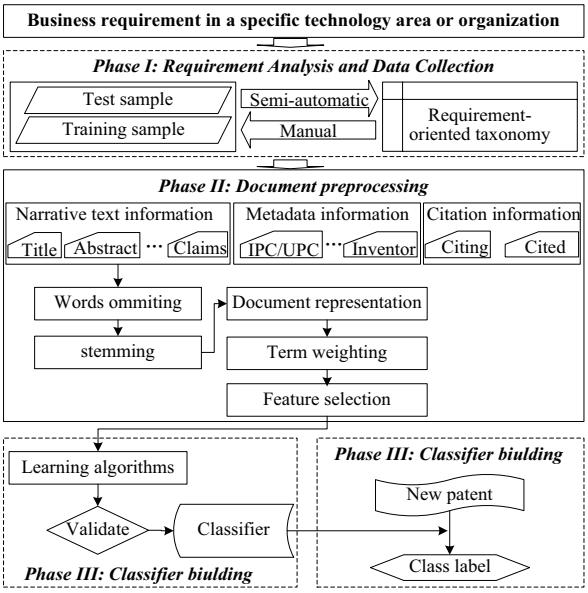

Fig.4. Automatic classification processes of patent classification to a requirement-oriented taxonomy using supervised machine learning techniques

3.2.1. *Requirement analysis and data collection*

In this phase, a retrieval strategy should first be made to collect raw patent documents, including the choice of patent database, search system and search words. Popular patent databases include the WIPO, EPO, USPTO, JPO, etc. The database of the USPTO is one of the favored sources to conduct a patent search because the US market is an important market for technology-transfer and international trade, which is combined with the territoriality of patent protection, luring inventors to file patent applications in the US. A good search strategy may be a combination of keywords and classification codes such as IPC and UPC. After a database is selected, the next step is to collecting raw patents according to the specific objective of patent analysis and patent management.

The requirement-oriented patent classification proposed in the paper is designed to overcome the weakness of conventional patent classifications for in-depth patent classification. As a requirement-oriented scheme to conduct more comprehensible patent classification for a specific technology area or organization, the first task is to analyze the business objectives and possible requirements. After an extensive analysis and discussion, a requirement-oriented taxonomy of the target technology area should be defined. With the help of techniques such as term clumping[26] and topic modeling[27], it becomes a semi-automatic or even automatic task for domain experts to build the taxonomy. It is worth mentioning here that Hu, et al[7] present an approach to automatically construct a KOS of patent documents based on term clumping, Latent Dirichlet Allocation (LDA) model, K-Means clustering and Principal Components Analysis (PCA). Automatic classification of patents documents based on the computer generated KOS was realized. According to the earlier discussion above, the taxonomy is often a mix of both technical and application topics.

Based on the data collection and the topic hierarchy developed above, the last step of phase one is to manually build a suitable size of sample, which is of significance to build and validate a classifier in the third phase. As a result of phase I, a requirement-oriented taxonomy and a dataset of manually labeled patent documents are built.

3.2.2. *Document preprocessing*

Preprocessing of patent documents includes word segmentation, stop words omitting using a stop words dictionary, word stemming, document representation, feature selection and word/term weighting. The goal of the first phase is to select the most efficient features to build a VSM (Vector Space Model) using term weighting and feature selection techniques. Every patent document is then represented by a feature vector for classification.

As shown in Figure 4, information of a patent document is grouped into narrative text information, metadata information, and citation information. Narrative text information includes title, abstract and claims etc. of a patent document. Metadata of a patent mainly includes the patent code, IPC codes, application date and inventor, etc. Citation information refers to the cited patents and citing patents of a patent document. As mentioned in the literature review, these sections are widely used to represent a patent document in patent retrieval applications. Derived from these, different combinations of these sections are indexed to represent a patent document in this paper while conducting patent classification. Word segmentation, extraction of classification codes and reference code, removing stop words and stemming of narrative text are performed using a stemming algorithm such as the Porter stemmer (Porter, 1980), which is the most popular stemming algorithm.

In order to limit the dimensionality of the feature space, feature selection, also known as variable selection, attribute selection or variable subset selection, is then performed to retain the most relevant features and improve classification performance[28]. Information Gain (*IG*) is frequently employed as one of the best term-goodness criterion feature selection approaches in the field of machine learning. It measures the number of bits of information obtained for category prediction by knowing the presence or absence of a term in a document[29]. It is defined in Eq. (1):

$$IG(T_k) = H(C) - H(C|T_k)$$

$$= -\sum_{i=1}^{n} P(C_i)\log P(C_i) + P(T_k)\sum_{i=1}^{n} P(C_i|T_k)\log P(C_i|T_k)$$

$$+ P(\overline{T_k})\sum_{i=1}^{n} P(C_i|\overline{T_k})\log P(C_i|\overline{T_k}) \quad (1)$$

Where $IG(T_k)$ is the information gain of term T_k, $H(C)$ denotes the information entropy of the dataset, $H(C|T_k)$ denotes the conditional entropy, n denotes categories number, $P(T_k)$ represents the number of documents in which T_k occurs divided by the total number of documents (document frequency of T_k), $P(\overline{T_k}) = 1 - P(T_k)$, $P(C_i|T_k)$ denotes the conditional probability of a document belonging to category C_i when T_k occurs, while $P(C_i|\overline{T_k})$ denotes the conditional probability of a document belonging to category C_i when T_k doesn't occur.

To construct a feature vector for a document, term weighting is conducted to evaluate the importance of a term relative to a document. As the term implies, the widely-used term weighting technique term frequency and inverse document frequency (*TF-IDF*), developed by Salton and Buckley [30], calculates values for each word in a document through an inverse proportion of the frequency of the term in a particular document to the percentage of documents the term appears in. by taking both the term frequency and document frequency into consideration, term with high *TF-IDF* numbers imply a strong relationship with the document they appear in. It is defined as Eq. (2):

$$TFIDF(T_k, D_p) = TF(T_k, D_p) \times log\, P(N|n_{T_k}) \quad (2)$$

where $TF(T_k, D_p)$ denotes the term frequency of term T_k occurs in patent document D_p, and $log\, P(N|n_{T_k})$ represent the total number of patent documents divided by those in which T_k occurs (the inverse document frequency).

3.2.3. *Classifier building*

At this phase, training process is firstly performed using various supervised machine learning algorithms based on the features vectors of patent documents. Classifiers are validated and compared by testing dataset afterwards. Four standard classification performance metrics, accuracy, precision, recall and F-measure are popularly used to evaluate the performance of the classifiers. The metrics have been widely used in information retrieval and machine learning studies. Classification accuracy was used to evaluate the overall performance, as shown in Eq. (3):

$$accuracy = \frac{number\ of\ correctly\ classied\ patents}{total\ number\ of\ claasified\ patents} \quad (3)$$

Precision, recall, and *F*-measure were used to evaluate the classification performance. For instances of class i:

$$\text{precision}(i) = \frac{\text{number of correctly classied patents for class } i}{\text{total number of patents identified as class } i} \quad (4)$$

$$\text{recall}(i) = \frac{\text{number of correctly classied patents for class } i}{\text{total number of patents in class } i} \quad (5)$$

Precision and recall evaluate whether a classification is successful. If both parameters yield high scores in a classification experiment, the approach's performance is considered ideal. However, precision and recall usually conflict with each other, so the de facto standard measure F_1 is used to balance the precision and recall scores, as shown in Eq. (6):

$$\mathrm{F}_{1(i)} = \frac{2 \times \text{precision}(i) \times \text{recall}(i)}{\text{precision}(i) + \text{recall}(i)} \quad (6)$$

Since multi-label text classification resolves to binary text classification, the averaging F_1 is used to evaluate the overall performance over the different categories. The averaging F_1 computes the F_1 measure for each category and then takes the average over the per-category F_1 measure [31]. Given a training set with m categories, assuming that the F_1 value for the *ith* category is $F_{1(i)}$, the averaging F_1 is defined as:

$$\text{Averaging } \mathrm{F}_1 = \frac{\sum_{i=1}^{\mathrm{m}} \mathrm{F}_{1(i)}}{\mathrm{m}} \quad (7)$$

3.2.4. *Patent classification*

In the last phase, automatic classification of new patents based on one of the satisfactory classifiers above is conducted. We propose to use combinations of narrative text information, metadata and citation information as the representation of patent documents. Each patent is labeled by at least one of the classification topics. The issue of multi-label classification is discussed in the following.

3.3. *Multi-label classification*

Multi-label classification is the task of assigning an object simultaneously to one or multiple classes. It is quite common in conventional patent classification. Take IPC as an example, the US patent "Common circuitry supporting both bit node and check node processing in LDPC (Low Density Parity Check) decoder" has four different IPC codes from two sections: H03M13/00 (coding, decoding or code conversion, for

error detection or error correction; coding theory basic assumptions; coding bounds; error probability evaluation methods; channel models; simulation or testing of codes); G06F11/00 (error detection; error correction; monitoring); H03M13/03 (error detection or forward error correction by redundancy in data representation, i.e. code words containing more digits than the source words); H03D1/00 (demodulation of amplitude-modulated oscillations). Techniques and innovations in a patent can be used in a number of different applications; therefore, multi-label classification occurs widely in patent topic classification proposed in this paper.

From the extensive literature on multi-label classification, You et al [32] proposed a Multi-label Embedded Feature Selection (MEFS) to improve multi-label classification for music emotions. Shen et al [33] summarized three models to train multi-label data: MODEL-s, which only labels the main class, simplifying multi-label classification to single-label classification; MODEL-*i*, which ignores the multi-label data when training the classifier; MODEL-*n*, which considers the combination of multiple labels of an item as a new class and builds a model for it. In addition, they proposed a novel model, MODEL-*x*, which uses the multi-label data more than once when training, using each example as a positive example of each of the classes to which it belongs. As suggested by Shen, the document is only used as a positive example of each class it belongs. Say, one document D is associated with both topic *i* and topic *j*, labeled as the class *i* and class *j*. When we train a binary classifier for the class *i*, D is only labeled as positive although it also belongs to the class *i* which is among the negative classes. Similarly, D is only labeled as positive when we are training a binary classifier for the class *j* and is labeled as negative when training binary classifiers for any other classes. The documents which do not belong to the class being trained are all labeled as negative for this binary classifier.

The purpose of the paper is to propose the requirement-oriented patent classification scheme for a specific technology area or organization, and compare the automatic classification performances based on different features of a patent document. A number of studies have proposed many methods to calculate the membership degree of an instance to each category. In this paper, we adopt the MODEL-*x* to train a binary classifier for each class and analyze the performance of each class, and the overall performance of a classifier is measured by the average performance of all classes.

4. Case Study and Experiment Analysis

In order to validate the effectiveness of the method and compare the classification performances based on different document features, in this section we conduct a case study on the SoC technology area.

4.1. *Case description*

A system on a chip or system on chip (SoC) is an integrated circuit (IC) that integrates all components of a computer or other electronic system into a single chip. It may contain digital, analog, mixed-signal, and often radio-frequency functions, all on a single chip substrate. Thanks to its high integration and flexibility, SoC is widely applied in the information and communication technology industry.

4.2. *Data and experiment*

4.2.1. *Requirement analysis and data collection*

We have been working on a project to analyze the field of SoC technology. According to the project's tech-mining requirements, we mainly focus on three parts of the SoC technology: SoC on digital television, SoC on personal mobile terminal, and the adaptable multimode RF SoC. As mentioned earlier, the conventional patent classifications, such as the IPC and UPC, are too complicated and too general to meet the detailed requirements of such a specific technology area. According to the specific requirement for SoC patent analysis, we build an application-technical mixed taxonomy with the help of a team of SoC experts. The taxonomy has four layers and 201 leaves in total.

Meanwhile, a number of 14,414 patents related to the SoC technology are collected automatically by our collection system from USPTO Patent Full-Text and Image Database, filed from 1976 to 2010. Our team of SoC experts manually classified all the 14,414 patents. Each patent document is labeled by 1 to 3 different categories.

4.2.2. *Document preprocessing*

Several patent documents have missing field values such as title, thus we just remove them from the dataset in our experiment. Meanwhile, in our dataset, the 14414

documents are associated with 37476 labels in total. Therefore, every patent is averagely associated with 2.6 labels, which lead to the issue of multi-label classification as described in section 3.3. In the paper we just train a binary classifier for each class and analyze the performance of each class using MODEL-*x* proposed by Shen[33].

Not all patent documents in our dataset can be used for the experiment. For example, there are classes that very little to no patent documents classified to them. The case study is aimed to validate whether it is practical to automatic classify patents using traditional machine learning techniques for a real world requirement-oriented patent classification task, where the user-defined categories are so specific and similar to each other. So we just choose classes with more than 85 patent documents classified to them to conduct the experiment. After filtering, 28 classes covering 8316 documents are chosen at last, listed in Table 1. For a class, positive documents are documents that are labeled by the class, while negative documents are those not labeled by it. The Table shows that for the 28 selected classes, the negative class contains many more examples than the positive class, with an imbalance ratio from 3.14:1 to 95.69:1. In the area of concept-learning, this type of data set is said to be exhibit a class imbalance[34]. A popular approach to deal with class imbalance is to re-sample classes: under-sampling the majority class to match the size of minority class or over-sampling the minority class to match the size of majority ones[35]. In the paper, we simply choose under-sampling to do the experiment. Thus, each category contains 86 patent

Table 1. Class ID, explanation, number of positive and negative documents for each class.

Class ID	Class explanation	Number of positive documents	Number of negative documents
	SoC on digital television		
c01	ATSC demodulation chip -> FEC decoding -> RS+TCM decoder	116	8200
c02	ATSC demodulation chip -> FEC decoding -> decoder/scrambler	86	8230
c03	DVB-C decoding -> AGC	293	8023
c04	DVB-C decoding -> synchronization	1091	7225
c05	DVB-C decoding -> de-mapping	249	8067
c06	DVB-C decoding -> sampling	178	8138
c07	Key technology in SOC design -> deep sub-micron technology -> power consumption	988	7328
c08	Key technology in SOC design -> design method research -> SOC, IC, integrated circuit, or application-specific integrated circuit design method	90	8226
c09	Source decoder chip -> MPEG-2	138	8178
c10	Display control and driver -> WXGA & FHD resolution supported	461	7855
c11	Video post-processing -> multimedia playback -> multimedia software technology	206	8110
c12	Video post-processing -> multimedia playback -> data compression and decompression	96	8220
c13	Video post-processing -> format conversion -> resolution, image conversion technology	217	8099
c14	Video post-processing -> quality enhancement -> ACM active color management	96	8220
c15	Video post-processing -> quality enhancement -> Y/C separation and compensation technology	147	8169
c16	Video post-processing -> quality enhancement -> interlaced scanning technology	133	8183
	SoC on personal mobile terminal		
c17	Navigation -> tracking and capture -> multipath interference suppression	86	8230
c18	Navigation -> tracking and capture -> high dynamic capture technology	410	7906
c19	Navigation -> measurement calculating -> differential positioning	94	8222
c20	Mobile communication -> GSM -> CDMA	190	8126
	Adaptive multimode RF SoC		
c21	Circuit design -> clock generator -> large tuning range voltage controlled oscillator	1062	7254
c22	Circuit design -> clock generator -> loop filter design	245	8071
c23	Circuit design -> clock generator -> high speed frequency divider design	205	8111
c24	Manufacturing process -> Bi CMOS	2010	6306
c25	Manufacturing process -> GaAs	860	7456
c26	Manufacturing process -> SiGe	222	8094
c27	Mobile communication -> low power consumption RF Front-End circuit design techniques	100	8216
c28	Mobile communication -> signa interference	449	7867

documents in the experiment dataset.

After stop words removing and word stemming using the Porter stemmer is performed. Referring to Zhang[36], term weighting is performed at word and code (e.g. the UPC code and reference patent number of a patent document) level by calculating the *TF-IDF* of the attribute in each document. In this paper, we choose Title, Abstract, Claims, IPC, UPC, Reference, Title+ Abstract, Title+ IPC, Title+ UPC, Title+ Reference, Abstract+ IPC, Abstract+ UPC, Abstract+ Reference, IPC+ Reference, and UPC+ Reference as the representation of a patent document to conduct comparison experiments. In order to retain the most relevant features and thereby limit the dimensionality of the feature space, we just choose IG as the feature selection criterion, which is also frequently used in other text classification tasks[19]. The number of attributes of every presentation scheme after feature selection is listed in Table 2.

Table 2. Attribute number after feature selection with IG.

Presentation	After_IG	Presentation	After_IG
Title	93	Title+UPC	226
Abstract	238	Title+Reference	225
IPC	129	Abstract+UPC	666
UPC	95	Abstract+IPC	700
Claim	314	Abstract+Reference	591
Reference	20	IPC+Reference	261
Title+Abstract	589	UPC+Reference	227
Title+IPC	200		

4.2.3. *Classifier building and validation*

In this study, we chose title, abstract and claims of a patent document as the representation of narrative text information respectively, IPC and UPC as metadata information respectively, and reference as citation information. Three classical machine learning techniques, i.e. decision tree (DT), Naïve Bayes (NB), and support vector machine (SVM), are adopted to build classifiers for patent topic classification of the SoC area. 10 fold cross validation is adopted to validate the quality of these algorithms. Accuracy and average F_1 value are used to evaluate the overall performance of classifiers.

4.3. *Result and Analysis*

Table 3 lists the classification performances based on single sections of a patent, namely title, abstract, claims, IPC, UPC and Reference.

From the perspective of document presentation, patent classification based on narrative text information performs better than metadata information and citation information, no matter which classification algorithm is adopted. Moreover, both accuracy and average F1 are high enough when using title and abstract as the document representation, not less than 0.65. That means the highly concise title and human generated abstract can well reflect the topic category of a patent.

While from the perspective of machine learning algorithms, SVM works best, NB next, and DT the worst generally when using narrative text information as the document representation. While using IPC, UPC or reference patent number as the document representation, the best performance is obtained by NB. DT still works the worst. Taking time into consideration, NB is the most efficient algorithm while SVM takes most of the time in our experiment.

The results above show that, compared with metadata and citation information, the title and abstract can well reflect a patent's topic category. As a result, we use them to represent the narrative text information, IPC and UPC as the representation of metadata information, and reference as the representation of citation information. In the next experiment, we take their combinations as document representation to test the performance when classifying patent documents using different combinations of narrative text, metadata and citation

Table 3. Classification performances based on single sections of a patent document using three different machine learning algorithms.

	DT		NB		SVM	
	accuracy (%)	average F_1	accuracy (%)	average F_1	accuracy (%)	average F_1
Title	65.03	0.65	66.57	0.667	67.86	0.678
Abstract	84.01	0.84	84.43	0.844	88.54	0.885
Claims	60.51	0.603	67.11	0.669	70.47	0.705
IPC	46.97	0.507	49.96	0.541	48.51	0.527
UPC	37.21	0.396	40.45	0.433	39.16	0.423
Reference	5.57	0.046	7.89	0.075	7.60	0.075

Table 4. Classification performances based on single sections of a patent document using three different machine learning algorithms.

	DT		NB		SVM	
	accuracy (%)	average F_1	accuracy (%)	average F_1	accuracy (%)	average F_1
Title +Abstract	98.68	0.978	100.00	1	100.00	1
Title +UPC	99.95	1	99.96	1	99.96	1
Title +IPC	100.00	1	100.00	1	100.00	1
Title +Reference	99.95	1	100.00	1	100.00	1
Abstract +IPC	100.00	1	100.00	1	100.00	1
Abstract +UPC	99.90	0.999	99.90	0.999	100.00	1
Abstract +Reference	100.00	1	100.00	1	100.00	1
UPC+ Reference	100.00	1	100.00	1	100.00	1
IPC+ Reference	100.00	1	100.00	1	100.00	1

information as the document representation.

Table 4 lists the classification performances based on different combinations of Title, Abstract, UPC, IPC, and Reference. Compared with Table 3, we draw a conclusion that satisfactory classification performances can be easily obtained using the combination of any two kinds of the above information as the document representation.

Preprocessing of metadata information and citation information is much simpler than narrative text information. At the document preprocessing phase, word segmentation, stop words omitting, and word stemming need to be done for narrative text information such as title, abstract and claims, while metadata information such as IPC and UPC or citing patent number can be weighted directly as a term. What's more, IPC and UPC are given by examiners; patent citations are valuable since they are directly related to a patent's economic value. Furthermore, we conducted a series of experiments to see how the performance for classifying patents based on IPC+ Reference and UPC+ Reference using NB will be when different size of attributes are retained. The performances are illustrated in Table 5. The Table indicates that the more attribute number is, the higher average accuracy and F1 value is obtained, and an average F1 value bigger than 0.8 can be obtained when attribute number is more than 60.

It is obvious to conclude from Table 3 and Table 4 that patent classification to the requirement taxonomy performs poorly when using a single section of whether metadata information or citation information, but perfect performances are obtained when the document representation is based on the combinations of any two of the above narrative text information, metadata information and citation information. Table 5 indicates that in the process of patent topic classification, using the combination of metadata information and citation information as the document representation can obtain a relatively high classification accuracy with a small scale of features. Considering the merits of metadata and citation information of a patent described earlier, the process of document preprocessing can consequently be simplified dramatically, and at last enhancing the efficiency of classification.

In summary, according to the specific requirement for SoC patent analysis, categories of the SoC technology is divided into three blocks: SoC on digital television, SoC on personal mobile terminal, and the adaptable multimode RF SoC. An application-technical mixed taxonomy with four layers and 201 leaves is defined by

Table 5. Average F1 value for different attribute number when classifying patents based on *IPC+ Reference* and *UPC+ Reference* using Naïve Bayes.

NB		30	40	50	60	70	80	90	100	150
IPC+Reference	accuracy(%)	57.018	71.179	71.179	81.894	85.465	89.037	92.608	92.61	100
	average F1	0.55	0.69	0.695	0.807	0.846	0.885	0.925	0.925	1
UPC+Reference	accuracy(%)	56.312	70.598	84.302	85.797	92.691	96.263	100	100	100
	average F1	0.541	0.686	0.836	0.85	0.924	0.962	1	1	1

the domain experts. It is more comprehensible and adaptable to the patent analysis purpose. Automatic patent classification into the requirement-oriented taxonomy defined by domain experts based on metadata and citation information performs well.

5. Conclusion and Future Study

We present a supervised requirement-oriented patent classification scheme aiming to a specific technology area or organization in this paper. Patents are classified into a requirement-oriented taxonomy of technical topic, application topic or application-technical mixed topic. In the paper, automatic process of the method based on metadata and citation information using supervised machine learning techniques consists of four phases.

A case study on SoC technology using DT, NB and SVM validates the effectiveness of the novel classification scheme. The experiments are conducted with term weighting technique of TFIDF and feature selection technique of IG. We draw three main conclusions:

(i) In the case of classifying patents by using a single section as the document representation, narrative text information such as title and abstract is much better than metadata and citation information;

(ii) Properly using the combination of narrative text information, metadata information and citation information as the document representation, can dramatically improve the classification performance;

(iii) Relatively high classification accuracy using Naïve Bayesian classifier based on the document representation of IPC+Reference or UPC+Reference can be received. Given the fact that patent metadata (especially classification code) and citation information (namely cited or citing patent number) can be very easily extracted and processed, using the combination of these information as the document representation can not only obtain perfect performance, but also simplify the document preprocessing process in a large scale, which is of great significance for real world requirement-oriented patent classification tasks.

The requirement-oriented taxonomy is aimed towards a specific technology area or organization. It is easy to comprehended and adaptable to various patent analysis requirements. Patent classification using metadata and citation information of patents is effective and easy to conduct.

However, we have just presented a practical requirement for classifying patent documents of a specific domain or organization into a user-defined taxonomy through our patent mining practices, and conducted a range of experiments on the SoC technology area to conclude the most effective document representation for our patent classification task. To utilize the method in constructing a real world requirement-oriented patent classification system for a specific technology area or organization, there are more practical issues need to be considered. More case studies and experiments need to be done to validate a more general conclusion. In addition, unsupervised methods are to be proposed to automatically build a requirement-oriented patent taxonomy and then classify a target patent set.

To sum up, the paper is just a start for the requirement-oriented patent classification research and we have a lot of work to do in our future study. First, more work is needed to improve the semi-automatic procedure of defining the requirement-oriented taxonomy using techniques such as term clumping and topic modeling. Second, the problem of multi-classification of patents, which we solve by training a binary classifier for each class and analyzing the performance of each class in the paper, should be addressed in the future. Last, words, classification codes and citation patent numbers are weighted without discriminative in the paper, so a good weighting method to treat them respectively needs to be studied further.

Acknowledgement

This research is financially supported by the National High Technology Research and Development Program of China (Grant No.2014AA015105), the General Program of National Natural Science Foundation of China (Grant No.71373019) and the National Key Technology R&D Program (Grant No. 2013BAH20F01). The authors would like to thank for the help from all teachers in co-lab of Technology Innovation.

References

1. Montecchi, T., Russo, D., & Liu, Y. (2013). Searching in Cooperative Patent Classification: Comparison between keyword and concept-based search. *Advanced Engineering Informatics*, *27*(3), 335-345.

2. Ma, J., & Porter, A. L. (2015). Analyzing patent topical information to identify technology pathways and potential opportunities. *Scientometrics*, *102*(1), 811-827.
3. Yoon, B., Park, I., & Coh, B. Y. (2014). Exploring technological opportunities by linking technology and products: Application of morphology analysis and text mining. *Technological Forecasting and Social Change*, *86*, 287-303.
4. Yang, C. S., Wei, C. P., & Chiang, Y. H. (2014). Exploiting Technological Indicators for Effective Technology Merger and Acquisition (M&A) Predictions. *Decision Sciences*, *45*(1), 147-174.
5. Ernst, H. (2003). Patent information for strategic technology management. *World Patent Information*, *25*(3), 233-242.
6. Lai, K. K., & Wu, S. J. (2005). Using the patent co-citation approach to establish a new patent classification system. *Information Processing & Management*, *41*(2), 313-330.
7. Hu, Z., Fang, S., & Liang, T. (2014). Empirical study of constructing a knowledge organization system of patent documents using topic modeling. *Scientometrics*, 1-13.
8. Terninko, J., Zusman, A., & Zlotin, B. (1998). *Systematic innovation: An introduction to TRIZ (theory of inventive problem solving)*. CRC press.
9. Loh, H. T., He, C., & Shen, L. (2006). Automatic classification of patent documents for TRIZ users. *World Patent Information*, *28*(1), 6-13.
10. Cong, H., & Tong, L. H. (2008). Grouping of TRIZ Inventive Principles to facilitate automatic patent classification. *Expert Systems with Applications*, *34*(1), 788-795.
11. Sureka, A., Mirajkar, P. P., Teli, P. N., Agarwal, G., & Bose, S. K. (2009). Semantic based text classification of patent documents to a user-defined taxonomy. In *Advanced Data Mining and Applications* (pp. 644-651). Springer Berlin Heidelberg.
12. Benzineb, K., & Guyot, J. (2011). Automated patent classification. In *Current challenges in patent information retrieval* (pp. 239-261). Springer Berlin Heidelberg.
13. Krier, M., & Zacca, F. (2002). Automatic categorisation applications at the European patent office. *World Patent Information*, *24*(3), 187-196.
14. Fall, C. J., Törcsvári, A., Benzineb, K., & Karetka, G. (2003, April). Automated categorization in the international patent classification. In *ACM SIGIR Forum* (Vol. 37, No. 1, pp. 10-25). ACM.
15. Kim, J. H., & Choi, K. S. (2007). Patent document categorization based on semantic structural information. *Information processing & management*, *43*(5), 1200-1215.
16. Koster, C. H., Seutter, M., & Beney, J. (2003, January). Multi-classification of patent applications with Winnow. In *Perspectives of System Informatics* (pp. 546-555). Springer Berlin Heidelberg.
17. Fall, C. J., Törcsvári, A., Fiévet, P., & Karetka, G. (2004). Automated categorization of German-language patent documents. *Expert Systems with Applications*, *26*(2), 269-277.
18. Tantug, A. C. (2010). Document Categorization with Modified Statistical Language Models for Agglutinative Languages. *International Journal of Computational Intelligence Systems*, *3*(5), 632-645.
19. Bhumika, Sukhjit S S, Anand N. (2013). A review paper on algorithms used for text classification. *International Journal of Application or Innovation in Engineering & Management*, *3* (2), 90-99.
20. D'hondt, E., Verberne, S., Koster, C., & Boves, L. (2013). Text representations for patent classification. *Computational Linguistics*, *39*(3), 755-775.
21. He, C., & Loh, H. T. (2010). Pattern-oriented associative rule-based patent classification. *Expert Systems with Applications*, *37*(3), 2395-2404.
22. Richter, G., & MacFarlane, A. (2005). The impact of metadata on the accuracy of automated patent classification. *World Patent Information*, *27*(1), 13-26.
23. Liu, D. R., & Shih, M. J. (2011). Hybrid‐patent classification based on patent‐network analysis. *Journal of the American Society for Information Science and Technology*, *62*(2), 246-256.
24. Li, X., Chen, H., Zhang, Z., Li, J., & Nunamaker, J. F. (2009). Managing knowledge in light of its evolution process: An empirical study on citation network-based patent classification. *Journal of Management Information Systems*, *26*(1), 129-154.
25. Li, X., Chen, H., Zhang, Z., & Li, J. (2007, June). Automatic patent classification using citation network information: an experimental study in nanotechnology. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries* (pp. 419-427). ACM.
26. Zhang, Y., Porter, A. L., Hu, Z., Guo, Y., & Newman, N. C. (2014). "Term clumping" for technical intelligence: A case study on dye-sensitized solar cells. *Technological Forecasting and Social Change*, *85*, 26-39.
27. Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*,*101*(Suppl 1), 5228-5235.
28. Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *Knowledge and Data Engineering, IEEE Transactions on*, *17*(4), 491-502.
29. Yang, Y., & Pedersen, J. O. (1997, July). A comparative study on feature selection in text categorization. In *ICML* (Vol. 97, pp. 412-420).
30. Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*(5), 513-523.
31. Aiolli, F., Cardin, R., Sebastiani, F., & Sperduti, A. (2009). Preferential text classification: learning algorithms and evaluation measures. *Information retrieval*, *12*(5), 559-580.

32. You, M., Liu, J., Li, G. Z., & Chen, Y. (2012). Embedded feature selection for multi-label classification of music emotions. *International Journal of Computational Intelligence Systems*, *5*(4), 668-678.
33. Shen, X., Boutell, M., Luo, J., & Brown, C. (2003, December). Multilabel machine learning and its application to semantic scene classification. In *Electronic Imaging 2004* (pp. 188-199). International Society for Optics and Photonics.
34. Jo, T., & Japkowicz, N. (2004). Class imbalances versus small disjuncts. *ACM SIGKDD Explorations Newsletter*, *6*(1), 40-49.
35. Sebastiani, F. (1999). A tutorial on automated text categorisation. In *Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence* (pp. 7-35). Buenos Aires, AR.
36. Zhang, X. (2014). Interactive patent classification based on multi-classifier fusion and active learning. *Neurocomputing*, *127*, 200-205.