Foundations and Trends® in
Information Retrieval
Vol. 7, No. 1 (2013) 1–97

DOI: 10.1561/1500000027

now
the essence of knowledge

Patent Retrieval

By Mihai Lupu and Allan Hanbury

Contents

Foundations and Trends® in
Information Retrieval
Vol. 7, No. 1 (2013) 1–97

DOI: 10.1561/1500000027

now
the essence of knowledge

Patent Retrieval

Mihai Lupu[1] and Allan Hanbury[2]

[1] *Vienna University of Technology, Favoritenstraße 9-11/188, Vienna, 1040, Austria, lupu@ifs.tuwien.ac.at*
[2] *Vienna University of Technology, Favoritenstraße 9-11/188, Vienna, 1040, Austria, hanbury@ifs.tuwien.ac.at*

Abstract

Intellectual property and the patent system in particular have been extremely present in research and discussion, even in the public media, in the last few years. Without going into any controversial issues regarding the patent system, we approach a very real and growing problem: searching for innovation. The target collection for this task does not consist of patent documents only, but it is in these documents that the main difference is found compared to web or news information retrieval. In addition, the issue of patent search implies a particular user model and search process model. This review is concerned with how research and technology in the field of Information Retrieval assists or even changes the processes of patent search. It is a survey of work done on patent data in relation to Information Retrieval in the last 20–25 years. It explains the sources of difficulty and the existing document processing and retrieval methods of the domain, and provides a motivation for further research in the area.

1

Introduction

Innovation is at the core of technological and societal developments. New ideas on how to make things better, faster, cheaper, and more reliable, or simply on how to make totally new things, are the result of many different economical, managerial, and cultural factors. In addition to all these factors, it stands to reason that technology moves forward on the basis of prior technology, and that therefore society as a whole benefits from public availability of detailed descriptions of technical innovation. For this reason, the patent system has been created to encourage inventors to share their know-how, in exchange for a temporary monopoly. Without approaching any controversial topic, this review looks, from the perspective of Information Retrieval (IR) researchers, at methods for benefiting from this amount of information.

The search for innovation, as expressed in patent documents, and for the purposes of obtaining new patents, has two facets: the *search for content* and the *search for legal information*. Most of us, as scientists, are very familiar with the search for content. While performed for different purposes, at its core lies the need to understand a technical process or entity. We achieve this by finding references to similar processes or entities, by analyzing its components and the more general

categories of processes or entities of which it is part. Second, there is the search for legal information related to the protection granted to the inventor for a specific invention. The two facets often intermingle in the different search use cases described in Section 1.4, but this review will focus on the former.

We begin the introduction with a brief description of the past and present of the patent system, in order to provide a context for everything that will be discussed further on. This will also give the reader the understanding of the specific terminology used in the following sections. We continue with an overview of the content of patent documents in Section 1.2, illustrated by analyzing an example of a patent in Section 1.3. Section 1.4 gives a description of the most important patent search types. Finally, Section 1.5 gives a brief overview of patent research in the IR community, and the sources of patent documents.

1.1 History and Present

The term *"patent"* stems from the Latin verb *patere* and means *"laying open."* As a noun, it is the short form of *letters patent*, an official document used in the middle ages by an authority to assign specific rights to a person or group. The first patent law in the sense that we would imagine it today, i.e., pertaining to inventions, was issued in Venice in 1474 [138], followed by the British Statute of Monopolies of 1623, the United States in 1790, and France in 1791 [11]. The full history of the patent law is certainly not the focus here, but rather the point that, when approaching this particular field, one has to take into account centuries of practice. The side-effect of this public disclosure of inventions is a library of cultural heritage documenting the development of human technologies from the middle-ages to the present day. All this, and more, is prior-art to any new patent application.

The different laws in different countries result in several possible definitions of a patent. According to the World Intellectual Property Organization (WIPO) [3],

> "a patent is the right granted to an inventor by a State, or by a regional office acting for several States,

> which allows the inventor to exclude anyone else from commercially exploiting his or her invention for a limited period, generally 20 years."

The conditions under which such a right may be granted may also show slight differences between authorities, but generally four conditions have to be met [11]:

novelty: The invention must not have been described or used before the application

inventive step: The invention must also not be a new but obvious combination of existing processes or entities

industrial applicability: It must be possible to build or use it in practice (e.g., no patent for a *perpetuum mobile*)

non-excluded material: It must not refer to areas explicitly excluded by law from patenting (e.g., natural products)

The modern practice of patent law starts with the Paris Convention of 1883 [207]. At the time of writing, there were 174 nations listed as contracting parties, the latest one being Thailand in 2008. The Paris Convention is the first in a series of international agreements that aim to make the patent system a truly global one. For even though most laws take prior art to be any public data anywhere in the world, the practice is essentially a national one, with only one true multinational authority, the European Patent Office (EPO).[1]

The Paris Convention lays down one of the fundamental properties of the current patent system, the *priority*. In essence, the Convention allows the inventor to claim priority on an invention at any patent office of a signatory country, based on a prior application he/she made in any other signatory country, generally within 12 months. This system results in the creation of links between documents issued by different patent offices, in different languages, essentially covering the same invention. It is a fundamental property that, as we will see in the following sections, has found its utility not only in the search methods,

[1] Even in the case of the EPO, the actual patents are issued by national offices, but the procedure is greatly simplified.

but also in machine translation, network analysis and evaluation of IR systems.

Priorities create the possibility of building patent *families* — the set of patents describing the same invention. Depending on how flexible one is in linking the documents based on their priority references, the families can describe a very specific invention, or a general technical field. Families are however not a legal concept, and just to illustrate this flexibility, let us note that the WIPO, in its Handbook of Industrial Property Information and Documentation [206], identifies five types (simple, complex, extended, national, and artificial), while the EPO, on its information Web site,[2] gives three definitions and provides links to how some commercial providers define their understanding of patent families.

This difference in definition notwithstanding, there are at least 250,000 common applications per year among *The Five IP Offices* (IP5)[3] (i.e., the same application filed at more than one IP5 office) [141]. This amounts to a considerable body of comparable multilingual data. And, given the way the patent system currently works (i.e., applications for the same invention made and examined at different patent offices), a set of independent searchers are creating relevance judgments in an ad-hoc pooling-like way.[4]

The size of patent corpora is relatively small when compared to the current web corpora (ClueWeb'09 is 25 terabytes compressed [1], while none of the patent corpora available reach the 1 terabyte mark). However, the research issues are still abundant. This review covers the vast majority of the research already done, as well as points out potential avenues for the future. When talking to a patent expert, it emerges that the work still needing to be done for patent retrieval is a mixture of technology features and legal or administrative issues. While this review certainly focuses on the former, the two are surprisingly difficult to extricate from each other. Often enough, procedures are put in

[2] http://www.epo.org/searching/essentials/patent-families/definitions.html

[3] Five patent offices that have agreed to a tighter collaboration in patent prosecution: European Patent Office (EPO), United States Patent and Trademark Office (USPTO), Japan Patent Office (JPO), Korean Intellectual Property Office (KIPO), and State Intellectual Property Office of the People's Republic of China (SIPO).

[4] There are international efforts underway to eliminate this apparent work duplication in order to speed-up patent prosecution.

place to do a good job of searching with the technology of the 1980s or even earlier. A classical example of this would be the creation of extremely long and complex Boolean queries [24]. These procedures then become part of what the community generally defines as "patent search" and research is done to adapt to this particular scenario. This is a factor to keep in mind when looking beyond the restricted confines of the described use cases.

Adams, in his presentation to the WIPO in 2009 [8] and later in his keynote at the Patent Information Retrieval Workshop in 2011,[5] identified three areas of development for improving search:

(1) Search strategy development — the human factor
(2) Database creation and maintenance
(3) Search engines and information navigation tools

The three areas all interact with and are dependent on each other, but the *Human Factor* and the *Database creation and maintenance* are not the subject of this survey. However, it is important for the IR community to understand that although the core algorithm, its supplementary features and its interfaces are very important, they are just a third of the complete process. Furthermore, studies on information navigation and visualization in the IR community are sparse. We will briefly cover them in Section 4.5.

The need for better search engines is however particularly acute now, as the number of patent applications grows and, together with it, the backlog of patent offices. For instance, as of January 2011, the United States Patent and Trademark Office (USPTO) had a backlog of 1.2 million patent applications [34].

1.2 Domains within a Domain

A common understanding of a *domain-specific search engine* is that it *"limits its index to pages corresponding to a particular subject area, publisher or purpose"* [183]. This definition covers all aspects of what one may define as domain-specific search, provided we are slightly liberal

[5] http://ifs.tuwien.ac.at/pair2011

in its interpretation. The "subject area" component refers to domains such as scientific publications, healthcare, biomedicine, chemistry, etc. The "publisher" could be perceived as a publication-specific medium. Text (hyperlinked or not), images, and combinations thereof such as news feeds, blogs or twitter search are examples that come to mind in this sense. Finally, "purpose" is better understood as *users* or *use case domains* and implies a connection to the user performing the search and his or her motivations and objectives. Let us therefore rephrase this definition:

Definition 1.1. Domain-specific search [engine | process] is a search [engine | process] that fixes one or more of the following three dimensions:

(1) **subject area** (e.g., chemical, biomedical, healthcare)
(2) **publication form or medium** (e.g., blogs, micro-blogs, books)
(3) **users or use case domain** (e.g., patent search, cultural heritage, expert search)

It should be noted that in reality the three axes are not quite orthogonal. Some use case domains require specific subject areas or publication forms. The lack of perfect orthogonality has however never been an obstacle in IR and we should take this definition in this spirit as well.

Figure 1.1 shows a graphical representation of the three axes. As an illustration of a domain, a volume of the space can be used to represent a domain in a qualitative way. The domain of healthcare is shown as an example of a domain that covers a more limited subject area, but targets a large number of users and user scenarios (both medical professionals and non-professionals regularly search for health and medical information in a large number of scenarios). In contrast, the patent domain covers chemistry, mechanical engineering, electrical engineering, and practically all other domains of industry applicable human knowledge, but focuses on a relatively small number of users and use cases. We develop the discussion on these use cases in the patent domain in Section 1.4.

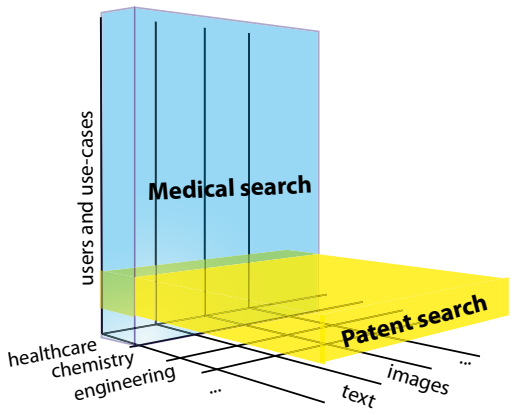


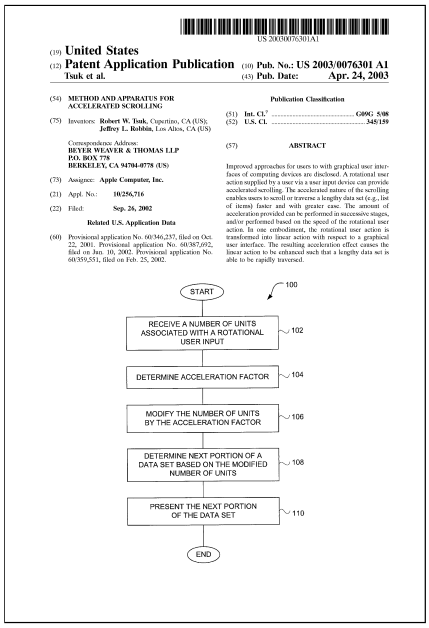
Fig. 1.1 The patent domain cuts across many scientific and technical domains.

1.3 A Patent Example

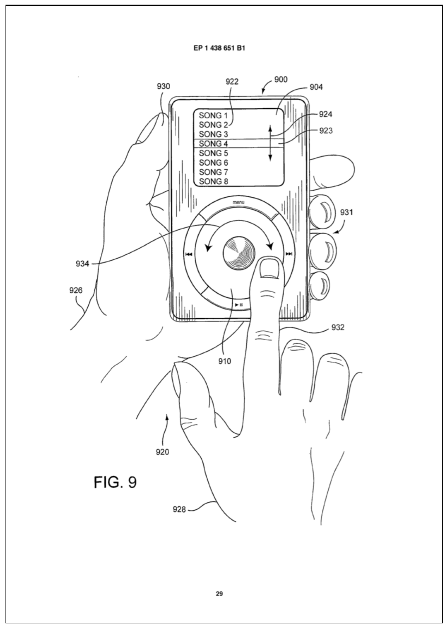
Before moving on with this survey, it is worth taking a closer look at the patent documents that we will often mention in later sections. As with any example, it does not cover all types of patents and aspects of the patenting process, but it provides the basic understanding necessary for subsequent sections. The reader familiar with the domain may skip this section.

Figure 1.2 shows two pages of the US patent application 10/256716, related to a very well-known consumer product. It is clear from Figure 1.2(b) that this patent application refers to the navigation mode of an iPod. The title (*Method and apparatus for accelerated scrolling*) and generally the first page is much less clear. In fact, this particular application was filed on September 26, 2002, after the first iPod was released, yet there is no mention of the device by its name in the text of the application. Instead, this patent application provides a detailed technical description of how the scrolling mechanism for the iPod works. For obvious reasons, we do not reproduce here the full text of the description, but it is available online.[6] The application requests

[6] http://1.usa.gov/Oq0Wr7



(a) The first page



(b) Page 12 of 28

FIG. 9

Fig. 1.2 Two pages of patent application 10/256716.

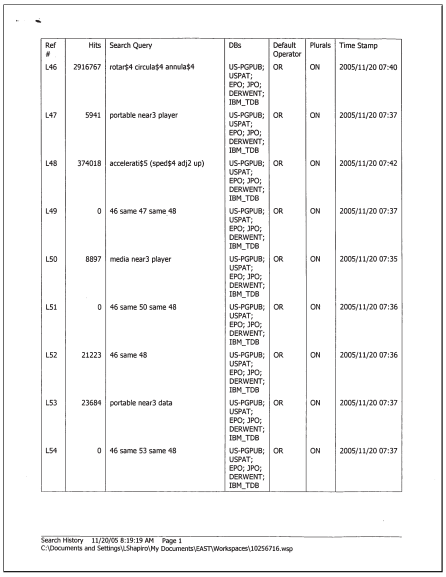
protection for a set of ideas, which it describes in 59 claims. Here are the first five:

(1) A method for scrolling through portions of a data set, said method comprising: receiving a number of units associated with a rotational user input; determining an acceleration factor pertaining to the rotational user input; modifying the number of units by the acceleration factor; determining a next portion of the data set based on the modified number of units; and presenting the next portion of the data set.
(2) A method as recited in claim 1, wherein the data set pertains to a list of items, and the portions of the data set include one or more of the items.
(3) A method as recited in claim 1, wherein the data set pertains to a media file, and the portions of the data set pertain to one or more sections of the media file.
(4) A method as recited in claim 3, wherein the media file is an audio file.
(5) A method as recited in claim 1, wherein the rotational user input is provided via a rotational input device.

As we can see, the claims are written in a particular style, sometimes referred to as *patentese* [19], resembling the language of many legal

contracts. In practice this is almost always the case. What one can also see from this example is that there are a number of internal references between claims. In practice, a claim which does not reference any other claim is called an *independent claim* and all others are called *dependent claims*. In the example above, Claim 1 is independent, while the following four are all dependent, forming a tree of references.

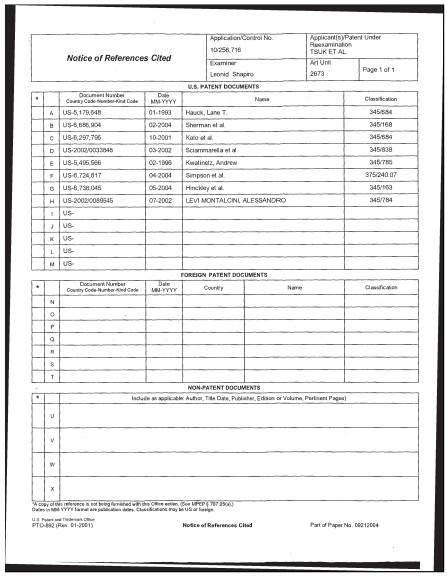
Following examination, this patent application was granted a patent in the United States, namely US7,312,785, issued over 5 years after the initial application. In this process, the examining office (i.e., the USPTO, in this case) published over a hundred documents, covering mostly the communication between the office and the applicants, as well as some procedural notes from the office. Among them, the most interesting for IR researchers are probably the *Examiner's search strategy and results*, the list of references cited by the examiner, and the series of decisions made by the office. In the end, the granted patent (US7,312,785) contains 40 claims only. Figure 1.3 shows two examples of documents published by the USPTO in relation to this application.



Ref #	Hits	Search Query	DBs	Default Operator	Plurals	Time Stamp
L46	2916767	rotar$4 circula$4 annula$4	US-PGPUB; USPAT; EPO; JPO; DERWENT; IBM_TDB	OR	ON	2005/11/20 07:40
L47	5941	portable near3 player	US-PGPUB; USPAT; EPO; JPO; DERWENT; IBM_TDB	OR	ON	2005/11/20 07:37
L48	374018	accelerati$5 (sped$4 adj2 up)	US-PGPUB; USPAT; EPO; JPO; DERWENT; IBM_TDB	OR	ON	2005/11/20 07:42
L49	0	46 same 47 same 48	US-PGPUB; USPAT; EPO; JPO; DERWENT; IBM_TDB	OR	ON	2005/11/20 07:37
L50	8897	media near3 player	US-PGPUB; USPAT; EPO; JPO; DERWENT; IBM_TDB	OR	ON	2005/11/20 07:35
L51	0	46 same 50 same 48	US-PGPUB; USPAT; EPO; JPO; DERWENT; IBM_TDB	OR	ON	2005/11/20 07:36
L52	21223	46 same 48	US-PGPUB; USPAT; EPO; JPO; DERWENT; IBM_TDB	OR	ON	2005/11/20 07:36
L53	23684	portable near3 data	US-PGPUB; USPAT; EPO; JPO; DERWENT; IBM_TDB	OR	ON	2005/11/20 07:37
L54	0	46 same 53 same 48	US-PGPUB; USPAT; EPO; JPO; DERWENT; IBM_TDB	OR	ON	2005/11/20 07:37

Search History 11/20/05 8:19:19 AM Page 1
C:\Documents and Settings\LShapiro\My Documents\EAST\Workspaces\10256716.wsp

(a) Examiner's strategy



Notice of References Cited

Application/Control No.	Applicant(s)/Patent Under Reexamination
10/256,716	TSUK ET AL.
Examiner	Art Unit
Leonid Shapiro	2673

Page 1 of 1

U.S. PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	Classification
	A	US-5,179,648	01-1993	Hauck, Lane T.	345/684
	B	US-6,686,904	02-2004	Sherman et al.	345/168
	C	US-6,297,795	10-2001	Kato et al.	345/684
	D	US-2002/0033846	03-2002	Sciammarella et al.	345/838
	E	US-5,495,566	02-1996	Kwatinetz, Andrew	345/785
	F	US-6,724,817	04-2004	Simpson et al.	375/240.07
	G	US-6,738,045	05-2004	Hinckley et al.	345/163
	H	US-2002/0089545	07-2002	LEVI MONTALCINI, ALESSANDRO	345/784
	I	US-			
	J	US-			
	K	US-			
	L	US-			
	M	US-			

FOREIGN PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

NON-PATENT DOCUMENTS

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	
	V	
	W	
	X	

*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.

U.S. Patent and Trademark Office
PTO-892 (Rev. 01-2001) **Notice of References Cited** Part of Paper No. 06212004

(b) List of citations

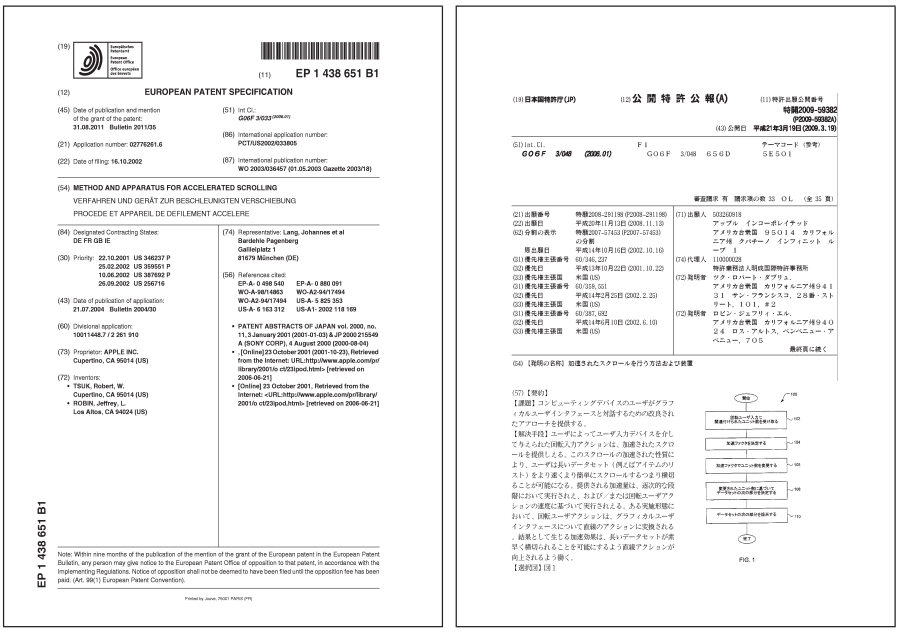
Fig. 1.3 Some documents published by the USPTO in relation to patent application 10/256716.

The US patent offers the owner a monopoly over the manufacture and licensing of the invention only in territories under the US jurisdiction. This is why, 20 days after filing the application with the USPTO, another application was filed with the WIPO: WO03/036457 on October 16, 2002. Even though the WIPO is not a patent office (i.e., it does not grant patents), it is the entry point to the so called *PCT Route*, which is a system designed to facilitate the acquisition of protection in different jurisdictions. The name of the route comes from the Patent Cooperation Treaty (PCT), which establishes the procedures under which an application filed with the WIPO reaches the national patent offices which can grant patents. Without going into the full details of the system,[7] here is the brief description of the route:

(1) an application is filed with the WIPO
(2) an International Search Report (ISR) is created by a Search Authority (an accredited search organization, generally a national patent office), and published. This phase is referred to as *Section 1*
(3) an optional further search is performed by a Search Authority (Section 2)
(4) the applicant decides, based on the search reports, whether to proceed to the national phase. This means that the patent application is passed to the set of offices where the applicant desires protection (they are called *designated states*)
(5) upon receiving the application, the national patent offices start their own examination procedures, optionally taking into account the ISR created in the previous steps.

Whether the application goes through the PCT route or not, in situations where protection is sought in different jurisdictions for the same invention, a so-called family of patents results. Figure 1.4 shows the European and Japanese patents corresponding to the US patent discussed before.

[7] Full details about the PCT and PCT applications are available at http://www.wipo.int/pct/en/. The details of the PCT route are much more complicated than the five bullet points presented here.



(a) European granted patent

(b) Japanese application

Fig. 1.4 Family members of US patent 7,312,785.

All of the application and granted patent documents will generally have the same structure, consisting of:

- **Bibliographical data**: title, metadata related to the specific publication at hand, the inventors, assignees, agents or applicants, as well as relations to other documents
- **Abstract**: a very brief summary of the invention
- **Description**: a detailed description of the invention, including prior work, examples, related technologies
- **Claims**: the legal description of the invention. Adams [11] defines the claims as a *"Sequence of paragraphs at the end of a patent application defining the scope of monopoly sought. After substantive examination, the same section of the granted patent defines the legal rights of the proprietor."*

We will often make references to these sections in the coming sections.

1.4 Patent Search Processes

While, in principle, the search process is always about finding relevant documents to satisfy a particular information need, patent search has specialized into different processes, differentiated as a function of the input (an idea, a disclosure of innovation, a patent application, a claim, a granted patent) and the needed output (a large set of scientific publications covering a domain, a set of patents, a single patent). In relation to patent search one will therefore often hear names such as *State of the art*, *Pre-filing patentability*, *Novelty*, *Freedom to operate*, *Validity*, or *Due diligence* search. Their precise names and definitions vary between different practitioners.[8] Alberts et al. [16] describe in detail five of these search types, but also demonstrate the variability in the definition, by providing a table with seven search types. Adams [11] adds another type of search (*Alerting*) and slightly regroups the rest.

Generally, these types of search are also related to eDiscovery because of their legal nature, which puts a large emphasis on finding *all* relevant documents. The greatest difference between the practice of patent searchers and legal staff is perhaps the amount of metadata available to patent searchers. As we have seen in the previous example, each published patent document is the result of a specific process and comes associated with a rich set of metadata.

The different types of patent search are summarized in Table 1.1, which shows the type of search and alternative names for it, as well as the search specification (what the search begins from) and the corpora in which the search is conducted. A short description of each search type is given in italic text. Figure 1.5 shows the life-cycle of an innovation, with the searches that are directly related to it. The figure describes the path from having an idea to do something new, i.e., an innovation, to obtaining (and defending) a patent. It follows four of the six types of search described in Table 1.1, in the order in which they occur, and shows the most important documents that are a result of this process. The rectangles denoting the searches also indicate who typically

[8] By practitioners, we understand here all those who deal with patents in their professional life. This generally includes corporate librarians, information specialists, private patent searchers, patent examiners at any patent office, and patent lawyers.

Table 1.1. Types of patent search.

Search type	Other names	Search specification	Corpora
State of the art	Technology survey	An idea	All public documents
To obtain a general understanding of the field surrounding the innovation at hand			
Pre-filing patentability		A fairly well defined innovation disclosure	All public documents
Similar to above, but with a more precise request for information and potentially more focus on patent documents			
Patentability	Novelty, Prior Art	A patent application	All public documents until the date of the application
Identify whether a specific patent application satisfies the conditions for granting			
Freedom to operate	Infringement, Right-to-Use, Clearance	A product and related methods or technologies	The set of patents in force in a particular jurisdiction
Identify any patent in force in a particular jurisdiction which may prevent a product from being commercialized in that jurisdiction			
Validity	Invalidity, Enforcement Readiness, Opposition	A granted patent	All public documents prior to the priority date of the patent in question
Identify whether a granted patent satisfied the granting criteria at the earliest priority date (i.e., the moment when a first application was registered for the invention described therein)			
Patent portfolio search	Due diligence, Patent landscape	A company, a technology area	All public documents
Obtain a general understanding of the patents, both in force and expired, in a specific technology area and/or jurisdiction			

performs them. Note that a search represented by a rectangle could take place over a number of hours or even weeks. The diamonds represent decision points, at which the question "Relevant item(s) found?" is asked, relating to the search just performed. If the response is positive, then a previous step in the process must be repeated.

Figure 1.5 does not show the *Freedom to operate* and *Patent portfolio* because, as can be seen from Table 1.1, these use cases do not

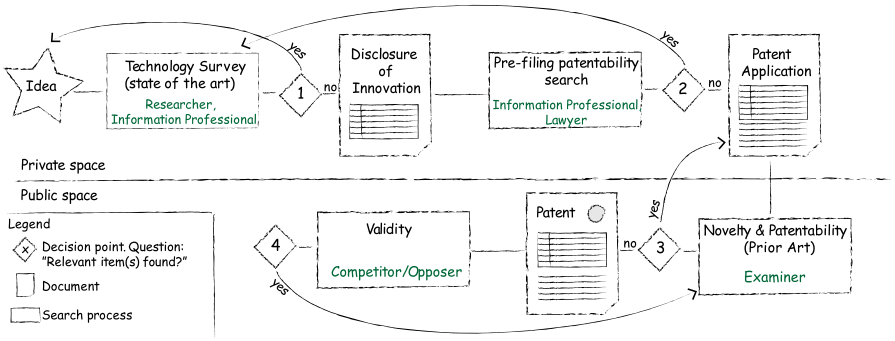


Fig. 1.5 High-level view of the life-cycle of a patented idea.

have a document at the basis of their request for information. These two types of search can in fact be performed at any time in the course of the development of an idea into a patented product, as well as any time thereafter.

Finally we should note that in Table 1.1, we refer to *all public documents* as a particular corpus, but what should be clear to the reader is that theoretically any publicly disclosed knowledge, not necessarily in written form, can be used to invalidate a patent.

The IR scientist reader may by now realize that for the core IR engine design, these different types of patent search do not appear to make a difference. The principal differences, as we have listed them, have to do with the target data collection or the form in which the request for information is expressed. The differences lie in the attitude with which the search process is conducted and the tools that assist the user in achieving the different objectives of these different patent searches.

Understanding these processes is important for the success of the resulting search system and surveys among professional patent searchers have been conducted in this sense. Hansen [77] have interviewed patent examiners at the Swedish patent office. Tseng and Wu [191] have interviewed 43 patent searchers, 18 of which agreed to a follow-up experimental observation of their search behavior. Most recently, Azzopardi et al. [20] followed up with an online survey which received 81 responses. Characteristically, the set of patent search types

identified was also different from the sets of Alberts et al. [16] and Adams [11] mentioned before.

Based on these and similar surveys in the context of the PROMISE project[9] [94] (Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation), the scenario outlined in Listing 1.1 has been crystallized to model the prior art search performed by a patent examiner. This is perhaps the most common type of search, if we do not take into account the technology survey done in the context of research environments. Nevertheless, the patent examiner at a patent office is not the only actor in the field. As Figure 1.5 and Table 1.1 indicate, scientists, librarians, and lawyers are also potentially involved.

Such user studies reveal aspects that the previous taxonomy did not identify. The different types of search identified by patent experts themselves have to do, as we have seen, with the final objectives of the search task. From an IR researcher's point of view, however, the different tasks

Listing 1.1. Prior Art Search use case.

```
1.  User receives a patent application document to evaluate
2.  User enters the search system
3.  User enters a text query, potentially with Boolean operators
    and specific field filters
4.  System presents a result list, sorted by relevance, with
    snippets, metadata information, and links to full documents
5.  User inspects and assesses all the documents
6.  User clicks on one element of the list for further inspection
7.  System presents the full document, with any metadata,
    attached images and text
8.  User inspects and assesses. Finds the document potentially
    relevant and saves it to a bucket
9.  User clicks on the 'Back' button to return to the list of
    results
10. System presents the list, with the already viewed documents
    visibly identifiable
11. Jump to Step 6, unless new query query is required or User
    satisfied
12. Based on a potential new understanding, User inputs a new
    query
13. Jump to Step 3, unless User satisfied
14. User saves the list and creates a search report
15. Use case ends
```

[9] http://www.promise-noe.eu

are distinguished by the formulations of the queries (Boolean, keywords only, full text, metadata, images), by the number of sessions required in the process, and by the level of collaboration in achieving the desired goal. The core difficulty for the IR scientist has to do with the patent documents themselves, even if many search types require the corpus to be the entire, publicly available, human knowledge.

1.5 Patents in the IR Community

This section reviews the beginnings of IR research in the patent domain, and lists a number of sources of information on IR in the patent domain, as well as sources of patent collections for use in IR research. The first recognition of the IR community of the need for special attention to be dedicated to the issues of patent retrieval was in the organization of the first Workshop on Patent Retrieval by Kando and Leong [97]. However the subject had been approached before by Sheremetyeva and Nirenburg [171] and by Larkey [110]. Other studies, for specific domains (particularly chemistry [32, 99]) and for business analysis (for example, [42]) had appeared even before, but outside of the core IR field.

1.5.1 Sources of Knowledge

Work in this domain has been encouraged through evaluation campaigns, first through the NII[10] Test Collection for IR Systems (NTCIR) [57, 60, 93, 145, 146], and later through the Text Retrieval Conference (TREC) (for the chemical domain [129]) and continuing at the moment of writing this review, through the Cross-Language Evaluation Forum (CLEF) [156, 157, 158, 161].

This survey relies heavily on the proceedings of the SIGIR Workshop on Patent IR in 2000 [97]; the ACL workshop on Patent Corpus Processing [4]; the special issue of the Information Processing & Management journal [59]; the PaIR series of workshops [5], generally co-located with the International Conference on Information and Knowledge Management (CIKM); the Advances in Patent Information Retrieval (AsPIRe) Workshop [76] co-located with the European Conference on

[10] National Institute of Informatics (Japan).

Information Retrieval (ECIR) in 2010; as well as the reports from the evaluation campaigns mentioned above. A number of other articles have appeared in various journals, specific to the technology described, and at least another survey has appeared in the *World Patent Information* (WPI) journal [35]. The WPI journal is an important source of domain related information and while its audience is mostly made up of patent information professionals, the IR researcher interested in the domain would find its articles interesting. Most recently, in 2011, an edited book has collected a series of articles focusing on the patent domain [127].

A number of other symposia do not have proceedings, but publish the slides online. The Information Retrieval Facility[11] (IRF) has organized a symposium between 2007 and 2010, bringing together IR researchers and IP professionals.[12] The various patent offices also organize information events and training sessions.[13]

1.5.2 Sources of Data

Working in the patent domain implies having access to collections of patent data. Aside from those made available in the various evaluation campaigns already mentioned previously, patent data can be obtained directly from the patent offices, or from research collections.

While all patent offices make available patent data (it is one of their core responsibilities), it is not always easy to obtain it. The USPTO has its full text database available online[14] for manual search, and for bulk download via Google.[15] The European Patent Office[16] is one of the most active in this area, offering both researchers and commercial organizations free access to their data as well as data they have collected from other offices, via their Open Patent Services (OPS).[17] A fair-use policy applies in this case.

[11] http://www.ir-facility.org
[12] http://www.irfs.at
[13] http://www.wipo.int/meetings/en/index.jsp, http://www.epo.org/learning-events.html, http://www.uspto.gov/products/events/index.jsp
[14] http://patft.uspto.gov/
[15] http://www.google.com/googlebooks/uspto.html
[16] http://www.epo.org
[17] http://ops.epo.org

The Matrixware Research Collection (MAREC), first made available by the IRF, consists of approximately 19 million patent documents in XML format, covering four patent offices (USPTO, EPO, JPO, and WIPO). MAREC is now available for download under a Creative Commons license.[18] Further sources are the datasets, queries, and relevance judgments made available in the NTCIR patent track, and CLEF-IP and TREC-CHEM evaluation campaign tracks.

1.6 Structure of the Survey

The rest of this survey looks in detail at the various aspects of IR in the patent domain. After this introductory section, we continue in Section 2 with a detailed description of evaluation best practices in the field. We do this because, on one hand, evaluation is based on the understanding of search processes just described, and, on the other hand, because in this way we lay the ground for the discussions of results in future sections. The main focus of the survey is in Section 3, on text indexing and retrieval. We cover there both bag-of-words approaches, as well as those supported by Natural Language Processing (NLP) methods. Section 4 follows up on the text retrieval discussion with details on metadata associated with patent documents and how such metadata assists the search process. We cover both existing metadata, as well as experiments on creating new metadata.

Section 5 moves away from textual information and introduces the specific issues related to image and chemical structure retrieval in the patent domain. We discuss the importance of the information contained in the non-textual parts of the patent, as well as algorithms that have been developed to make use of this information in search.

Finally, we summarize the domain in the Conclusions section, and provide a set of research and development trends observed in recent years in relation to patent IR.

[18] http://www.ifs.tuwien.ac.at/imp/marec.shtml

2

Evaluation

Laboratory-style evaluation of search algorithms is a well-established practice in the IR Community [78]. This style of evaluation has been supported for IR in the patent domain through the organization of patent IR tracks in the CLEF, TREC, and NTCIR evaluation campaigns, with the corresponding availability of corpora, queries, and relevance judgments.

Throughout all the papers reviewed, numerical results of evaluation experiments are cited to demonstrate the quality of the new methods, or of their applicability to this particular domain. It is not new in IR to ask ourselves to what extent these numbers are significant for the practice of the end users [167]. In the patent domain however, the users play perhaps a greater role than in the web or news domain. The usual arguments for laboratory-style and user-based evaluations apply perfectly well to this domain, but it is important to realize that here the user is a person who will use the search system as the primary tool in order to do his or her job. In most cases, the user would have also attended some training on how to use a particular system most effectively. The decision to change to a new system will not be taken lightly and is not transparent at all. If in the case of a web search engine, the developer

may change the ranking scheme without necessarily informing the end user, changing the retrieval engine behind a patent search service may have significant practical and potentially legal consequences.

We have recently discussed this gap and complementarity between the practitioners' perspective on evaluation and that of the IR community [124]. This is not to say that the practitioners do not support evaluation campaigns such as those organized at NTCIR, CLEF, or TREC. On the contrary — such campaigns would in general not even be possible without the support and guidance of patent experts.

The question is how do we know what exactly is needed? What are the factors that lead to a positive assessment of a system? The answer begins with the word *"system."* If we look at the literature published by professional users assessing search services, it tends to mix together, within the same pool of judgments, distinct factors like collection coverage, search performance, efficiency, and usability [18, 52, 140]. The most common practice of assessing a system, or demonstrating its performance, is through the very "unscientific" way of showing an example. An example has the benefit that it is concise, easy to follow, and conveys a clear message. But this is hardly the way to make a decision.

In the PROMISE project, the research partners started from these examples and scenarios, and, through interviews and use case modelling, designed more appropriate, and yet still laboratory-style, evaluation campaigns. Some aspects remain particularly difficult to model. Trippe and Ruthven [189] for instance, answer the very fundamental question of *"What is success?"* with *risk minimization*. The argument is that what the searcher is ultimately trying to achieve is to minimize the risk of having a patent application rejected, infringing on someone else's in force patent, being excluded from a market because of a new patent, etc. They map the fundamental metrics of IR effectiveness, precision and recall, to the different search tasks and how they are ordered on the "risk scale." They observe that precision is more appropriate as a match, insofar as it orders the tasks in the same way as the importance of risk minimization.

This is to some extent surprising as the common mantra of Patent IR requires "high recall." It is however an expression that only in the

mind of the IR researcher is tightly connected to "low precision." For the professional searcher, as for any user for that matter, desirable is high recall *and* high precision (i.e., that the set of results is ordered by relevance [160]). What makes the user needs of a patent search system special, compared with a generaly search engine, is the lower tolerance to errors in this ranked list. In the practice of patent search, the way to achieve high precision and high recall is often through a set of high-precision queries. This may explain for instance the popularity that Boolean search systems still enjoy.

In this section, we provide an overview of approaches to both laboratory-style and user-centered evaluation in the patent domain.

2.1 Laboratory-style Evaluation

With patent related tasks in NTCIR, TREC, and CLEF, there is now a significant amount of data about how IR systems perform on such corpora and there are a number of test collections for the evaluation of future systems [58]. This section describes the approaches to obtaining relevance judgments and the evaluation metrics used in patent IR evaluation campaigns.

2.1.1 Relevance Judgments

For the patent domain, it is rarely feasible to obtain manual relevance judgments from patent search experts. With ever increasing backlogs, the patent offices have few resources to assist research, while private practice tends to be outside the budget of research centers. An alternative is to use "surrogate" relevance judges, for example the use of chemistry students in the TREC-CHEM campaign. More commonly however, approaches have been devised to make use of the citation information provided by patent examiners in search reports for the creation of relevance judgments.

With the exception of the NTCIR-3 patent retrieval task [93] and the Technology Survey (TS) task of TREC-CHEM 2009–2011 [126, 129], all ad-hoc retrieval tasks on the patent domain have used the examination report provided by the patent office as a source of relevance judgments.

The requests for information in the NTCIR-3 patent retrieval task started from a newspaper article and consisted, among others, of four text fields:

- **article**: the original newspaper article that generated the request for information
- **supplement**: additional information to the article
- **description**: a short description of the request for information
- **narrative**: a long description of the request for information

Therefore, the relevance assessments had to be created manually.

For the Technology Survey task of TREC-CHEM, the desire to include scientific articles meant that manual evaluations had to be done anyway, for even though scientific articles may also be cited in relation to a patent application or even a granted patent, there are technical difficulties in extracting and therefore using these citations, since there is no universal format for referring to non-patent literature [82]. For all other campaigns however, evaluation is based on citations, but only patent citations.

The use of citations from the examination reports is probably not optimal. The examiner is only required to find one relevant document which invalidates the claims in the application, even if hundreds may exist. This, added to the fact that practice and language familiarity may restrict the search of the examiner, results in a clearly incomplete relevance judgment set.

In a recent study [134], Magdy et al. looked at the number of words that a given patent application had in common with its citations. They observed that the percentage of applications which have no words in common with any of their cited documents is higher than the percentage of applications which have no words in common with the top five non-relevant (i.e., not cited) documents returned by a system. This may seem surprising, but it is part of the "features" of the collection. In fact, the study above does not take into account the fact that not all documents in patent test collections contain all sections, nor that the words used in a patent describe a technology which obtained a common

Table 2.1. Sources and types of citations for patent documents.

Source	Type	Relevance
Applicant		Prior work, which does not destroy novelty. Relevance level: 1
Examiner	A	Found by the examiner, but still not novelty destroying. Relevance level: 1
Examiner	Y	Partially novelty destroying. In combination with other citations may render the new invention obvious and thus not patentable. Relevance level: 1–2
Examiner	X	Novelty destroying citations. Will generally imply a change in the claims of the patent application. Relevance level: 2–3
Opposition		A citation introduced through an opposition procedure after grant. Relevance level: 2–3

name long after the patent was published (for example, *"touchscreen"* versus *"capacitive sensing device"*).

Documents cited in relation to a patent may come from different sources and these sources may come with assigned levels of relevance. Table 2.1 shows the different sources and types. The relevance levels mentioned in the table are for indicative purposes only, with a higher value indicating higher relevance.

The opposition citations are particularly interesting because they are, as a rule, extremely accurate. The opposition procedure is available at most of the major patent offices to account for the fact that the examiner might have missed significant documents. Practically, it crowdsources the examination. As competitors monitor the IP landscape in which they operate, they will put in effort to oppose a granted patent. At the same time, opposing a patent without success will strengthen the patent and make it much more difficult to litigate after a failed opposition. The procedure is therefore not taken lightly and the citations submitted by the competitor are, in general, extremely relevant. The problem in using them for an evaluation campaign is that they are comparatively rare.

To increase the number of citations, the CLEF-IP [161] and TREC-CHEM campaigns have used patent families to identify related documents to both the citing and cited documents. This is particularly useful if the task at hand is a general state-of-the-art search, where even tangentially related documents are interesting for the searcher. In terms of ranking systems however, the use of expanded citations with

patent family members has not been shown to produce significantly different rankings compared to those obtained in evaluations based only on direct citations [128].

When citations are not available, either because the topic is not a patent document (as in NTCIR-3) or the results are expected to be more than patents (TREC-CHEM TS task), manual relevance judgments must be created, and best-practices of standard evaluation campaigns should be followed [166].

2.1.2 Metrics

One of the questions in doing the evaluation is to what extent existing metrics apply to the new domain.

In general, the patent domain-focused evaluation campaigns organized so far have used well-known metrics. Practically all have reported results in terms of Mean Average Precision (MAP), and precision and recall at different levels. The different relevance levels have been taken into account by also reporting nDCG values (Normalized Discounted Cumulative Gain). The incompleteness of the citation lists has lead to the reporting of bpref (binary preference) values as well. For the invalidity task, a more appropriate metric might be Mean Reciprocal Rank (MRR) as, in principle, only one document is needed to invalidate a claim.

While all of these metrics are informative and useful, there has also been an attempt to introduce a more specific metric. The Patent Retrieval Evaluation Score (PRES) [133, 132] is a recall-focused evaluation metric, designed to map the recall-oriented perception of patent retrieval effectiveness into a numeric value. It takes into account the fact that a professional patent searcher is likely to go deeper into the set of results to find a relevant document, and therefore does not penalize as heavily as MAP or MRR the existence of non-relevant documents at the top ranks, but instead penalizes systems that return fewer relevant documents more heavily.

The problem with all of the metrics is that we still do not know how well they correlate to user satisfaction in this particular domain. A very small study was presented in [128], but the results were inconclusive because only 12 topics could be manually evaluated by experts.

2.2 Humans in the Loop

User-based evaluation is important in any IR evaluation study, but notoriously difficult to do. The problem is intensified in this domain, where privacy and confidentiality is paramount, hourly costs are in line with lawyers' practices, and communication between end users and researchers is limited and difficult.

Despite these difficulties, there have been initial steps, both in including users in the evaluation of laboratory-style campaigns and in designing interactive campaigns such as the PatOlympics [125].

The PatOlympics invited research teams to present their patent retrieval systems in a competitive-demo session, where patent experts use the systems to answer a particular request for information. The comparability of the results in the end is supported by the provision of a common corpus for all participants, the equal opportunity to interact with the patent expert, and the pooling of results identified as relevant during the event itself. The disadvantages are: first, the expert's perception of the topic changes as he or she sees more documents, so if a team's turn to interact with an expert is at the end of the event, the expert will be more likely to know exactly what to look for. This is mitigated by the fact that the teams and experts interact in a round-robin fashion, so all teams will experience both the advantage and the disadvantage of this change in an expert's familiarity with the topic. Second, and most significant, the number of topics is extremely limited (2–5 topics) due to the limited time available for such an event.

Finally, let us observe that humans are in the loop in the adoption of techniques in practice. While getting insights into internal company systems is practically impossible, recently some systems based on published research have been made available to the public [85]. Their success, and implicitly evaluation, remains to be decided by the users they attract.

2.3 Summary

The definitions of success given by the IR researchers and the IP professionals differ. They are not conflicting, but complementary. The patent

domain is not unique in its need to consider the needs of the users, but in this case, the users are knowledgeable about the data, the search systems and even the expected results [16]. If in web or newswire search each of us can, with reasonable accuracy, imagine what the user does and needs, patent searchers are trained professionals, whose search results are difficult to replicate with *ad-hoc* search. That being said, the need to clearly identify the real performance of core components is as present as ever, and there is a huge number of resources at our disposal. The examination reports, performed at patent offices around the world, are invaluable. Only the peer-review process of scientific literature comes close to it, but it lacks the structure and unique global identifiers that the patent system has.

The importance of users in this domain has lead to systems being evaluated by comparison with best practices in the field. This opens up a difficult modelling problem: changes and improvements in search technology, as triggered by evaluation campaigns, lead to changes in best practices. The users are always right, but a balance is to be found between a tradition of best practice, and innovative practices.

3

Text Retrieval

Text retrieval is the core business of many in the IR community and consequently this particular approach to patent retrieval has seen a lot of attention in the past 10 years. This section discusses the differences between "general" text retrieval problems (and we refer here to web and news corpora examples) and those in the patent domain.

Before we begin, we need to observe that full text search is a relatively new experience in the patent retrieval systems. The incumbents here are manually created indices of semantic data and it is important to understand the difficulties that a full-text search system has to face in improving existing practice [9, 10]. The manually created indices, such as summaries, keywords or chemical lists, cannot cope with an ever growing amount of data, but from the user's perspective, they are extremely important because they provide a level of quality still unmatched by automated processes. The requirements for Information Retrieval systems are therefore very demanding, particularly in the areas of query syntax and information extraction.

This section begins with a discussion on the differences between patent and non-patent literature and of the challenges in the processing of patent literature. This is followed in Section 3.2 by a review of current

approaches to processing and indexing patent documents. Finally, the important topics of query syntax and multilinguality are discussed in Sections 3.3 and 3.4, respectively. We should note that although it is technically still text, the extensive metadata associated with a patent and its use in patent processing and search are not discussed here, but in Section 4.

3.1 Patent versus Non-patent Literature

The motivation of this survey is the need to process patent documents differently from other collections. In the first section, we have seen the background of this problem (i.e., the legal patent system and its motivations). We have seen that this background generates specific search tasks which define what "patent retrieval" means for the practitioner. This section explores the effects of this background on the text itself and, indirectly, on the Information Retrieval researcher and the methods he or she develops to tackle the problem at hand. We compare the text of a patent collection with that of a general corpus consisting, for example, of news or web data.

3.1.1 Term Distributions

The largely empirical methods of Information Retrieval rely heavily on assumptions about the relationship between the frequency of occurrence of terms in a document and the topic of the document. It has often been said, and there are many anecdotal examples, that patent documents are more "difficult" than other documents used in IR tests. We start by looking at the collection term frequencies. It is well known that term frequencies follow a Zipfian distribution, and that the plot of frequencies against ranks on logarithmic axes should be, approximately, a line. Figure 3.1 shows this plot for the CLEF-IP 2011 corpus [157] and for the American National Corpus (ANC)[19] [88]. For the latter, the frequency figures are provided by the corpus creators online. Although the figures vary between the two corpora,[20] the plot is made such that

[19] http://americannationalcorpus.org/

[20] Technically, the CLEF-IP is not a corpus, but just a collection of documents, since it does not have any linguistic annotations.

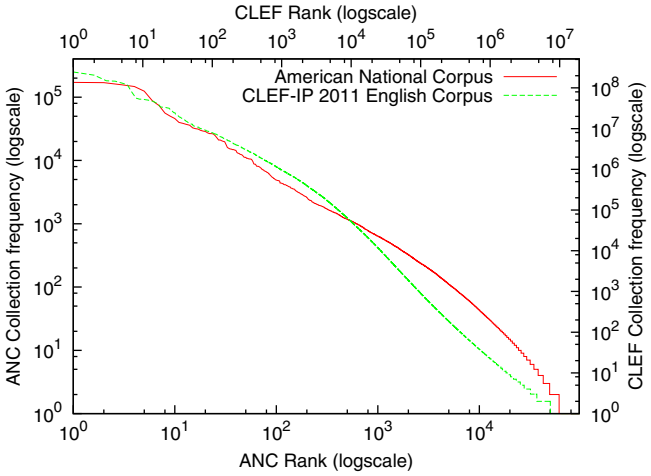


Fig. 3.1 Term frequencies in patents versus standard American English.

the two extremes (top left corner and lower right corner) match. They appear to be similar, but we should not forget that these are logarithmic plots. What matters here is the second derivative of the curves: the rate at which the frequency changes. Compared to a corpus of "regular" American English, the patent corpus decreases the frequency much faster. This means that it is rich both in high frequency and low frequency terms, but poor in average frequency terms. One of the reasons for this is the presence of text obtained by Optical Character Recognition (OCR) in the corpus [185] and the errors introduced by the OCR process, but also the nature of the text. As is often the case in legal texts, the authors of the Claims sections do not shy away from repeating the same term over and over again, wherever deemed necessary to clarify the object or procedure being claimed.

The second aspect to observe is the specificity of the terms used in this corpus. Given the relationship between term specificity and document frequency (DF) [96], we can make an inference about the specificity of each section by looking at the average DF of the terms within it. Figure 3.2 shows the distribution of the average document frequency of terms in the three sections of a patent document, as a percentage of the total number of sections of that type. For comparison, the figure also shows the average DF in documents of the Open American National

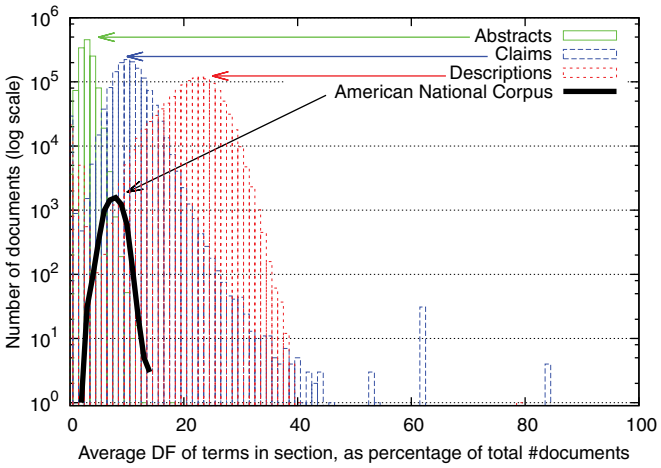


Fig. 3.2 Average Document Frequency distribution, per sections(s).

Corpus (the free version of the ANC). Extreme cases are visible in the patent data, such as those four documents whose *average* DF is 80% of the total number of documents. Even without these extremes, most Description sections have an average DF of over 25% of the size of the corpus.

3.1.2 Natural Language Processing Issues

In addition to the issues related to the frequency and distribution of terms, a text processing method may need to take into account the structure of the sentences, the frequency and distribution of noun phrases (NP) rather than terms, and the frequency of out-of-vocabulary terms.

Among the research done in this area, most emphasis was placed on the Claims section of patents. This has two reasons: first, the claims are the legally binding description of the invention; second, because they are legally binding, they are written in the "patentese" mentioned in the Introduction. Claims tend to have longer sentences, a complex structure and use uncommon terms [173]. While these facts about claims have been mentioned consistently in numerous publications, numerical observations about them are given in [150, 197].

Verberne et al. [197] used a set of 67,292 Claims sections extracted from the 400k patent document corpus made available for the AsPIRe workshop [76], to which they applied a sentence splitter (using both the full stop and the semi-column as delimiters), followed by the AEGIR dependency parser [106]. They compared their results with statistics from the British National Corpus (BNC) [112]. They found that the average Claims sentence has 53 words, while the median is 22. Comparing with the BNC, where the peak of the distribution is at 10 words, patent claim sentence length distribution peaks at 20 words per sentence. To put a figure on the out-of-vocabulary issue, they compared the coverage of the patent corpus by the CELEX lexical database [22]. In terms of token coverage, this was found to be better (96%) than a general corpus (92%). However, in terms of types (i.e., unique tokens), the coverage dropped to 55%, suggesting that there is a high token-to-type ratio. This observation was later detailed by the authors in a follow-up article [150]. We will come back to this shortly. This issue is not specific to English or western patenting practices. A similar observation was made about Japanese claims by Shinmori et al. [173], though without an extensive study. An extensive comparison of issues between English and Japanese in patent practice and language is presented by Lise [117].

Interestingly, Verberne et al. did not find significant differences between the lexical frequencies of ambiguous words (in this case, words which can be assigned multiple parts of speech). Instead, they found that multi-word terms in patent claims tend to not appear even in specialized dictionaries (*"we found that fewer than 2% of the two-word NPs from* SPECIALIST *occurs in the* MAREC *subcorpus"*).[21] This again supports the intuition that patent claims, by definition, describe in existing, common words, new things, the expression of which is therefore not found in existing dictionaries.

A third challenge mentioned by Shinmori [173] is the complicated syntactic structure of the claims. While the results here are obtained from a small set of 100 short sentences, and therefore are inconclusive, Verberne et al. argue that the main cause of failure for syntactic

[21] The SPECIALIST lexicon is a lexicon covering both common English terms and biomedical vocabulary [37].

analyzers is that the claims are not mainly formed of clauses (i.e., do not always have a subject and a predicate), but rather noun phrases. This is a feature of patent practice because of the restriction of the USPTO that the claims have to consist of a single sentence [2]. The resulting text is not always easy to read, and often difficult to process automatically. The subject and predicate are often in the "title" of the claims section, which can be *"I claim:"* or *"We claim:"*.

3.1.3 Document Length Issues

Finally, we observe that patent documents tend to be quite lengthy in nature. Iwayama et al. [92] use the test collections of NTCIR-3 to compare a patent corpus consisting of Japanese patent applications and a news corpus consisting of Mainichi newspaper articles. Their results show that the patent documents have approximately 24 times as many terms as news documents. At the same time, the standard deviation of the distribution of lengths for patent documents is 20 times that of news articles.

Length normalization methods have, however, been available for over 15 years, and a version thereof is part of the standard BM25 model (for a full discussion, see [159]). Iwayama and colleagues do not in fact observe a difference in the relative performance of the nine retrieval models they try on the two collections. The question is therefore only if there are ways to better manage huge documents in order to increase the absolute retrieval scores. We will examine this problem in more detail in the next section on patent document processing, but it is worth taking a moment to understand why patent documents are longer.

As discussed in [159], increased length of a document may come from one of two sources: either the document talks about a unitary topic, but is verbose in doing so; or the document covers more topics. The assumption in patents is that they are simply verbose. This is supported by the idea that a patent covers one invention and therefore one topic. However, we should not forget that what exactly an invention is may be differently understood at different patent offices. For instance, the USPTO allows an application to be amended by the applicant with a

so-called *Continuation-in-Part application* (CIP).[22] The EPO does not have such an option, and in this case the applicant is generally advised to file a new application altogether. Additionally, each examiner may decide whether a document describes a single invention or not, in which case it is possible that one invention is assigned one patent by one office and multiple patents by another. This means that, despite the intuition that a patent should cover a single invention, and that therefore the issue of long patents is only related to the verbosity of the text, we do have to consider the other variant as well. Namely, we have to take into account that a patent document may be discussing sufficiently different topics to consider each as a separate piece of information.

3.2 Patent Document Processing

Based on the observations above, in this section we will discuss approaches to process the documents to improve retrieval. In the course of doing so, we will mention the results obtained by various research groups, but the reader should note that often these are based on different collections, and are therefore rarely comparable. Also by processing, we understand here weighting in the case of Vector Space Models (VSM) or Probabilistic Models (PM), or model creation in the case of Language Models (LM).

We follow, in large, the structure of the previous section, focusing first on bag-of-words approaches, followed by NLP-based approaches.

Before we begin however, we should note that state-of-the-art Information Retrieval models are not easy to improve upon. In [196], Verberne and D'hondt use the text of the claims in an off-the-shelf Lemur instance using TF*IDF, and rank 6th and 35th in the 70 runs submitted to CLEF-IP 2009, according to normalized Discounted Cumulative Gain (nDCG) and Mean Average Precision (MAP), respectively. Figure 3.3 shows the result obtained by an equally off-the-shelf Solr/Lucene instance in the CLEF-IP 2011 test collection (outside of the officially submitted runs). Both the Lemur/Indri and Solr/Lucene results show that even a "basic" IR instance can achieve an average, or

[22] A CIP application follows up on a previous application and is generally used to protect enhancements of a previously disclosed method or device.

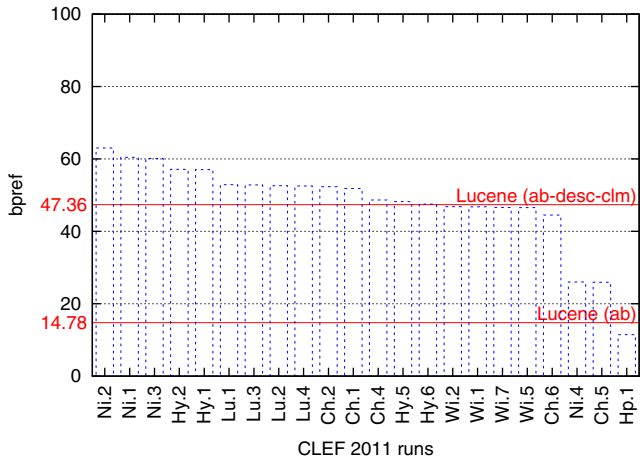


Fig. 3.3 Solr/Lucene performance on the CLEF-IP 2011 test collection, when using the full text, as well as just the abstracts. The other runs (seen here as columns) are here for illustrative purposes only, and therefore without full names.

above average performance. This figure shows the bpref value (similar plots had been obtained for MAP), which is appropriate for the search report-based qrels used in the patent domain, as described in Section 2.1.2.

3.2.1 Bag-of-words Approaches

Before we talk about words, we need to mention tokenization. Most articles cited in the following do not identify specific changes to the tokenization procedure, although this is arguably important. Particularly in chemistry patents [129], the tokenization procedure must take care not to include artificial tokens extracted from textual representation of chemical formulae into the general index.

Iwayama et al. [93] are the first in presenting a study comparing different weighting techniques and their results as applied to patent and news corpora. Table 3.1 shows the nine scoring functions used[23] (common notations, see [93] for details).

[23] It is unclear why BM25F was not used in this case, as fields are always present in the target documents.

Table 3.1. Retrieval models analyzed by Iwayama et al. [93].

Model	Weight
hits	$b_{q,t} \times b_{d,t}$
baseline	$f_{q,t} \times b_{d,t}$
tf	$f_{q,t} \times \frac{f_{d,t}}{dlf_d}$
idf	$f_{q,t} \times idf_t$
tf.idf	$f_{q,t} \times idf_t \times \frac{f_{d,t}}{dlf_d}$
log(tf)	$(1 + \log(f_{q,t})) \times \frac{1+\log(f_{d,t})}{1+\log(\text{ave}f_d)}$
log(tf).idf	$(1 + \log(f_{q,t})) \times idf_t \times \frac{1+\log(f_{d,t})}{1+\log(\text{ave}f_d)}$
log(tf).idf.dl	$(1 + \log(f_{q,t})) \times idf_t \times \frac{1+\log(f_{d,t})}{1+\log(\text{ave}f_d)} \times \frac{1}{\text{ave}dlb+S\times(dlb_d-\text{ave}dlb)}$
BM25	$f_{q,t} \times \log\left(\frac{N-n_t+0.5}{n_t+0.5}\right) \times \frac{(K+1)\times f_{d,t}}{K\times\left((1-b)+b\frac{dlf_d}{\text{ave}dlf}\right)+f_{d,t}}$

There are no surprises in the results: models that perform well on news corpora (BM25, log(tf).idf.dl) perform well on the patent corpora too, relative to the other models. Also to note from their experiments is that using the full text of the patent documents in the search process, rather than restricting to specific sections (claims, abstracts)[24] improves results. This was also subsequently observed in [50, 61, 68]. The only reversal of ranking that grabs attention is that when using the original article that generated the topic (see the description of NTCIR topics in Section 2.1.1), aided by the "supplementary memorandum" added by the topic creators, the ranking of the top two performing models is reversed. In fact, it decreases the performance of BM25 by almost 40%. The difference must lie in the fact that while the article and supplementary data were not written with a search task in mind, the description and narrative, added by Iwayama and colleagues, was more specific.

Fujita [61] continues this analysis by comparing a BM25-variant weighting with a language modelling approach with different smoothing methods (Dirchlet or Jelinek–Mercer) and priors (document-length or IPC[25]). This particular study focuses on the effects of the document

[24] The article does not show results for using the description section of the patents.
[25] International Patent Classification, see Section 4.2.

length on the scoring of the retrieval model. In the end, Fujita concludes that in both cases (probabilistic and language model), retrieval is improved when the model penalises long documents. In the case of the BM25, this is done by setting the b parameter to larger values (0.9 to 1 is suggested here for the patent domain, compared to just 0.3 to 0.4 for news corpora).

A number of subsequent systems have used TF*IDF [74, 83], BM25(F) [28, 56, 68, 136, 154], language modelling [95, 134, 215], or a combination thereof [214]. Comparing them directly is not always possible. An exception is made by the systems participating in or using the collections of evaluation campaigns. Even in these cases however, they most often combine some form of similarity metric or term weighting, with some form of query generation. We discuss the latter in Section 3.3.

3.2.2 Latent Semantic Analysis

Before moving forward, we also mention distributional semantics, of which Latent Semantic Indexing (LSI) is the poster-child. The aim of such technologies is to discover patterns between terms, or between terms and documents, and use them in the ranking process.

Latent semantic indexing is known to improve retrieval effectiveness [62], but in the patent domain its effects are still unclear. While there exists at least one provider who uses it [163],[26] LSI has not been shown to improve results in the patent domain. In fact, there are few studies in this sense. Chen et al. [43] observed that their system, based on Differential LSI (DLSI), a variant proposed earlier by them, does not perform well and suggest a scheme to use DLSI as a first-level filtering scheme and apply a dynamic programming method to find syntactic templates matching the query. No effectiveness results are presented here, but a later study confirmed that LSI did not improve results compared to the standard VSM [139]. Instead, Latent Semantic Analysis may find its way as an assisting technology, rather than relying fully on its ability (or inability) to detect document similarity [64].

[26] At the time of writing of this review however, no reference to LSI was to be found on the company's Web site (http://www.freepatentsonline.com).

A possible problem with distributional semantics methods, as well as with the bag-of-words approach in the patent domain, is what exactly should be considered as a word. As mentioned in the previous section, at least in the Claims section, but also in the other parts of a patent document, multiple, common words are used together to define a new concept. Tseng et al. [190] propose a method to generate word n-grams by traversing backwards through the text after tokenization and keeping maximally repeated sets of words. A set is taken to be maximal if none of its ordered subsets of length at least two occurs more frequently than itself. The proposed processing also includes the use of stop-words and stemming, but it is unclear which stop-words were used. In general, a domain specific stop-word list would be preferable [33].

The expansion from bag-of-words to larger indexing units is particularly tempting in this domain. As D'hondt et al. [51] observe, this may become too specific and damage recall. Nonetheless, given that syntactic structures are, as we observed in the previous section, some of the most distinctive characteristics of patent documents, the next section lists some of the approaches in this sense.

3.2.3 Natural Language Processing

The use of NLP techniques to enhance retrieval is certainly not an attempt specific to the patent domain. The inherent problems of the bag-of-words model and the potential of NLP have been discussed at length. We found the tutorial of Smeaton at ESSIR'95 useful for an overview [176]. This particular section deals with the applications of such processing for the purposes of indexing (e.g., tokenization) and weighting.

Given that parsers are even less likely to provide reliable results on the complex syntactic structure of patent documents and in particular of patent claims, the question is what to expect of NLP techniques in this domain, and in particular for retrieval. This being said, experiments at the University of Hildesheim [28] do show improvements in recall when using phrases rather than terms, but still are outperformed by other systems not using phrases [158].

A central focus in the research community is on processing the claims section of a patent, as that is where one has both the legally binding content and the most differences with respect to newswire or web corpora. The purpose is, on one hand, to improve retrieval, but often to simply make the claims readable to the public without a degree in Law. Surprisingly enough, one of the first research publications in this area handled not the problem of understanding the written claims, but rather that of writing them. As Sheremetyeva indicates, *"the difficulty of the task is not constrained only by syntax and style. A claim must be composed so as to make patent infringement difficult"* [171, 172]. This means that the claim must be neither too restrictive, nor too broad, neither too clear, nor too opaque. The existence of such an automated system to generate claims would indicate that perhaps the claims are still not as difficult as commonly thought. If they can be generated automatically, maybe they can also be reverse-engineered to the core concepts. Unfortunately, reading just a few claims is enough to understand that this is not the case. While the patent claim generation system that resulted from this research is available commercially as AutoPat,[27] it is unclear to what extent it is adopted, and to what extent an applicant would have to change the output of the system to adapt it to their needs.

Making the claim text more understandable to the reader involves two aspects. First, the structure of the claim section needs to be made visible. Second, terms may need explanations.

The claim section of a patent generally includes several claims, which are often connected to each other in a dependence relationship. As mentioned in Section 1.3, a claim is said to be *independent* if it does not make reference to another claim and *dependent* otherwise. The process of uncovering the structure of the claim section results in a tree (potentially a forest), with independent claims as roots, and dependent claims as intermediary or leaf nodes. This kind of structure leads to the use of rhetorical structure theory in order to link cascading definitions [170, 173]. Furthermore, this tree can be refined by using claim-specific cue words in order to recognize lists or

[27] http://www.lanaconsult.com/index.html

relationships between terms. Both Shinmori et al. [173] and Sheremetyeva [170] use such keywords. Shinmori first identifies six types of relations in claims (PROCEDURE, COMPONENT, ELABORATION, FEATURE, PRECONDITION, and COMPOSE) and then identifies for each a set of cue words. Sheremetyeva first identifies the cue words and then associates a range of morphological, semantic and syntactic properties to them. In both cases, a combination of dependency parsers and context free grammars is used to generate the enriched claim structure.

Later, Parapatics and Dittenbach [153] took advantage of developments in the area and used an off-the-shelf parser [49] to obtain a high percentage of successful parses, after a claim segmentation process similar to the ones just mentioned. In the process, they describe a very comprehensive list of types and patterns in claims.

As mentioned, the utility of such techniques is not found in retrieval, but rather in making the claims understandable to the common reader. While all the articles of the previous two paragraphs mention paraphrasing and summarization as applications, the articles do not show specific results. Bouayad-Agha et al. [36] make this step in the context of the PatExpert project.[28] Alternatively, the conceptual graphs extracted by the methods just described may be compared automatically, as in [211]. The utility of this latter approach for the patent users is however still to be determined.

NLP methods may also be used to identify more precisely terms of interest. Such terms may carry particular information, and may need further explanation for the reader. In relation to the claim structure detection method, Shinmori et al. [173] also propose the use of the description section to provide context to specific terms. They use a fairly intuitive method based on heuristic rules to identify the terms, and keyword search on the description section to extract context.

One of the issues observed in attempting to process claims using parsers is that a specific lexicon is needed. Sheremetyeva had generated one, but did not provide details of coverage [170]. Oostdijk et al. [151] create such a lexicon semi-automatically, and show that it covers 86.5%

[28] http://www.patexpert.org/

of the types found in the MAREC patent corpus, compared to only 60.4% coverage obtained by a general purpose lexicon.

3.3 Query Creation, Modification, and Enhancement

Regardless of how well the indexer manages to process the patent collection, the ultimate results will also depend on asking the right questions, or, in this case, on generating the most effective queries. The nature of the domain, where a typical use case scenario implies beginning the search upon receiving a patent application, increases the necessity to process the request for information (e.g., a full patent application, an innovation disclosure) in order to be able to efficiently search the index.

Sometimes however, some documents may not be retrieved at all, by any query. This is the study of *Findability*, first introduced to the patent domain by Azzopardi and Vinay [21] and expanded by Bache and Azzopardi [23]. The idea is that the index of a particular document may be such that, regardless of the terms used in the query, there will always be other documents obtaining a higher score, given a specific ranking function. The relationship between the queries and the different retrieval methods is explored in depth by Bashir and Rauber [27].

As a rule, all papers describe some form of query generation when discussing patent retrieval. Even the off-the-shelf Lucene implementation used in the generation of Figure 3.3 incorporates within its "More Like This"-functionality a term selection mechanism. It is clear that query generation is significant, but there are only a handful of studies where the matter is specifically targeted. For the rest, we can hypothesize that the query generation method played a specific part in the more or less successful runs, but it remains impossible to tell which component of the retrieval system had the most significant impact on the results.

We can hypothesize, for instance, that some approaches may be too simplistic. For instance, Wanagiri et al. [202] use the top 10 keywords according to the TF*IDF version of Salton and Buckley [165]. This appears to be provide too little information and most likely also too corrupted by the direct use of the term frequency in the weight. We can

learn more from Xue and Croft [208, 209] who cover more weighting schemes and different fields. In particular, they find that the optimal query length, for their test collection, is around 30 terms. However, since the collection used is not a standard one, results are not comparable. At the same time, Xue and Croft show that using a particular field of the US patent format, the *brief summary of the invention*, achieves the best results, followed by using the full text. This field is unfortunately only a feature of the US patent applications.

Prior to these, Itoh et al. [89] attempt to weight terms in the query based on their perceived significance in the target corpus, combined with their significance in the query. To measure this significance, different term weighting schemes are used. The proposed method does not take into account the fact that some terms, while being important to the definition of the request for information, may not necessarily appear in the target set at all. For query term selection purposes, it would seem more useful to weight them based only on the genre to which the query belongs, rather than the genre of the target collection.

An enhanced version of this is to use language models for both the query and the target collection. Perez-Iglesias et al. [154] select the query terms based on the difference between the language model of the topic document and the collection. They consider of interest the terms which appear frequently in the query, but not so frequently in the collection. Results are average. A similar approach was taken by Mahdabi et al. [136], although here the language model was more refined.

The problem with having a full document as a query is that, as described in Section 3.1.3, it might refer to multiple topics. Even in the case of a single invention, different components of the new device or process may have their own prior art and generating one query for all may produce poor results for all. This is particularly important in the case of technology survey tasks, where all aspects of a field are requested (though arguably, in this case, the queries will be generated by humans). Mahdabi et al. continued their study using language modelling for the query and collection with a study on applying a summarization technique, a version of TextTiling, and generating a model for each summary [135]. The results reported are however inconclusive. The problem is that splitting the request for information and recombining

the search results obtained for each such sub-request introduces a new set of parameters to be investigated. The subtopics discovered in a patent application may be of different significance. It is perfectly reasonable that the Description section talks in detail about existing work, but none of the relevant documents for such a subtopic is likely to be cited in the examination procedure as invalidating the application at hand. Takaki et al. [182] show that with different weights assigned to the different topics, the retrieval results can be improved. Even without weighting however, Ganguly et al. [63] obtain good results by segmenting the Description section using TextTiling, as Mahdabi before, but also using the Title, Abstract, and Claims sections as additional subqueries. The difference here is that they add, for each of the subqueries, a pseudo-relevance feedback loop. We will come back to this shortly.

We return briefly to the use of natural language processing techniques, discussed in the previous section, to note that while all the query generation methods mentioned so far rely on terms, encouraging results have been recently obtained using phrases. An early result was presented by Osborn et al. [152] in the late 90s, when they showed that using an NLP-based grouping of terms results in a performance increase compared to the bag-of-words approach, even if the increase is smaller than in a non-patent collection (i.e., Tipster in their experiments). Later, D'hondt et al. [51] use a query term extraction based on triples generated after a dependency parser was run on the topic document. Verma and Varma use key-phrase extraction techniques based on the results of a part-of-speech tagger [199].

A less strict version of the query generation methods described above would be to simply allow the user to select the terms [116]. Even more, complex Boolean queries may be automatically suggested to the professional used to this form of interaction with the system [100].

Finally, there are those methods that not only use the request for information or the target data to select a proper query, but also the results obtained in a first try, i.e., pseudo relevance feedback methods. Adaptations of the original Rochio method, as well as a new method, capable of taking into account the degree of relevance, are applied to the NTCIR-3 test collection [101]. The results do not show an improvement in retrieval performance in this case. This may be traced back to the fact

that patent retrieval does appear to produce less good results than a retrieval method applied on a general domain, since a pseudo-relevance feedback (PRF) method will only enhance what is found in the top retrieved documents. However, it may be that the documents are too long and cover too many topics, as discussed in Section 3.1.3. This is suggested by Ganguly et al. [63], who first use a query segmentation procedure, based on TextTiling, similar to that of [135], and apply PRF on the results of each sub-query. Then, they merge the results via round-robin. In this case, the PRF method is also quite standard, so we can assume with some degree of confidence that the significant improvements shown are due to the more precise queries. A good survey of query expansion techniques, including PRF, as applied to the patent domain, is presented by Magdy and Jones [131].

A final note on a method not reviewed in the above survey, but still of interest, despite the poor results. Sahlgren et al. [164] used Random Indexing to identify terms to use for query expansion in Japanese. While the method had performed well in non-patent Cross-Language Information Retrieval (CLIR), it is unclear still whether its failure in the context of the NTCIR patent retrieval task comes from the patent domain or from the lack of familiarity of the authors of this particular study with the Japanese language. We move on to the issues of multiple languages next.

3.4 Multilinguality

As mentioned in Section 1, the patent system is an essentially multilingual one. Hence, machine translation and cross-lingual Information Retrieval (CLIR) are essential. This is reflected in the focus given to CLIR in two evaluation campaigns (NTCIR and CLEF-IP). As a result, there are a number of systems available for searching patents in different languages. Also as a result, there exist lexicons and parsers dedicated to the patent domain in several widely used languages, such as English, Japanese, Spanish, French, and German [36, 151, 173]. Essentially the problem of handling different languages is not a patent specific one, but the specific genre (see Section 3.1.2) is an additional factor on top of everything else that has been discussed in the literature [149].

So far, language specific studies are extremely limited. Among non-Asian languages, one of the features that exists only in some languages but not in others (notably, not in English) is compound words. German, Dutch, and the Nordic languages use such words, which consist of more than one constituent (e.g., *"Patentamt"* (DE) — *"Patent office"* (EN)). Using decompounding is found to help retrieval performance both for Swedish [17] and German [113]. Jochim et al. [95] compare the use of phrase and term translations for retrieval, and find that the first improves French topics, while the second German ones. The explanation is that without decompounding, terms in German have the informational content of phrases in English or any other non-compounding language.

The familiar approach of query translation is extensively used in the patent domain as well. Fujii and Ishikawa use it for Japanese-English CLIR [56] in the context of NTCIR-3. Additionally, they have used patent families to return a document different to the one actually retrieved, but within the same area of technology. In fact, the use of patent families for creating parallel or comparable corpora for statistical machine translation has found a considerable appreciation [122]. A summary of all other cross-lingual experiments at NTCIR-3 is in [102].

However, all of these observations would potentially apply to any other domain, and a comprehensive book on CLIR provides context to all of them [149]. What is specific to the patent domain, but has not been thoroughly analyzed in any of these works, is how good the translation is. Statistical Machine Translation (SMT) systems are only as good as the training data they are trained on, and they must be adapted to the patent domain [41]. But perhaps a full-blown SMT is not necessary for query translation. After all, what is needed is not necessarily a properly formed sentence, but the correct keywords. Such a bilingual thesaurus may be learned directly from the corpus at hand, using hypernym and hyponym relations [148].

Query translation is not the only option for cross-lingual retrieval. Recently, Li and Shawe-Taylor presented a study, also on the NTCIR-3 collection, using Kernel Canonical Correlation Analysis (KCCA) [115]. The retrieval results show indistinguishable MAP values compared to monolingual runs.

Finally, let us note that translation may not necessarily be between languages, but also between genres [147]. The same method used by Nanba et al. [148] to learn a bilingual thesaurus is used again by the team, in conjunction with references between patents and scientific articles, to learn the different ways of expressing the same thing in the two different genres. For instance, their method is able to learn that a "TV Camera" may be referred to in a patent as a "photographic device," "image shooting apparatus," or "image pickup apparatus."

3.5 Summary

From the use of more-or-less off-the-shelf IR models, methods or tools, we have seen that the bag-of-words approach, without any enhancement related to metadata present in the patent, stands its ground. What seems to consistently work is using the full text of the document, but using different fields (using here the Lucene terminology) for the different sections. This means that the term weighting is done per field and this makes perfect sense considering the different genres in the different sections of the document. Additionally, the studies presented in this section have provided ambiguous results in terms of the utility of using metadata (IPC, inventors, assignees, etc.) in the search process. However, this deserves a more careful analysis, and we look in more depth at the use of metadata in Section 4.

Latent semantic methods, although apparently used in a few commercial services, show inconclusive results. Given their success in other IR domains, their study should be encouraged in order to clearly assess their utility in the case of patent search.

In terms of query generation, there are positive results that significantly improve retrieval performance, but also negative experiences. Interestingly, Zhao and Callan [216] indirectly reflect on one of the fundamental properties of patent retrieval systems: the experience of the professional searcher. They describe how Indri performed better using automated queries than using manual queries generated by an inexperienced user. While the study is only on the six topics of the Technology Survey task of TREC-CHEM 2011 [126], and an additional one from PatOlympics [125], it shows both that a state-of-the-art search system

is able to assist the casual searcher, but must allow the experienced one full flexibility.

The most visible difference between patent and non-patent literature was observed in the application of NLP methods. The studies mentioned in this section indicate that the nature of the genre (using many common words, but in new combinations, together with relatively many hapax terms) is a probable cause of the perceived reduced effectiveness of both bag-of-words and distributional semantics methods, when compared with results on web or newswire corpora. Such linguistic studies are unfortunately rather sparse, as they require a significant amount of manual annotation and text analysis, together with potentially large computational resources.

In fact, an issue that was not directly addressed in this section is that of scalability. The global set of patent documents is large. Estimates for the total number of patent documents available worldwide are difficult to make, but commercial collections claiming worldwide coverage range in the area of 70–90 million documents.[29] This is not larger than the web collections that the IR community is currently handling, and scalability becomes a particular issue here only insofar as the method at hand is running significantly more complex algorithms than those of standard IR. For instance, Klampanos et al. [103] and Urbain and Frieder [193] propose distributed systems in order to manage more complex processing: probabilistic logic in the first case, and chemical information extraction in the second. These are problems that are not specific to the patent domain and which receive a well deserved large share of attention in the community.

[29] Precise numbers are not available, and these are estimates we obtained in private conversations.

4

Metadata

One of the main characteristics of patent search is the availability of a rich metadata, created in the granting process by both the applicants and the examiners. We start with a description of this metadata, and of the uses one can make of it for retrieval. We then continue with a brief survey of technologies which assist in creating additional metadata, such as classification and information extraction. We conclude this section with a section on visualizations for patent retrieval.

4.1 Existing Metadata

In Section 3, and particularly in Section 3.2 on patent document processing, we have presented a number of systems that use various IR models to calculate similarities between queries and documents. We have carefully avoided however discussing those enhancements which use metadata in this process. We come back to this now and start by showing in Table 4.1 a list of common fields generally present in all patent documents. This kind of information is generally not present in general purpose IR collections, and has been indicated by patent search professionals to be very useful in patent retrieval.

Table 4.1. Metadata generally associated with a patent document.

Field	Explanation
ID	Unique, global identifier
Family ID	An identifier for the family to which the current document belongs
Publication Number	Potentially the same as the ID, unless the emitting authority has a different practice
Publication Date	Date when the present document was made public
Application Number	Many authorities assign a number upon receiving a patent application, and another number to the published document describing that application
Application Date	The date when the application was received
Priority	One or more patent IDs indicating prior applications for this invention, according to the Paris Convention
Earliest Priority Date	Earliest application date of the set above
IPC	International Patent Classification Label
ECLA/F-terms/other	Other classifications, as assigned by national patent offices
Title Language	Language of the title
Patent Citations	List of patent IDs cited by the applicant, examiner or in an opposition procedure, in relation to the current document
Non-Patent Citations	As above, but for non-patent literature
Applicant Name	Name of the individual or company that applied for the patent
Assignee Name	Name of individual/company owning the patent
Inventor Name	Name of the individual(s) who invented the object or process disclosed
Agent Name	Name of the individual or company acting on behalf of the above
Abstract Language	Language in the abstract
Description Language	Language in the description
Claims Language	Language in the claims

A very simple way to increase effectiveness in the case of Novelty or Validity searches is to make sure that, unless the collection already restricts the date range we are searching on, we do it ourselves [215]. For validity search in particular, considerable attention must be paid to when the potentially invalidating document was published. Any document, regardless of its topical relevance, will be irrelevant if made public after the earliest priority date of the patent whose validity we are now verifying.

One of the most complex, but also most useful, metadata are the classification fields (IPC, ECLA, etc.). These are covered in more detail in the next section.

4.2 Classification

Patents are classified by the patent offices into large hierarchical classification schemes based on their area of technology. The use of patent classification has two major benefits [6]. The first is that the classifications provide access to concepts rather than words, such that even if the same word or phrase is commonly used in two technology areas, patent classifications will provide the context of its use. In effect, they allow the search space of patents to be reduced, by allowing the user to exclude from the search process patents in classes not related to the search topic at hand. The second major benefit is the language independence provided by classifications, as classification symbols can be mapped to multiple languages. This allows patent searchers to conduct reasonably effective retrieval even in languages that they do not understand. At present, a large amount of the patent classification in patent offices is done manually, although due to the ever increasing numbers of applications, the use of automated systems is becoming more attractive.

A patent classification scheme must be able to classify the whole body of technological knowledge. As the technological knowledge of the world is continuously developing and expanding, a classification scheme must also be able to include devices or ideas that did not exist when it was created. Furthermore, patent classification schemes are periodically revised to take the changes to technological knowledge into account. The changes can involve both creating new categories as well as fusing categories that have received little use.

At present, a number of patent classification schemes are in use. The *International Patent Classification*[30] (IPC) is a hierarchical patent classification system maintained by the WIPO. This scheme will be described in more detail, as it is widely used and also forms the basis of

[30] http://www.wipo.int/classifications/ipc/en/

other patent classification schemes. Other patent classification schemes will then be briefly described.

4.2.1 International Patent Classification

At the top level, the IPC has 8 *Sections* indicated by the letters A to H, shown in Table 4.2. The names of the remaining four levels of the IPC hierarchy are shown on the left of Table 4.3, along with the total number of categories at each level. An example of a path down the hierarchy, along with the symbols used to represent the path, is shown on the right of the table.

Many of the categories also have notes guiding the classification, specifying exactly what is covered and what is not covered by the category. For example, subclass G04G (electronic time-pieces) has the note: "This subclass does not cover electronic time-pieces with moving

Table 4.2. The eight IPC sections.

Section	Description
A	Human necessities
B	Performing operations; Transporting
C	Chemistry; Metallurgy
D	Textiles
E	Fixed constructions
F	Mechanical engineering; Lighting, Heating, Weapons, Blasting
G	Physics
H	Electricity

Table 4.3. The five levels in the IPC hierarchy, with the total number of categories in each level in the second column. On the right of the double line is a specific example of the codes and corresponding titles for a single path down the hierarchy.

Level	No. of categories	Example symbol	Example title
Section	8	G	Physics
Class	129	G04	Horology
Subclass	631	G04D	Apparatus or tools specifically designed for making or maintaining clocks or watches
Main group	7392	G04D 3/00	Watchmakers' or watch-repairers' machines or tools for working materials
Sub-group	62493	G04D 3/04	Devices for placing bearing jewels, bearing sleeves, or the like in position

parts, which are covered by subclass G04C." Categories also have references to other categories that are related, for example, subclass G04D (shown in Table 4.3) has the references: "machine tools in general B23, B24; hand tools in general B25."

The IPC also provides categories to take into account the possibility of emerging technology not covered by any of the existing categories. For example, G04D 99/00 is: "Subject matter not provided for in other groups of this subclass." This provides a challenge for automated classification algorithms, as these categories are defined not by positive examples of their members, but by documents that do not fit into other categories.

The IPC is periodically revised. The first edition of the IPC came into force in 1968. Until 2005, it was revised every five years, but the revisions are now taking place more often.

4.2.2 Other Patent Classification Schemes

The *European Classification system* (ECLA) is built on top of the IPC (it is identical down to the main group level) and maintained by the EPO. The ECLA contains around twice as many categories as the IPC (around 140,000 at main group and sub-group level), and hence has narrower categories allowing more fine-grained classification. The *file index* (FI) used by the JPO is also an extension of the IPC, having a total of 170,000 categories [168]. The *F-terms* classification, also used by the JPO, is a completely independent classification system and is used in addition to the IPC and FI classifications. It allows documents to be classified from a number of technical viewpoints by assigning *term codes* to them [168]. A document to be classified is first assigned a *theme*, of which there are over 2,500, each of which is mapped to a set of FI codes. Each theme has a defined collection of viewpoints for specifying possible aspects of the inventions under the theme (e.g., material, operation, product, purpose, etc.), and each viewpoint has a list of possible elements that can be assigned to it. An F-term consists of a pair of a viewpoint and its element [91].

The *United States Patent Classification* (USPC) system maintained by the USPTO is an exception in that it is not based on the IPC.

Adams [6] has pointed out that even though both EPO and USPTO patents are also classified into IPC categories, there can be significant differences in the IPC categories assigned to an EPO patent and its USPTO equivalent. This is because USPTO patents are first classified by the USPC system, and then automatically assigned to IPC categories by a mapping between the systems, which most likely has shortcomings. However, a recent initiative of the EPO and USPTO to create a common classification system for technical documents resulted in the *Cooperative Patent Classification*[31] (CPC). The CPC is based on the ECLA, containing more than 200,000 categories. It has been under development since October 2010, and was officially launched on January 2, 2013. From January 2015, it is planned to be the only classification scheme used by both patent offices.

4.2.3 Uses for Automated Patent Classification

Automated patent classification in patent offices may reduce the workload of human classifiers. Due to the nature of the patent field (e.g., extensive use of neologisms, rapid technological development, complex definitions of some classes), fully automating the classification is a challenge. There are however many opportunities to support the manual classification [29, 178]. For instance, *pre-classification* is the task of distributing the incoming patent applications among the various possible groups of examiners based on their contents, which could be seen as a classification at class level of the IPC. This task has a high potential for success, as the number of classes is on the order of 100, and errors in the initial routing can be caught at the more detailed classification done manually at the next step. In *interactive classification*, the system proposes classes to the examiner and then allows the examiner to refine the classification. An example of such a system is the IPCCAT[32] provided by the WIPO. A third application for patent offices is in the *re-classification* of patent documents. After the periodic re-organizations of the patent classification schemes, all patents in the database of a patent office must be re-classified using the new scheme. This task

[31] http://www.cooperativepatentclassification.org
[32] https://www3.wipo.int/ipccat/

is currently done manually, and can require a long time to complete, making it attractive for automated assistance in the classification.

Classification codes are usually applied to patent documents at two levels. Earlier, this was usually in the form of a single code for the *main classification* (the main area to which the invention belongs), and potentially some codes for the *other classifications* or *secondary classifications*, relating to other aspects expressed in the patent [14, 53]. Furthermore, earlier IPC classifications were done based mainly on the claims of the patent document [205]. With the reform of the IPC in January 2006, it is recommended to classify additionally based on inventive aspects found in the description, examples, or drawings. These features are classified with one or more codes referred to as *invention information*, superseding the single IPC code of the *main classification*. Other content of documents, which is of lesser importance but still has search value, can be classified as *additional information* [205].

A final promising application is automated classification of non-patent documents into a patent classification scheme. These documents, including scientific literature and even web pages, should also be taken into account in prior art searches. Their lack of classification into the schemes used by the patent offices makes it cumbersome for professional users to narrow down the search space. Automated classification of these documents could lead to more effective retrieval of non-patent prior art by patent examiners.

4.3 Generating Metadata

In addition to the metadata already associated with a document, there is ample space to automatically create new metadata, correct existing metadata, or to assist the user in associating metadata to a particular document. Automated patent classification is the area of metadata generation that has received the most attention, due to the importance of the applications it has in patent offices and among patent searchers, as described in the previous section. Therefore, we focus now on text-based patent classification. We will then continue in Section 4.3.2 with a discussion of information extraction technologies for populating ontologies specific to the patent domain.

4.3.1 Text-based Patent Classification

Patent document classification is challenging for the following reasons: (1) there is a large imbalance in the distribution of documents in categories, as the number of inventions varies in different parts of the taxonomy; (2) most patents are assigned to multiple categories — a multi-classification task; and (3) the codes are assigned at two levels of importance — primary categories and secondary categories.

In 1999–2000, the EPO carried out tests for pre-classification of patents. They provided data to research groups and companies for testing their classification systems. The summary report of these tests [213] does not give a comparison of the systems participating but names some of them [87, 107].

In 2002, the WIPO made publicly available a dataset for training and testing IPC classification: the WIPO-alpha dataset,[33] containing over 75,000 patent documents in English.[34] This dataset was introduced in [53], and results for single-label classification into the IPC main classification for each patent, using common approaches such as Support Vector Machine (SVM), naïve Bayes and k-Nearest Neighbour classifiers, were presented. This dataset was mostly used for the development of single-label classifiers that take the hierarchical structure of the class labels into account [186], with many papers focusing on the development of kernel methods for this task [39, 162, 169, 192, 200]. Unfortunately, the majority of the latter papers only report experimental results on the subset of the dataset classified in Section D of the IPC, making it unclear what the scalability of these methods to the full dataset is. The task of multi-label classification — assigning more than a single class to a patent document — is done using SVMs in [40], in which classification results for each IPC Section are reported individually. The only paper that considers the problem of different levels of importance of the categories in multi-label classification (the *preferential text classification* task) is [14]. It begins by proposing a new metric for evaluating preferential text classification,

[33] http://www.wipo.int/classifications/ipc/en/ITsupport/Categorization/dataset/

[34] The WIPO-de dataset, available on the same webpage, contains over 110,000 patent documents in German.

formulates baseline solutions for this task in terms of well established classification problems, and finally adapts the Generalized Preference Learning Model [13, 15] to the task. Experiments on the WIPO-alpha dataset for preferential classification at subclass level show the advantage of explicitly modeling the primary categories as being in competition with the secondary categories. A study focusing on the scalability of the SVM and Balanced Winnow classification algorithms on a corpus of 1.2 million patent applications is presented in [30].

In recent years, IR evaluation campaigns have organized patent classification evaluation tasks. Such a task was first introduced in the NTCIR-5 with the task of classifying patents by F-terms [91], with two subtasks: the first subtask was to classify the patent into a theme and the second to classify a patent for which the theme has been provided using F-terms. The F-term subtask was run again in the NTCIR-6 [90], with the same training data but different test data. The classification of a patent by F-term given a theme is more difficult than the classification by theme — the highest MAP for classification by theme in NTCIR-5 was 0.69, obtained using a k-Nearest Neighbour (kNN) classifier. The highest MAP for classification by F-terms (for exact match to the F-terms) in NTCIR-5 was 0.50 obtained by a kNN classifier, and in NTCIR-6 was 0.49 obtained by a group of naïve Bayes models.

The NTCIR-7 campaign investigated a cross-genre classification task: the classification of research publications into the IPC, based on training with patent data. This task continued in NTCIR-8 [146]. English and Japanese patent data were provided for training, and tests were done on both English and Japanese research papers. Both monolingual (same training and test language) and cross-lingual classification evaluation was done. In NTCIR-7, only classification at subgroup level was evaluated, while in NTCIR-8, classification at three levels was evaluated. The best Japanese monolingual runs in NTCIR-8 had the following MAP values: 0.80 at subclass level, 0.64 at main group level, and 0.45 at subgroup level. All were obtained using various refinements of the kNN algorithm, including the additional use of a learning to rank algorithm.

The 2010 CLEF-IP track included a new task on patent classification [158], which was run again in 2011 [157]. In these tasks, over

1 million patents were provided as training data. The aim was to classify patent documents in the test set at IPC subclass level. The results from the 2011 task show that systems are able to assign a single subclass to a patent effectively, with the best runs, obtained using the Winnow classifier, having a precision@1 of between 0.82 and 0.85. A recall@5 of over 0.85 was also obtained, but with a correspondingly low precision@5 of around 0.55. A second classification task added in 2011 is similar to the F-term classification task in the NTCIR-5 and 6 — given the IPC subclass for a patent, return the group/subgroup classifications. Unfortunately only one group participated in this task, but as for the NTCIR task, this task proved difficult, with the best run having a precision@1 of only 0.54 (using a kNN approach).

4.3.2 Information Extraction

In addition to all the metadata already available in a patent corpus, there are large numbers of entities present in the text, entities which, if identified, assist not only the search process, but also the user in understanding the document. The process of information extraction will often involve NLP methods, some of which were presented in Section 3.2.3, but is not limited to them.

The domain of patents, as mentioned in the Introduction, covers many technological domains, each of which has, presumably, its own complex ontological system. A general approach can be taken, using gazetteers, rule-based systems or machine learning, to extract a common denominator consisting of measures and units, places and names of people, publication venues or organizations [12]. Such information can then be integrated into a complex retrieval system to allow retrieval on both the full text and the semantic data [46].

A deeper focus on specific domains is likely to be more useful, particularly since companies, as well as individual users of a patent search system, are also specialized on one or a small number of technology areas.

For most domains there are only few publications targeting the patents for information extraction. There is more work in scientific publications, which provide useful methods to create semantic data from

full-text. In some domains there is a lot more explicit semantics than in others. For instance, in the biomedical and healthcare domain, there are a number of works using the data available from PubMed Central or Medline [44, 143]. The lessons learned from these studies, such as linguistic patterns predominantly used in these domains, can be applied, in particular contexts, to the patent domain [108]. We will discuss more about biochemical information extraction methods in Section 5.4 in the context of non-textual information.

All the information extracted from the text, in combination with the existing metadata, can be put together to create a patent domain-specific ontology, such as the one attempted in the PatExpert project.[35] The PatExpert ontology represents both the explicit metadata already presented in the previous sections, as well as any other data that could be extracted directly from the text, or assigned to specific entities via a classifier, or, more generally, a machine learning method [65, 66]. Taduri et al. build on this work, but opt for a simpler ontology, which can be used to map several sources of domain knowledge [179, 180, 181]. However, the adoption in practice of these methods and tools is still unclear.

4.4 The Use of Metadata

Whether manually or automatically created (or a combination thereof), the metadata can be used inside the retrieval process in addition to being available to the user for navigation. Here, we review such uses, as described in the literature.

4.4.1 The Use of Citations

The most successful use of metadata to date is that of the citation lists in order to learn patterns of relevance. This was used for the first time by Fujii in the context of the NTCIR-6 [55]. Fujii used the citation network as a voting mechanism similar to PageRank and showed moderate, but significant improvements in comparison with text-only based retrieval. Lopez and Romary in the 2009 CLEF-IP campaign [121] also

[35] http://www.patexpert.org/

used the citations and achieved a vastly superior effectiveness score compared with all other methods. The proposed system, PATATRAS, is however much more complex than just using the citations and yet again it is difficult to distinguish where exactly the significant performance boost comes from. Given that other components had been used before or concurrently in other systems, it is reasonable to assume that the use of citations had a positive effect in this case as well. As a consequence of these results, the use of citations has been adopted by other groups as well, with similarly positive results [68, 72].

The method was further enhanced in the following year [120] with, among others, a machine learning method to expand the list of citations explicitly mentioned as metadata, with citations occurring in the text of the patent application itself [119].

The use of citations in the retrieval process may be disputed, and it has been in the various fora where these methods were presented, because the evaluation of the system also uses citations as relevance judgments. Of course, these relevance judgments were not part of the training set for any of the systems, and in our view the use of citations for system training or result re-ranking is valid scientific practice as long as care is taken not to mix the training set with the evaluation set inadvertently, via long citation chains.

4.4.2 The Use of Classifications

Another component that was a part of the PATATRAS system [121], as well as of many others, was the use of classification codes. The principle is very simple and goes back again to the multitude of sub-domains within the patent domain: it is perhaps not useful to search for prior art to an aquatic toy (US Patent 3113517) among the chemical patents. Then again, this intuition is not always true. Perhaps the aquatic toy just mentioned was floating because of a particular material it was made of and which might therefore have prior art in chemistry.

The use of IPC codes can be particularly useful in the case of a multilingual collection, since the codes themselves are language independent [50]. A similar observation was made by several other participants in the CLEF-IP campaign [61, 69, 83, 98, 199]. In general,

the participants use the IPC annotations, as they are the only set that is common across different patenting authorities, but Harris et al. investigated the different utility of the IPC, USPC, and ECLA systems [79, 80, 81]. The observations made indicate that the IPC is in fact less useful for retrieval than the other classifications. This is, in hindsight, expected, since the IPC is designed for uniformity across different patenting authorities and practice, while the national classifications are designed for retrieval.

4.5 Visualizations

Visualization of complex data is a general problem, not particularly specific to the patent domain. That being said, the patent domain has often used it, in combination with text mining, particularly in the field of business analytics [38, 42, 175, 203]. Such visualization tools can be divided into *patent graphs* that use structured data, such as the metadata described earlier, and *patent maps* that use unstructured text data [190].

Some patent graph examples are shown in Figures 4.1 to 4.3, kindly provided by M-CAM Inc. Figure 4.1 shows a visual tracking of the

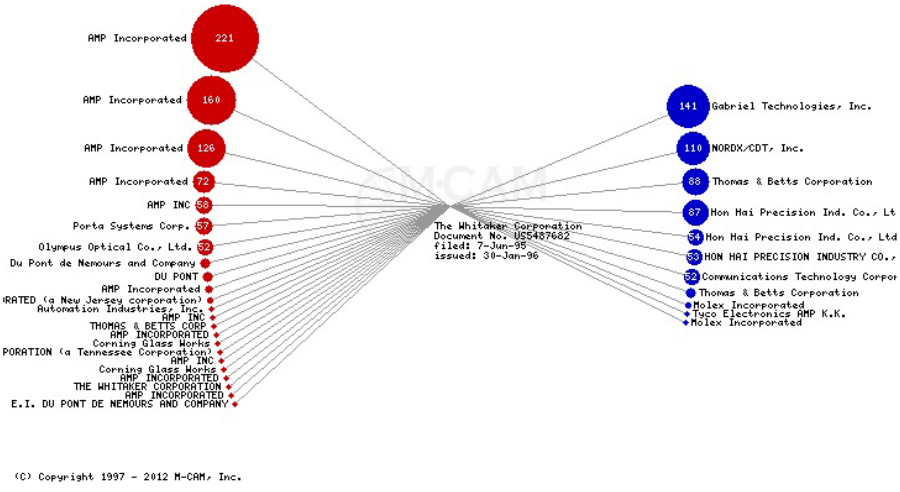


Fig. 4.1 Cited and citing patents, grouped by applicant.

citation relationships between cited prior art, the subject patent and subsequent cited art. Figure 4.2 depicts the prosecution history of the issued patents examined in the context of the subject patent. The bars at the bottom of the graphic use unstructured data to display the innovation space around the subject patent, which was not cited by the applicant or examiner. Finally, Figure 4.3 provides a geographic visualization of a patent's family data allowing comparison across country information.

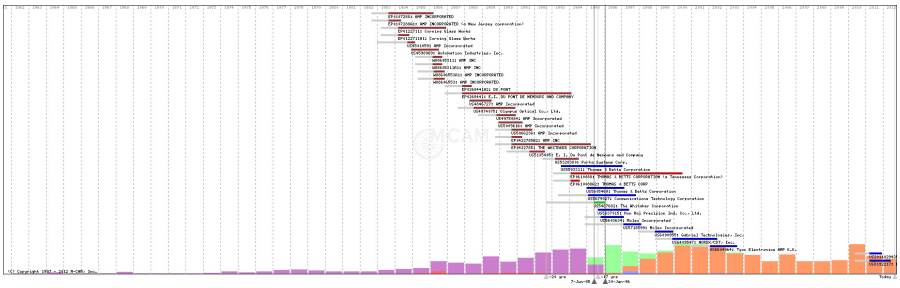


Fig. 4.2 The prior and current art of a patent, including cited documents.

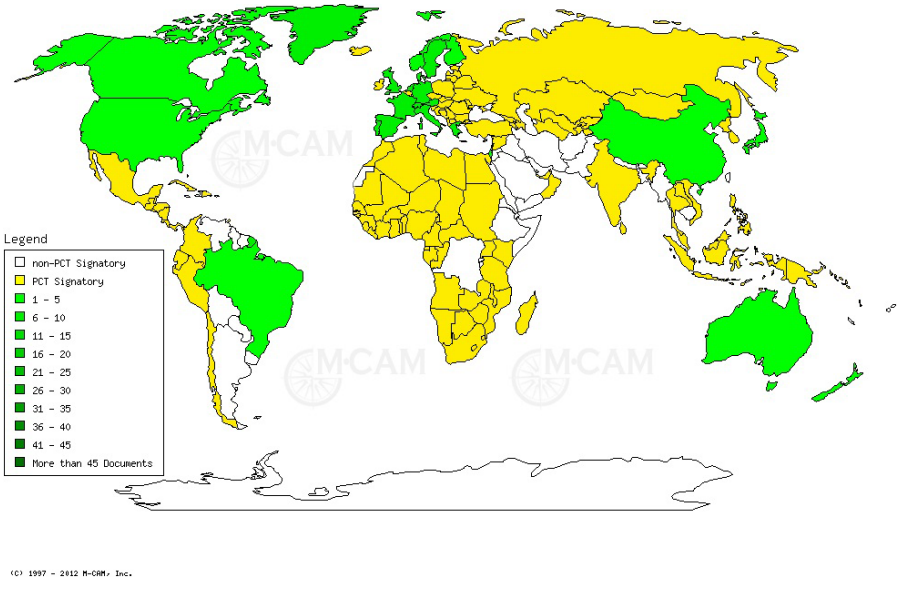


Fig. 4.3 A global view of the patent family.

An example of patent maps are the self-organizing maps used in [109] to assist the user in the analysis of a patent collection. In general however, a professional user is likely to take advantage of both graphs and maps [67, 105].

There are however many more tools for visualization in the commercial space, to which we have little access, and even fewer possibilities to evaluate. A sample of these tools is described and evaluated by professional users in [212]. Their conclusions are sometimes surprising, as they are extremely practical and have little to say about the core effectiveness of any method. The perceived strengths are phrased in a fairly general way (e.g., *"Sophisticated semantic analysis"* or *"Patent mapping, clustering and citation analysis"*) while the potential limitations obviously come from practical needs and constraints (e.g., *"Requires in-house training; high cost"* or *"Difficult to understand Themescape labels"*). This again illustrates the gap between the understanding of evaluation in the IR and IP communities, as discussed in the introduction to Section 2.

4.6 Summary

The patent domain compensates its use of complex texts with the presence of extensive metadata. Searches in patent collections were done well before modern Information Retrieval established itself, and a consequence of this is the existence of detailed information about who invented, wrote or owns the intellectual property described, when protection was sought, as well as what other similar disclosures exist (be they patent or non-patent literature) and the kind of technology it describes. All this information has been shown to help increase the effectiveness of the search methods, and in some cases it is vital.

In addition to content specific knowledge representations, recent works attempt to model the patent domain itself (the documents, the authors and other parties involved, the different dates and priorities, etc.) as an ontology on which to build data exploration tools. Among these tools, visualization methods assist the user in tasks where the search target is not necessarily very clearly defined (e.g., patent landscape search).

Automated classification of patents has a number of important practical applications, and is of immense interest to patent offices as a means of increasing efficiency in the face of a mounting number of patent applications. Automated classification at levels of the classification hierarchies with fewer classes (e.g., IPC subclass) has acceptable performance, but automated classification at the levels with more numerous classes (e.g., IPC group/subgroup) is still a challenge. The adoption by the EPO and USPTO of the CPC, with over 200,000 categories, makes the classification task even more challenging.

Results from the NTCIR and CLEF evaluation campaigns show that simpler classification algorithms (kNN and Winnow) produce the best results. These algorithms are also suitable as they can be easily applied to large amounts of data. The kNN approach can be implemented by indexing and retrieving similar patents from the training data, while the Winnow approach has a very simple model. The main disadvantage of a more complex model such as an SVM is that it requires more time for training [30, 198] and hence does not scale well with the size of the data. None of the evaluation campaigns have yet evaluated the task of classifying a patent into primary and secondary classes (preferential classification) — it would be interesting to test the scalability of preferential classification algorithms to large amounts of data and with many classes.

5

Beyond Text

The majority of the work on patent retrieval has focused on the text and metadata of the patent. However, patents also contain non-textual information in the form of images, or, as they are referred to in the patent domain, *drawings*. The USPTO "Nonprovisional (Utility) Patent Application Filing Guide"[36] makes the following statement, strongly encouraging applicants to include drawings:

> A patent application is required to contain drawings, if drawings are necessary to understand the subject matter to be patented. Most patent applications contain drawings. The drawings must show every feature of the invention as specified in the claims. Omission of drawings may cause an application to be considered incomplete and no application filing date will be granted by the USPTO.

In some cases, key information may be only in the drawings and not in the text. For example, in mechanical gearing and power

[36] http://www.uspto.gov/patents/resources/types/utility.jsp

transmissions, inventions often involve the relative orientation of various known parts, and the information is typically conveyed by reference to drawings [178]. This underlines the importance of the drawing content in the information contained in the patent, although this content has largely been ignored in patent IR systems.

Chemical structures can be represented in multiple ways in a patent, including a number of methods of encoding them in text as well as in drawings. To assist in searching, it is essential that these various representations are converted to comparable "chemical entities."

After an introduction to the characteristics of drawings in patents, we review the work that has been done in patent drawing retrieval, patent drawing classification and the retrieval of chemical structures.

5.1 Characteristics of Drawings in Patents

Drawings in patents are usually in black and white, as patent offices often make the submission of color drawings a more complex process. The USPTO, for example, requires an explanation of why the color drawings are necessary and the payment of an additional fee.

Patents contain different types of figures. The patent drawing ontology created by Vrochidis et al. [201], shown in Figure 5.1, specifies the five classes of figure that occur in patents: photo, diagram (including block, state, and circuit diagrams), flowchart, technical drawing, and graph. For the CLEF-IP 2011 patent image classification task, nine classes of patent image were identified by analyzing the MAREC

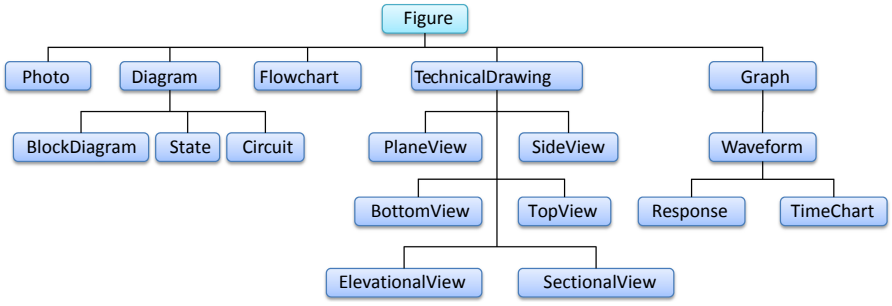


Fig. 5.1 Patent Drawing Ontology (from [201]).

dataset: abstract drawing, graph, flowchart, gene sequence, program listing, symbol, chemical structure, table, and mathematics. Examples of these are shown in Figure 5.2. Drawings in older patents are often done by hand, although the increasing use of computerized drawing packages is leading to more homogeneous drawing styles in more recent patent applications. In electronic versions of older patents, the quality of the drawings is usually poor, due to the less advanced scanning techniques available when they were scanned. Even in more recent patents that have been submitted electronically, workflows in some patent offices lead to a loss of the vectorized images and their replacement by bitmap images.

There is a strong association between patent drawings and patent text. As in the majority of technical documents, drawings are numbered consecutively and referred to as "FIG. x" in the text (note that tables are usually also considered to be drawings). However, the figure number indication is usually part of the drawing, so the relation is not always simple to extract automatically due to the widely varying fonts used in the drawings, as demonstrated in the examples in Figure 5.3. Some patents also contain a section headed *Brief Description of the Drawing*, which contains the brief captions of the drawings, for example, "FIG. 2 is a partial view of a cutting head assembly." Further linking between drawings and text is provided by part numbers. These part numbers are referenced in the text — applicants are discouraged by the patent application guidelines from including words in the drawing (in the case of the EPO, this is to allow the straightforward reuse of drawings in translated versions of the patent). For example, for the drawing in the top row of Figure 5.2, the first part of the corresponding passage from the description is:

> As FIG. **2** shows in detail, each membrane support plate **6** comprises a frame, for example rectangular frame **14** made of profiled tubes or bars. The frame **14** is, in the case shown, mounted on a support base **15** . . .

List [118] presents the current challenges faced in searching drawings in patents from a patent search professional's points of view — at

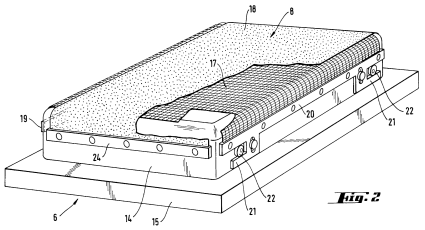
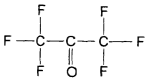
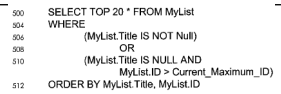
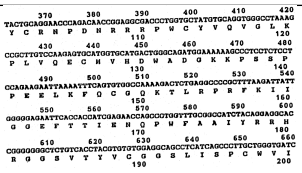
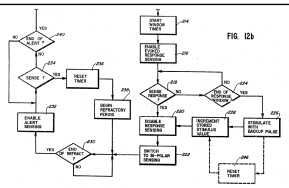
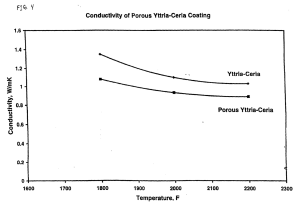
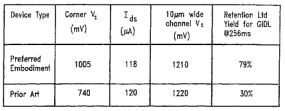
1	drawing	
2	chemical structure	
3	program listing	
4	gene sequence	
5	flowchart	
6	graph	
7	mathematics	$$\sigma(y) = \frac{1}{1 + \exp(-y)}$$
8	table	
9	character	

Fig. 5.2 The nine classes of patent drawing used in CLEF-IP 2011.

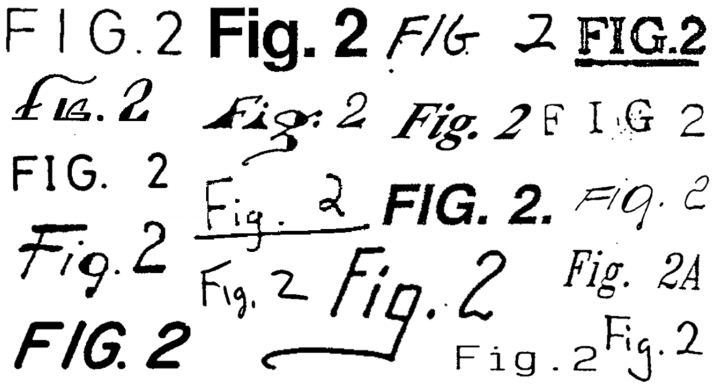


Fig. 5.3 A collage of drawing identification labels from patent drawings (from [75]).

present the majority of drawing-based search is done by manual visual comparison of the drawings. Adams [7] describes the crucial role that non-textual information plays in patents. He points out that physical scale models were required to be submitted with early US patents, and that the complexity of many inventions is poorly captured by the stack of black and white cross-section drawings on paper required today. He argues that it would make sense for patent offices to accept electronic 3D models and other electronic supplementary material to make the invention disclosures more intelligible and hence make the patent searching process more effective. Areas in which the submission of 3D structural data is useful include: mechanical devices, chemistry, and biology. Many patent offices already accept or are considering accepting 3D Computer-Aided Design (CAD) models and structural models with patent applications [194].

5.2 Computer-aided Patent Drawing Retrieval

The search for similar patent drawings is mostly based on references and the text in the captions. Some work has been done on automatically linking text in drawings (e.g., figure numbers) to paragraphs in the patent text [114, 130] to make it easier for a reader to switch between drawings and their descriptive text.

As described above, patent drawings are in general binary. The majority of work done on image similarity retrieval focuses on

photographs, and uses features such as color and texture [48, 177], which are not applicable to patent drawings in a useful way. Nevertheless, the classic Content-Based Image Retrieval (CBIR) approach is usually adopted for drawing retrieval systems for patents: a feature representation is chosen and this is applied to calculate the distance between all drawings in a patent database, irrespective of the type of drawing.

Some early work focused on developing feature representations specific to binary drawings in patents [86, 187]. This early work was however only tested on small datasets of around 200 drawings. More recent work on drawing search in patents has been done in the PatExpert[37] project [204], in which the PatMedia[38] system was developed [174, 201]. A new hierarchical feature representation for the patent drawings was developed and shown to perform well on a dataset of 2,000 patent drawings, outperforming three earlier methods for binary image retrieval [219, 137, 210]. The PatViz system for applying visual analytics to patent search makes use of the PatMedia system for image similarity calculation [105].

Beyond binary drawings, approaches have been described specifically aimed at design patents [217] and at flower patents [47]. The latter (patents covering new families or sub-families of flowers) is an area in the patent domain to which common color and texture-based image retrieval approaches are applicable due to the use of color photos of flowers.

A drawback of all published papers on patent drawing retrieval is the relative small scale of their experiments, not representative of the huge number of drawings that need to be searched through in the patent domain. In 2011, for the first time, a patent drawing retrieval task was offered in CLEF-IP [157]. A total of 23,444 patent documents (in XML format) from three IPC subclasses were provided to participants together with their 291,566 image files (i.e., an average of 12.5 images per patent). The three IPC subclasses were chosen to limit the set of drawings to a reasonable number, and were chosen based on input from

[37] http://www.patexpert.org
[38] http://mklab-services.iti.gr/patmedia/

patent searchers as to which classes often require visual comparison of drawings to locate prior art. Participants had the choice of using either visual information, textual information or combined visual and textual information in the retrieval experiments. As queries, 211 patents (along with their drawings) were provided. The common CLEF-IP approach of creating relevance judgments based on citations was followed. Unfortunately only one group, Xerox Research Centre Europe, participated [45]. As feature representation for the patent drawings, they used the Fisher vector representation [155]. Experiments were done using image features only, text features only, and a combination of text and image features. As is to be expected, the performance of the retrieval using only image features was poor, giving a best MAP of 0.035. The text-only runs significantly outperformed the image-only retrieval, producing a best MAP of 0.203. The fusion of the text and image runs with a late fusion approach did result in an improved MAP of 0.212, demonstrating the usefulness of including visual information in the ranking determination.

5.3 Image Classification

For the work on patent drawing retrieval presented above, a global image feature representation was chosen for all drawings. However, as already mentioned in Section 5.1, patents contain different classes of drawings with different characteristics. It was pointed out in [75] that a large amount of work on the classes of drawings found in patents has been done on technical documents in general. This includes work on technical drawings [188], 2D plots [123], and charts [218]. A more detailed review of the approaches applied to these drawing types is in [31]. A promising approach to improve drawing retrieval in patents is to adapt the analysis and matching to each class of patent drawing, for example compare flowcharts to each other with a similarity metric designed for flowcharts.

In order to accomplish this, it is necessary to classify each patent image by type. Vrochidis et al. [201] do such a classification based on the text in the drawing captions, but the caption does not always contain the name of the image class. In order to evaluate the classification

of patent drawings into classes based on visual information, an image classification task was organized in CLEF-IP 2011 [157]. The task was to classify images into the nine classes depicted in Figure 5.2 using only visual information. As training data, 38,087 images pre-classified into the nine classes were provided. 1,000 images were later released as test data. Two groups participated [45, 142], with the best true positive classification rate obtained being 91%. This was obtained by making use of the Fisher vector representation mentioned above followed by a linear classifier.

5.4 Chemical Structure in Patents

Chemical patents, which are of key importance to the pharmaceutical and agrochemical industries, have to contain information about the chemical molecules of interest. This information can be represented in a number of formats. Formats that can be represented by text include the molecule name, SMILES (Simplified Molecular Input Line Entry Specification), InChI (International Chemical Identifier), and a connection table. The structure can also be represented in an image, as shown in the second row of Figure 5.2 [84]. Chemical structures are made up of a limited set of symbols, reducing their variability. Therefore, while it is still challenging to interpret these structures, it is simpler than interpreting a mechanical drawing, or even a plot, in which there is much more flexibility as to the objects depicted [75].

Recently, the extraction of chemical structures from images in patents has attracted particular interest [54, 195]. To evaluate how well chemical structure information can be extracted from images, an Image-to-Structure task was organized in TREC-CHEM 2011 [126]. Overall, all participants correctly recognized over 60% of the structure images, with the best runs recognizing over 90%.

The difficulty in interpreting chemical structures lies more in mapping the discovered structure to known compounds, or to other, similar representations of the same compound. These techniques are an important component of *chemoinformatics* which is "concerned with the application of computational methods to tackle chemical problems, with particular emphasis on the manipulation of chemical structure

information" [111]. This process can be assisted by information present in the text. In fact, the chemical domain is to a large extent a privileged domain in terms of the resources that patent searchers have at their disposal. The advantage of searchers in this domain is the long history and practice of chemical retrieval. The importance of chemistry in all aspects of human life led to the appearance, over 50 years ago, of a manually created index, which is now part of the Chemical Abstracts Service (CAS).[39] Efforts on making this process automatic have also been extensive [99].

More recently, Klinger and his colleagues at Fraunhofer SCAI explored the detection of chemical formulae in text using Conditional Random Fields (CRF) [104], and then the same research group used it successfully in the TREC-CHEM campaign [71, 72, 73]. The results show that using the chemical entities in the retrieval process always improves results. They do not say, however, what was the effectiveness of the entity recognition process. We may assume that it was similar to the one on scientific articles, since that study [104] briefly mentioned a smaller test on a hand-selected collection of patents. But even Klinger and his colleagues were surprised at the apparent small loss in F_1 values, given that scientific articles and patents still present some considerable distinctions in terms of presence of chemical and biomedical entities [144].

As Klinger before, Grego et al. [70] use CRF on a patent collection consisting of European patents, with manual annotations as the gold standard. They also use Oscar3, a general purpose chemical information extraction engine developed in the Peter Murray-Rust group in Cambridge.[40] They show lower F_1 scores than Klinger, to some extent confirming our expectation that the patent corpus is more difficult. The comparison cannot be made conclusively, since they looked for partially different forms of entities (Klinger et al. looked exclusively at IUPAC names, while Grego et al. at more general chemical entities), and they have slightly different definitions of what a match is (particularly with respect to partial matches).

[39] http://www.cas.org/
[40] http://www-pmr.ch.cam.ac.uk/wiki/Oscar3

The systems presented so far do not approach the patent collection any differently from the scientific articles. Some fine-tuning of these methods may be needed for the patent domain, but the difference lies mainly after the entity recognition phase, in the search and use of the so-found entities. Chemical patent search requires the system to allow for partial structures in the queries. This is for two reasons: First, because patents, in their desire to cover as many structures as possible, use a form of regular expressions for the patent domain — Markush structures [25, 26] — to represent possibly infinite sets of chemical formulae. Second, because innovation is never totally new, especially in a field as old as chemistry, and therefore there may always be a set of sub-structures that would be easy to combine to invalidate the novelty of a new structure. An excellent overview of these issues has recently been written by Holliday and Willet [84].

5.5 Summary

The use of non-text information in patent IR has received relatively little attention. Image retrieval systems that have been developed for patents have tended to process each drawing in the same way, regardless of its content. A promising route to improving patent drawing retrieval is to make use of the work that has been done on figure analysis and retrieval for specific types of figures in technical documents. This will require that a pre-classification by type of the patent drawings be done, but good results in the CLEF-IP 2011 drawing classification task have shown that this is feasible. As this task used only visual features from the drawings, adding text from the drawing captions has the potential to improve the drawing type classification further. For drawings such as circuit diagrams, chemical structures and flowcharts, which are built on a limited vocabulary of symbols, automatically extracting the "meaning" of the drawing is feasible. However, in many cases, it is extremely difficult to automatically extract the complex semantic concepts represented from the abstract drawings available as 2-bit depth bitmaps in patents.

In the chemical domain, advantage can be taken of the large amount of work done in the area of cheminformatics. However, beyond the

general problems of chemical information processing and retrieval, the patent domain introduces here a form of regular expressions — Markush structures — designed to cover many (potentially infinite) compounds with similar functionality. Calculating similarity based on these structures can be a computationally expensive task.

Until now, the separate modalities in the patents have generally been retrieved separately. For example, individual images are retrieved from multiple patents based only on image similarity. Taking both information about text and images into account by modality fusion in the retrieval has produced promising results in the CLEF-IP 2011 evaluation. There remains much research to be done on multimodal retrieval of patents. In particular, a representation for all of the modalities of a patent that would allow similarity to be calculated at patent level should be developed. The importance of the various modalities in calculating patent similarity and how this changes with respect to query and search type should also be investigated.

The potentially increasing use of digital 3D models in patents will require the use of methods for retrieving such models based on their content. A recent survey of these methods is available in [184].

6

Conclusions

This survey serves two purposes. First, it reviews the work already done in the field of patent IR. Second, it introduces the specificities and peculiarities of this domain to IR researchers looking for a new challenge for their methods and algorithms.

We have covered as much as possible of the research focused on patent data, and provided the reader with a reference guide for the different aspects of this domain. We have naturally focused on Information Retrieval issues, but touched on other aspects related to information management, such as classification, information extraction, and visualization. We did not cover business aspects of patent research, such as those aiming to evaluate patents and patent portfolios.

In the end, we observe, without surprise, that systems that perform well on general IR test data also perform well on patent data, in relative terms. In absolute values, they all perform apparently less well on the patent data. Such values cannot be directly compared, but practice and the experience of professional searchers corroborate such a tentative conclusion. The lower effectiveness figures are, on the one hand, the result of the genre and domain of patent documents. On the other hand, they are also a result of the very specific use cases on which they are

evaluated. In practice, several other factors than topical relevance are considered, particularly in such specific searches as validity or patent landscape search.

For such use cases, a combination of metadata analysis, information extraction and ontologies (including classifications) are used to boost effectiveness scores. Such explicit semantic methods are particularly interesting because they have the potential to assist the user in evaluating the risk present in each search, in a similar way to what Boolean search is now doing (i.e., providing visible proof of why a document has been retrieved). While evaluation campaigns are excellent at predicting average performance of many queries, the patent professional needs guarantees on each individual search result, and often the best guarantee is an understanding of the limits of the search system.

The case of image search and processing for the patent domain is particularly difficult. While for text there are a number of options that have been explored and which show promising results, the semantic gap between the 2-bit color depth abstract drawings and the complex concepts they depict is still extremely large. Only in certain specific cases, like chemical structure images, has some progress been made.

6.1 Adoption of IR Technologies

Based on the current survey, we can see a large increase in the production of research results using patent data, and in addressing patent search issues. To what extent these results are adopted by the industry and incorporated in end-products is unclear. Certainly, there are economic and business reasons that lead to the adoption by a certain provider of a particular technology. There is however a more significant issue which surfaced repeatedly in our discussions with patent professionals: the searcher's professional success relies on the search system. With existing systems, the searcher is, through practice and experience, able to understand their weak points and compensate. This is for instance the attractiveness of Boolean search results: the reasons for the presence or absence of documents in the search results are easy to understand, and the results can therefore be calibrated through professional experience. The IR engines present in academia

have demonstrated their performance improvements over the long term (through averages over many queries), but are difficult to debug for any individual query.

Another aspect which is not necessarily viewed with great enthusiasm by the professional searcher is the lack of repeatability of results. A Boolean system will return the same set of documents when the collection changes (plus or minus the documents changed). A modern IR system will return mostly the same documents, but its capacity to adapt the term weights and the similarity function makes it such that the result is not guaranteed to be the same. Professional searchers are, in some cases, liable for the quality of their search and they might need to show that a previous search did not produce a particular document. Modern IR systems (as opposed to IR engines) need to cater to this need as well.

6.2 Trends in Patent IR

We have hinted already at interesting issues to focus on for future research and development. Let us now put them in a list. They are the result of going through the many papers covered in this review, but also of the many interactions we have had with patent searchers from different companies or patent offices, and of following discussions in patent search fora. We have tried to order the issues directly proportional to our perception of their importance for the domain, and inversely proportional to the amount of research already carried out, but this is still our subjective ranking. This should not be taken to imply that further study in all of the issues mentioned in this survey is not important.

Information Fusion. The information available in one patent document is multilingual and multimodal. The professional searcher adapts to this situation by using several different tools (or simply going manually through sets of patents to identify images of interest). Our initial attempt to encourage the combination of these two domains by proposing a task at CLEF-IP 2011 [157] showed that there are extremely few research groups able to tackle this challenging task. More consideration of non-textual information combined with textual information is

therefore needed. The general research question in this case would be: Is it possible and useful to treat all of the components (text, image, metadata, etc.) of a patent document as a single entity in the similarity calculation, instead of, for instance, calculating similarity of single images?

Two refinements of this information fusion issue deserve special mention. First, we remember from Section 5 that not all images are equal in this domain, and that similarity must take into account the specificities of these different types if it is to be of significant use. Second, while the use of citations and other metadata is mentioned in several works and discussed here, it is unclear from these initial studies whether their use has already been optimized, or if there is still space to optimize further.

Federated Search. As opposed to the previous issue, where the research problem assumes availability of all data in one site, the practice of patent search often requires searching in different databases. In most cases, such databases cannot be merged (perhaps less for technical than for policy or commercial reasons). There is a significant amount of work in distributed IR, but its application to the patent domain is not very visible. Some issues which may require special attention when adapting to this domain are: the different genres of the patent documents versus the non-patent literature and the links between these databases (i.e., citations from one collection to another).

User Modeling. Precisely understanding the needs of the user and the processes that he/she goes through to satisfy those needs is generally important for IR. An effort to model the patent searcher is already undergoing in the context of the PROMISE project mentioned in previous sections. Based on this experience, we can say that there is still a substantial amount of work to be done in this sense.

User Interfaces and Collaborative Search. Based on the current understanding of the search processes, and that which will still be developed in the future, better interfaces should be designed. The amount of work available in this sense is particularly limited, because the vast majority of systems used by professional searchers are commercial systems, and, as discussed, their evaluation is generally done via very

narrow examples. An exhaustive study in this sense would be highly desirable, to complement effectiveness evaluation campaigns such as CLEF, TREC, or NTCIR.

In some cases, the interfaces must allow for several users to collaborate on a search. Particularly in cases of very high importance to a company, it is not uncommon for an information specialist to collaborate with a domain specialist (a chemist, for instance) and a lawyer specialised in a particular jurisdiction. How exactly to optimize this process is largely unclear.

Unit of retrival. Finally, we should not forget that patent search is also about assisting the searcher in understanding a domain or a document. In addition to document retrieval, identifying important paragraphs within each document, as well as important concepts and their links is still an issue in patent search, again due to the complexity of the documents to be processed.

Acknowledgments

We thank the Information Retrieval Facility (IRF), in particular John Tait, for initiating and guiding the research on patent retrieval that has led to a significant amount of progress in the field. We are also grateful to our colleagues at the IRF and subsequently at the Vienna University of Technology for inspiring discussions and collaboration on the subject of patent retrieval: Linda Andersson, Florina Piroi, Giovanna Roda, and Mike Salampasis.

We are particularly grateful for extensive information on the IP domain and professional patent searching from: Stephen Adams, Rosa Alentorn, Pierre Buffet, Barrou Diallo, Carlos Faerman, Monika Hanelt, Maike Houtrouw, Frasier Kennedy, Jane List, Teresa Loughbrough, Madeleine Marley, Susan McKee, Henk Thomas, Tony Trippe, and Gerard Ypma.

The work on this review has been partially funded by the following projects: PROMISE Network of Excellence (FP7-258191), funded by the European Commission; IMPEx (825846), funded by the Austrian Research Promotion Agency (FFG); and PLuTO (ICT-PSP-250416), funded by the European Commission.

Notations and Acronyms

ANC	American National Corpus
BNC	British National Corpus
CLEF	Cross-Language Evaluation Forum
CLEF-IP	CLEF Intellectual Property Track
CLIR	Cross-Language Information Retrieval
CPC	Cooperative Patent Classification
CRF	Conditional Random Field
DF	Document Frequency
DLSI	Differential Latent Semantic Indexing
ECLA	European Classification System
EPO	European Patent Office
FI	File Index (JPO classification scheme)
IDF	Inverse Document Frequency
IPC	International Patent Classification
IP	Intellectual Property
IP5	The Five IP Offices
IR	Information Retrieval
ISR	International Search Report
JPO	Japan Patent Office

KCCA	Kernel Canonical Correlation Analysis
KIPO	Korean Intellectual Property Office
kNN	k-Nearest Neighbor
LM	Language Model
LSI	Latent Semantic Indexing
MAP	Mean Average Precision
MAREC	Matrixware Research Collection
MRR	Mean Reciprocal Rank
nDCG	normalized Discounted Cumulative Gain
NII	National Institute of Informatics (Japan)
NLP	Natural Language Processing
NP	Noun Phrase
NTCIR	NII Test Collection for IR Systems
OCR	Optical Character Recognition
PM	Probabilistic Model
PRF	Pseudo-Relevance Feedback
PROMISE	Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation
SIPO	State Intellectual Property Office of the People's Republic of China
SMT	Statistical Machine Translation
SVM	Support Vector Machine
TF	Term Frequency
TREC	Text Retrieval Conference
TREC-CHEM	TREC Chemical IR Track
TS	Technology Survey
USPC	United States Patent Classification
USPTO	United States Patent and Trademark Office
VSM	Vector Space Model
WIPO	World Intellectual Property Organization

References

[1] "The clueweb dataset," http://lemurproject.org/clueweb09.php/index.php.

[2] "Manual of patent examination procedure, section 608.01(m)," Revision July 2010, http://www.uspto.gov/web/offices/pac/mpep/index.htm.

[3] "Understanding intellectual property," http://www.wipo.int/about-ip/en/.

[4] *PATENT '03: Proceedings of the ACL-2003 Workshop on Patent Corpus Processing*. Association for Computational Linguistics, 2003.

[5] *PaIR '11: Proceedings of the Workshop on Patent Information Retrieval*. ACM, 2011.

[6] S. Adams, "Comparing the IPC and the US classification systems for the patent searcher," *World Patent Information*, vol. 23, pp. 15–23, 2001.

[7] S. Adams, "Electronic non-text material in patent applications-some questions for patent offices, applicants and searchers," *World Patent Information*, vol. 27, pp. 99–103, 2005.

[8] S. Adams, "New methodologies for patent searching; what do we need?," in *Proceedings of the Global Symposium of Intellectual Property Authorities*, 2009.

[9] S. Adams, "The text, the full text and nothing but the text: Part 1 — standards for creating textual information in patent documents and general search implications," *World Patent Information*, vol. 32, pp. 22–29, 2010.

[10] S. Adams, "The text, the full text and nothing but the text: Part 2 — standards for creating textual information in patent documents and general search implications," *World Patent Information*, vol. 32, pp. 120–128, 2010.

[11] S. Adams, *Information Sources in Patents*. G. Saur, 3rd ed., 2011.

[12] M. Agatonovic, N. Aswani, K. Bontcheva, H. Cunningham, T. Heitz, Y. Li, I. Roberts, and V. Tablan, "Large-scale, parallel automatic patent annotation," in *Proceedings of Workshop on Patent Information Retrieval*, pp. 1–8, 2008.

[13] F. Aiolli, "A preference model for structured supervised learning tasks," in *Proceedings of IEEE International Conference on Data Mining*, pp. 557–560, 2005.

[14] F. Aiolli, R. Cardin, F. Sebastiani, and A. Sperduti, "Preferential text classification: Learning algorithms and evaluation measures," *Information Retrieval*, vol. 12, pp. 559–580, 2009.

[15] F. Aiollo and A. Sperduti, "Learning preferences for multiclass problems," in *Advances in Neural Information Processing Systems*, pp. 17–24, 2005.

[16] D. Alberts, C. B. Yang, D. Fobare-DePonio, K. Koubek, S. Robins, M. Rodgers, E. Simmons, and D. DeMarco, "Introduction to patent searching — practical experience and requirements for searching the patent space," in *Current Challenges in Patent Information Retrieval*, (M. Lupu, K. Mayer, J. Tait, and A. J. Trippe, eds.), Springer, pp. 3–43, 2011.

[17] L. Andersson, "A vector space analysis of Swedish patent claims with different linguistic indices," in *Proceedings of Workshop on Patent Information Retrieval*, ACM, pp. 47–56, 2010.

[18] M. Annies, "Full-text prior art and chemical structure searching in e-journals and on the internet — a patent information professional's perspective," *World Patent Information*, vol. 31, pp. 278–284, 2009.

[19] K. H. Atkinson, "Towards a more rational patent search paradigm," in *Proceedings of Workshop on Patent Information Retrieval*, 2008.

[20] L. Azzopardi, W. Vanderbauwhede, and H. Joho, "Search system requirements of patent analysts," in *Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 775–776, 2010.

[21] L. Azzopardi and V. Vinay, "Accessibility in information retrieval," *Advances in Information Retrieval*, pp. 482–489, 2008.

[22] R. H. Baayen, R. Piepenbrock, and L. Gulikers, "The CELEX Lexical database (release 2)," 1995.

[23] R. Bache and L. Azzopardi, "Improving access to large patent corpora," in *Transactions on Large-Scale Data- and Knowledge-Centered Systems II*, Springer, pp. 103–121, 2010.

[24] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Addison Wesley, 2nd ed., 2010.

[25] J. M. Barnard and G. M. Downs, "Use of Markush structure techniques to avoid enumeration in diversity analysis of large combinatorial libraries," http://www.daylight.com/meetings/mug97/Barnard/970227JB.html, last checked, March 2012.

[26] J. M. Barnard and P. M. Wright, "Towards in-house searching of Markush structures from patents," *World Patent Information*, vol. 31, pp. 97–103, 2009.

[27] S. Bashir and A. Rauber, "On the relationship between query characteristics and IR functions retrieval bias," *Journal of the American Society for Information Science and Technology*, vol. 62, pp. 1515–1532, 2011.

[28] D. Becks, T. Mandl, and C. Womser-Hacker, "Phrases or terms? the impact of different query types," in *Cross-Language Evaluation Forum (Notebook Papers/Labs/Workshop)*, 2010.

[29] K. Benzineb and J. Guyot, "Automated patent classification," in *Current Challenges in Patent Information Retrieval*, (M. Lupu, K. Mayer, J. Tait, and A. J. Trippe, eds.), Springer, pp. 239–261, 2011.

[30] K. Beuls, B. Pflugfelder, and A. Hanbury, "Comparative analysis of balanced winnow and svm in large scale patent categorization," in *Proceedings of Dutch-Belgian Information Retrieval Workshop (DIR)*, pp. 8–15, 2010.

[31] N. Bhatti and A. Hanbury, "Image search in patents: A review," *International Journal on Document Analysis and Recognition (IJDAR)*, 2012. doi:10.1007/s10032-012-0197-5.

[32] M. Blackman, R. Honeywood, and K. Milne, "Searching organic chemical structures: A comparison of online access to chemical abstracts (CAS) and the corresponding United Kingdom Patent Office search system (c2c)," *World Patent Information*, vol. 8, pp. 20–28, 1986.

[33] A. Blanchard, "Understanding and customizing stopword lists for enhanced patent mapping," *World Patent Information*, vol. 29, pp. 308–316, 2007.

[34] G. H. Blosser, N. Arshadi, and S. Agrawal, "A critical assessment of the USPTO policies toward small entity patent applications," *Technology and Information*, vol. 13, 2011.

[35] D. Bonino, A. Ciaramella, and F. Corno, "Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics," *World Patent Information*, vol. 32, pp. 30–38, 2010.

[36] N. Bouayad-Agha, G. Casamayor, G. Ferraro, S. Mille, V. Vidal, and L. Wanner, "Improving the comprehension of legal documentation: The case of patent claims," in *Proceedings of International Conference on Artificial Intelligence and Law*, pp. 78–87, 2009.

[37] A. Browne, A. McCray, and S. Srinivasan, "The specialist lexicon," Technical Report, National Library of Medicine, 2000.

[38] A. Buchanan, N. H. Packard, and M. A. Bedau, "Measuring the evolution of the drivers of technological innovation in the patent record," *Artificial Life*, vol. 17, pp. 109–122, 2011.

[39] L. Cai and T. Hofmann, "Hierarchical document categorization with support vector machines," in *Proceedings of International Conference on Information and Knowledge Management*, pp. 78–87, 2004.

[40] L. Cai and T. Hofmann, "Exploiting known taxonomies in learning overlapping concepts," in *Proceedings of International Joint Conference on Artifical intelligence*, pp. 714–719, 2007.

[41] A. Ceausu, J. Tinsley, A. Way, J. Zhang, , and P. Sheridan, "Experiments on domain adaptation for patent machine translation in the pluto project," in *Proceedings of Conference of the European Association for Machine Translation*, 2011.

[42] A. Chakrabarti, I. Dror, and N. Eakabuse, "Interorganizational transfer of knowledge: An analysis of patent citations of a defense firm," *IEEE Transactions on Engineering Management*, vol. 40, pp. 91–94, 1993.

[43] L. Chen, N. Tokuda, and H. Adachi, "A patent document retrieval system addressing both semantic and syntactic properties," in *Proceedings ACL Workshop on Patent Corpus Processing*, 2003.

[44] A. Chu, S. Sakurai, and A. F. Cardenas, "Automatic detection of treatment relationships for patent retrieval," in *Proceedings of Workshop on Patent Information Retrieval*, ACM, pp. 9–14, 2008.

[45] G. Csurka, J.-M. Renders, and G. Jacquet, "XRCE's participation at patent image classification and image-based patent retrieval tasks of the CLEF-IP 2011," in *Cross-Language Evaluation Forum (Notebook papers/ Labs/Workshop)*, 2011.

[46] H. Cunningham, V. Tablan, I. Roberts, M. Greenwood, and N. Aswani, "Information extraction and semantic annotation for multi-paradigm information management," in *Current Challenges in Patent Information Retrieval*, (M. Lupu, K. Mayer, J. Tait, and A. J. Trippe, eds.), Springer, pp. 307–327, 2011.

[47] M. Das, R. Manmatha, and E. M. Riseman, "Indexing flower patent images using domain knowledge," *IEEE Intelligent Systems*, vol. 14, pp. 24–33, 1999.

[48] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40, pp. 5:1–5:60, 2008.

[49] M.-C. de Marneffe, B. MacCartney, and C. D. Manning, "Generating typed dependency parses from phrase structure parses," in *Proceedings of International Conference on Language Resources and Evaluation (LREC)*, 2006.

[50] E. D'hondt and S. Verbene, "CLEF-IP 2010: Prior art retrieval using the different sections in patent documents," in *Cross-Language Evaluation Forum (Notebook Papers/LABs/Workshops)*, 2010.

[51] E. D'hondt, S. Verberne, W. Alink, and R. Cornacchia, "Combining document representations for prior-art retrieval," in *Cross-Language Evaluation Forum (Notebook Papers/Labs/Workshop)*, 2011.

[52] C. Emmerich, "Comparing first level patent data with value-added patent information: A case study in the pharmaceutical field," *World Patent Information*, vol. 31, pp. 117–122, 2009.

[53] C. J. Fall, A. Törcsvári, K. Benzineb, and G. Karetka, "Automated categorization in the international patent classification," *SIGIR Forum*, vol. 37, pp. 10–25, 2003.

[54] I. V. Filippov and M. C. Nicklaus, "Optical structure recognition software to recover chemical information: Osra, an open source solution," *Journal of Chemical Information and Modeling*, vol. 49, pp. 740–743, 2009.

[55] A. Fujii, "Enhancing patent retrieval by citation analysis," in *Proceedings International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 793–794, 2007.

[56] A. Fujii and T. Ishikawa, "Patent retrieval experiments at ULIS," in *Proceedings of NII Test Collection for IR Systems-3*, 2002.

[57] A. Fujii, M. Iwayama, and N. Kando, "Overview of patent retrieval task at NTCIR-4," in *Proceedings of NII Test Collection for IR Systems-4*, 2004.

[58] A. Fujii, M. Iwayama, and N. Kando, "Test collections for patent-to-patent retrieval and patent map generation in NTCIR-4 workshop," in *Proceedings of International Conference on Language Resources and Evaluation (LREC)*, 2004.

[59] A. Fujii, M. Iwayama, and N. Kando, "Introduction to the special issue on patent processing," *Information Processing and Management*, vol. 43, pp. 1149–1153, 2007.

[60] A. Fujii, M. Iwayama, and N. Kando, "Overview of the patent retrieval task at the NTCIR-6 workshop," in *Proceedings of NII Test Collection for IR Systems-6*, 2007.

[61] S. Fujita, "Revisiting document length hypotheses: A comparative study of Japanese newspaper and patent retrieval," *ACM Transactions on Asian Language Information Processing*, vol. 4, pp. 207–235, June 2005.

[62] G. W. Furnas, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum, "Information retrieval using singular value decomposition model of latent semantic structure," in *Proceedings of SIGIR*, 1988.

[63] D. Ganguly, J. Leveling, and G. Jones, "United we fall, divided we stand: A study of query segmentation and PRF for patent prior art search," in *Proceedings of PaIR*, 2011.

[64] A. Gibbs, "Boolean patent search: Comparative patent search quality/cost evaluation super Boolean vs. legacy Boolean search engines," Technical Report, http://patentcafe.com, 2006.

[65] M. Giereth, S. Brügmann, A. Stäbler, M. Rotard, and T. Ertl, "Application of semantic technologies for representing patent metadata," in *International Workshop on Applications of Semantic Technologies*, 2006.

[66] M. Giereth, S. Koch, Y. Kompatsiaris, S. Papadopoulos, E. Pianta, L. Serafini, and L. Wanner, "A modular framework for ontology-based representation of patent information," in *Proceedings of the Conference on Legal Knowledge and Information Systems*, 2007.

[67] M. Giereth, S. Koch, M. Rotard, and T. Ertl, "Web based visual exploration of patent information," in *International Conference on Information Visualization, 2007 (IV '07)*, pp. 150–155, July 2007.

[68] J. Gobeill, E. Pasche, D. Teodoro, and P. Ruch, "Simple pre and post processing strategies for patent searching in CLEF intellectual property track 2009," in *Proceedings of the Cross-language Evaluation Forum Conference on Multilingual Information Access Evaluation: Text Retrieval Experiments*, Springer, pp. 444–451, 2009.

[69] E. Graf, I. Frommholz, M. Lalmas, and K. van Rijsbergen, "Knowledge modeling in prior art search," in *Advances in Multidisciplinary Retrieval*, Springer, pp. 31–46, 2010.

[70] T. Grego, P. Pezik, F. M. Couto, and D. Rebholz-Schuhmann, *Identification of Chemical Entities in Patent Documents*, Vol. 5518 of *LNCS*. pp. 942–949, Springer, 2009.

[71] H. Gurulingappa, B. Mueller, M. Hofmann-Apitius, and J. Fluck, "Information retrieval framework for technology survey in biomedical and chemistry literature," in *Proceedings of Text Retrieval Conference*, 2011.

[72] H. Gurulingappa, B. Mueller, M. Hofmann-Apitius, R. Klinger, H.-T. Mevissen, C. M. Friedrich, and J. Fluck, "Prior art search in chemistry patents based on semantic concepts and co-citation analysis," in *Proceedings of Text Retrieval Conference*, 2010.

[73] H. Gurulingappa, B. Müller, R. Klinger, H.-T. Mevissen, M. Hofmann-Apitius, J. Fluck, and C. Friedrich, "Patent retrieval in chemistry based on semantically tagged named entities," in *Proceedings of Text Retrieval Conference*, 2009.

[74] J. Guyot, G. Falquet, and K. Benzineb, "UniGE experiments on prior art search in the field of patents," in *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, Springer, pp. 502–507, 2010.

[75] A. Hanbury, N. Bhatti, M. Lupu, and R. Mörzinger, "Patent image retrieval: A survey," in *Proceedings of Workshop on Patent Information Retrieval*, ACM, pp. 3–8, 2011.

[76] A. Hanbury, V. Zenz, and H. Berger, "1st international workshop on advances in patent information retrieval (AsPIRe'10)," *SIGIR Forum*, vol. 44, pp. 19–22, 2010.

[77] P. Hansen, "The information seeking and retrieval process at the Swedish patent office: Moving from lab-based to real life work-task environment," in *Proceedings of Workshop on Patent Retrieval*, 2000.

[78] D. Harman, *Information Retrieval Evaluation*. Morgan & Claypool Publishers, 2011.

[79] C. G. Harris, R. Arens, and P. Srinivasan, "Comparison of IPC and USPC classification systems in patent prior art searches," in *Proceedings of Workshop on Patent Information Retrieval*, ACM, pp. 27–32, 2010.

[80] C. G. Harris, R. Arens, and P. Srinivasan, "Using classification code hierarchies for patent prior art searches," in *Current Challenges in Patent Information Retrieval*, (M. Lupu, K. Mayer, J. Tait, and A. J. Trippe, eds.), Springer, pp. 287–304, 2011.

[81] C. G. Harris, S. Foster, R. Arens, and P. Srinivasan, "On the role of classification in patent invalidity searches," in *Proceedings of Workshop on Patent Information Retrieval*, pp. 29–32, 2009.

[82] Z.-L. He and M. Deng, "The evidence of systematic noise in non-patent references: A study of New Zealand companies' patents," *Scientometrics*, vol. 72, pp. 149–166, 2007.

[83] B. Herbert, G. Szarvas, and I. Gurevych, "Prior art search using international patent classification codes and all-claims-queries," in *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, Springer, pp. 452–459, 2010.

[84] J. D. Holliday and P. Willet, "Representation and searching of chemical-structure information in patents," in *Current Challenges in Patent Information Retrieval*, (M. Lupu, K. Mayer, J. Tait, and A. J. Trippe, eds.), Springer, 2011.

[85] V. Hristidis, E. Ruiz, A. Hernández, F. Farfán, and R. Varadarajan, "Patentssearcher: A novel portal to search and explore patents," in *Proceedings of Workshop on Patent Information Retrieval*, ACM, pp. 33–38, 2010.

[86] B. Huet, G. Guarascio, N. J. Kern, and B. Mérialdo, "Relational skeletons for retrieval in patent drawings," in *Proceedings of International Conference on Image Processing*, pp. 737–740, 2001.

[87] D. Hull, S. Aït-Mokhtar, M. Chuat, A. Eisele, E. Gaussier, G. Grefenstette, P. Isabelle, C. Samuelsson, and F. Segond, "Language technologies and patent search and classification," *World Patent Information*, vol. 23, pp. 265–268, 2001.

[88] N. Ide and K. Suderman, "Integrating linguistic resources: The american national corpus model," in *Proceedings of International Conference on Language Resources and Evaluation (LREC)*, 2006.

[89] H. Itoh, H. Mano, and Y. Ogawa, "Term distillation in patent retrieval," in *Proceedings of ACL Workshop on Patent Corpus Processing*, 2003.

[90] M. Iwayama, A. Fuji, and N. Kando, "Overview of classification subtask at NTCIR-6 patent retrieval task," in *Proceedings of NII Test Collection for IR Systems-6*, 2007.

[91] M. Iwayama, A. Fujii, and N. Kando, "Overview of classification subtask at NTCIR-5 patent retrieval task," in *Proceedings of NII Test Collection for IR Systems-5*, Tokyo, Japan, 2005.

[92] M. Iwayama, A. Fujii, N. Kando, and Y. Marukawa, "An empirical study on retrieval models for different document genres: patents and newspaper articles," in *Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 251–258, 2003.

[93] M. Iwayama, A. Fujii, N. Kando, and A. Takano, "Overview of patent retrieval task at NTCIR-3," in *Proceedings of ACL Workshop on Patent Corpus Processing*, 2003.

[94] A. Järvelin, G. Eriksson, P. Hansen, T. Tsikrika, A. G. S. de Herrera, M. Lupu, M. Gäde, V. Petras, S. Rietberger, M. Braschler, and R. Berendsen, "Deliverable 2.2 revised specification of the evaluation tasks," Technical Report, PROMISE Network of Excellence, 2012.

[95] C. Jochim, C. Lioma, and H. Schütze, "Expanding queries with term and phrase translations in patent retrieval," in *Multidisciplinary Information Retrieval*, Springer, pp. 16–29, 2011.

[96] H. Joho and M. Sanderson, "Document frequency and term specificity," in *RIAO: Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*, 2007.

[97] N. Kando and M.-K. Leong, "Workshop on patent retrieval (SIGIR 2000 workshop report)," *SIGIR Forum*, vol. 34, pp. 28–30, 2000.

[98] I.-S. Kang, S.-H. Na, J. Kim, and J.-H. Lee, "Cluster-based patent retrieval," *Information Processing and Management*, vol. 43, pp. 1173–1182, September 2007.

[99] N. Kemp and M. Lynch, "Extraction of information from text of chemical patents. 1. identification of specific chemical names," *Journal of Chemical Information and Computer Sciences*, vol. 38, 1998.

[100] Y. Kim, J. Seo, and W. B. Croft, "Automatic Boolean query suggestion for professional search," in *Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 825–834, 2011.

[101] K. Kishida, "Pseudo relevance feedback method based on Taylor expansion of retrieval function in NTCIR-3 patent retrieval task," in *Proceedings of the ACL Workshop on Patent Corpus Processing*, 2003.

[102] K. Kishida, K.-H. Chen, S. Lee, H.-H. Chen, N. Kando, K. Kuriyama, S. H. Myaeng, and K. Eguchi, "Cross-lingual information retrieval (CLIR) task at the NTCIR workshop 3," *SIGIR Forum*, vol. 38, pp. 17–20, July 2004.

[103] I. Klampanos, H. Azzam, and T. Roelleke, "A case for probabilistic logic for scalable patent retrieval," in *Proceedings of Workshop on Patent Information Retrieval*, 2009.

[104] R. Klinger, C. Kolářik, J. Fluck, M. Hofmann-Apitius, and C. M. Friedrich, "Detection of iupac and iupac-like chemical names," in *Proceedings of International Conference on Intelligent Systems for Molecular Biology (ISMB)*, 2008.

[105] S. Koch, H. Bosch, M. Giereth, and T. Ertl, "Iterative integration of visual insights during scalable patent search and analysis," *Transactions on Visualization and Computer Graphics*, vol. 17, 2011.

[106] C. H. A. Koster, "Text mining for intellectual property," in *Proceedings of Dutch-Belgian Information Retrieval Workshop (DIR)*, 2010.

[107] C. H. A. Koster, M. Seutter, and J. Beney, "Multi-classification of patent applications with Winnow," in *Perspectives of System Informatics*, Springer, pp. 546–555, 2004.

[108] A. Krishnan, A. F. Cardenas, and D. Springer, "Search for patents using treatment and causal relationships," in *Proceedings of Workshop on Patent Information Retrieval*, New York, NY, USA, pp. pp. 1–10, 2010.

[109] J.-C. Lamirel, S. A. Shehabi, M. Hoffmann, and C. Francois, "Intelligent patent analysis through the use of a neural network: Experiment of multi-viewpoint analysis with the multisom model," in *Proceedings of the ACL Workshop on Patent Corpus Processing*, 2003.

[110] L. S. Larkey, "A patent search and classification system," in *Proceedings of ACM Conference on Digital Libraries*, pp. 179–187, 1999.

[111] A. Leach and V. Gillet, *An Introduction to Chemoinformatics*. Springer, 2007.

[112] G. Leech, "100 million words of english: The british national corpus," *Language Research*, vol. 28, 1992.

[113] J. Leveling, W. Magdy, and G. J. Jones, "An investigation of decompounding for cross-language patent search," in *Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1169–1170, 2011.

[114] L. Li and C. L. Tan, "Associating figures with descriptions for patent documents," in *Proceedings of the IAPR International Workshop on Document Analysis Systems*, pp. 385–392, 2010.

[115] Y. Li and J. Shawe-Taylor, "Advanced learning algorithms for cross-language patent retrieval and classification," *Information Processing and Management*, vol. 43, pp. 1183–1199, 2007.

[116] Y.-R. Li, L.-H. Wang, and C.-F. Hong, "Extracting the significant-rare keywords for patent analysis," *Expert Systems with Applications*, vol. 36, 2009.

[117] W. A. Lise, "An investigation of terminology and syntax in Japanese and US patents and the implications for the patent translator," http://www.lise.jp/patsur.html, 2011. Last visited: September, 4, 2012.

[118] J. List, "How drawings could enhance retrieval in mechanical and device patent searching," *World Patent Information*, vol. 29, pp. 210–218, 2007.

[119] P. Lopez, "Automatic extraction and resolution of bibliographical references in patent documents," in *Advances in Multidisciplinary Retrieval*, Springer Berlin/Heidelberg, pp. 120–135, 2010.

[120] P. Lopez and L. Romary, "Experiments with citation mining and key-term extraction for prior art search," in *Cross-Language Evaluation Forum (Notebook Papers/Labs/Workshop)*, 2010.

[121] P. Lopez and L. Romary, "Patatras: Retrieval model combination and regression models for prior art search," in *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, Springer, 2010.

[122] B. Lu, B. K. Tsou, T. Jiang, O. Y. Kwong, and J. Zhu, "Mining large-scale parallel corpora from multilingual patents: An English-Chinese example and its application to SMT," in *Proceedings of the CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pp. 79–86, 2010.

[123] X. Lu, S. Kataria, W. J. Brouwer, J. Z. Wang, P. Mitra, and C. L. Giles, "Automated analysis of images in documents for intelligent document search," *International Journal on Document Analysis and Recognition*, vol. 12, pp. 65–81, 2009.

[124] M. Lupu, "The status of retrieval evaluation in the patent domain," in *Proceedings of Workshop on Patent Information Retrieval*, 2011.

[125] M. Lupu, "Patolympics — an infrastructure for interactive evaluation of patent retrieval tools," in *Proceedings of CIKM Workshop on Data infrastructures for Supporting Information Retrieval Evaluation (DESIRE)*, 2011.

[126] M. Lupu, Z. Jiashu, J. Huang, H. Gurulingappa, I. Filipov, and J. Tait, "Overview of the trec 2011 chemical ir track," in *Proceedings of Text Retrieval Conference*, 2011.

[127] M. Lupu, K. Mayer, J. Tait, and A. J. Trippe, eds., *Current Challenges in Patent Information Retrieval. Information Retrieval Series*. Springer, 2011.

[128] M. Lupu, F. Piroi, and A. Hanbury, "Aspects and analysis of patent test collections," in *Proceedings of Workshop on Patent Information Retrieval*, 2010.

[129] M. Lupu, F. Piroi, J. Huang, J. Zhu, and J. Tait, "Overview of the trec chemical ir track," in *Proceedings of Text Retrieval Conference*, 2009.

[130] M. Lupu, R. Schuster, R. Mörzinger, F. Piroi, T. Schleser, and A. Hanbury, "Patent images — a glass-encased tool: Opening the case," in *Proceedings of International Conference on Knowledge Management and Knowledge Technologies*, pp. 16:1–16:8, 2012.

[131] W. Magdy and G. Jones, "A study on query expansion methods for patent retrieval," in *Proceedings of Workshop on Patent Information Retrieval*, 2011.

[132] W. Magdy and G. J. Jones, "Pres: A score metric for evaluating recall-oriented information retrieval applications," in *Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 611–618, 2010.

[133] W. Magdy and G. J. F. Jones, "Examining the robustness of evaluation metrics for patent retrieval with incomplete relevance judgements," in *Proceedings of the 2010 International Conference on Multilingual and Multimodal Information Access Evaluation: Cross-language Evaluation Forum*, Berlin, Heidelberg, pp. 82–93, 2010.

[134] W. Magdy, J. Leveling, and G. J. F. Jones, "Exploring structured documents and query formulation techniques for patent retrieval," in *Proceedings of the Cross-language Evaluation Forum Conference on Multilingual Information Access Evaluation: Text Retrieval Experiments*, Springer-Verlag, pp. 410–417, 2009.

[135] P. Mahdabi, L. Andersson, A. Hanbury, and F. Crestani, "Report on the CLEF-IP 2011 experiments: Exploring patent summarization," in *Cross-Language Evaluation Forum (Notebook Papers/LABs/Workshops)*, 2011.

[136] P. Mahdabi, M. Keikha, S. Gerani, M. Landoni, and F. Crestani, "Building queries for prior-art search," in *Multidisciplinary Information Retrieval*, Springer-Verlag, pp. 3–15, 2011.

[137] F. Mahmoudi, J. Shanbehzadeh, A.-M. Eftekhari-Moghadam, and H. Soltanian-Zadeh, "Image retrieval based on shape similarity by edge orientation autocorrelogram," *Pattern Recognition*, vol. 36, pp. 1725–1736, 2003.

[138] L. Molà, *The Silk Industry of Renaissance Venice*. JHU Press, 2000.

[139] A. Moldovan, R. I. Bot, and G. Wanka, "Latent semantic indexing for patent documents," *International Journal of Applied Mathematics and Computer Science*, vol. 15, 2005.

[140] G. Moradei and P. C. Contessini, "An evaluation of some semantic tools for simple patent searching," in *Proceedings of Patent Information Conference*, 2011.

[141] N. Morey, "Global business solutions for patent prosecution," in *Proceedings of the Symposium of Intellectual Property Authorities*, WIPO, 2011.

[142] R. Mörzinger, A. Horti, G. Thallinger, N. Bhatti, and A. Hanbury, "Classifying patent images," in *Cross-Language Evaluation Forum (Notebook papers/ Labs/Workshop)*, 2011.

[143] S. Mukherjea and B. Bamba, "Biopatentminer: An information retrieval system for biomedical patents," in *VLDB '04: Proceedings of International Conference on Very Large Data Bases*, VLDB Endowment, pp. 1066–1077, 2004.

[144] B. Müller, R. Klinger, H. Gurulingappa, H.-T. Mevissen, M. Hofmann-Apitius, J. Fluck, and C. Friedrich, "Abstracts versus full texts and patents: A quantitative analysis of biomedical entities," in *Advances in Multidisciplinary Retrieval*, Springer, pp. 152–165, 2010.

[145] H. Nanba, A. Fujii, M. Iwayama, and T. Hashimoto, "Overview of the patent mining task at the NTCIR-7 workshop," in *Proceedings of NII Test Collection for IR Systems-7*, 2008.

[146] H. Nanba, A. Fujii, M. Iwayama, and T. Hashimoto, "Overview of the patent retrieval task at the NTCIR-8 workshop," in *Proceedings of NII Test Collection for IR Systems-8*, 2010.

[147] H. Nanba, H. Kamaya, T. Takezawa, M. Okumura, A. Shinmori, and H. Tanigawa, "Automatic translation of scholarly terms into patent terms," in *Current Challenges in Patent Information Retrieval*, (M. Lupu, K. Mayer, J. Tait, and A. J. Trippe, eds.), Springer, pp. 373–388, 2011.

[148] H. Nanba, S. Mayumi, and T. Takezawa, "Automatic construction of a bilingual thesaurus using citation analysis," in *Proceedings of Workshop on Patent Information Retrieval*, ACM, pp. 25–30, 2011.

[149] J.-Y. Nie, *Cross-Language Information Retrieval*. Morgan & Claypool Publishers, 2010.

[150] N. Oostdijk, E. D'hondt, H. van Halteren, and S. Verberne, "Genre and domain in patent texts," in *Proceedings of Workshop on Patent Information Retrieval*, ACM, pp. 39–46, 2010.

[151] N. Oostdijk, S. Verberne, and C. Koster, "Constructing a broad-coverage lexicon for text mining in the patent domain," in *Proceedings of International Conference on Language Resources and Evaluation (LREC)*, 2010.

[152] M. Osborn, T. Strzalkowski, and M. Marinescu, "Evaluating document retrieval in patent database: A preliminary report," in *Proceedings of Conference on Information and Knowledge Management*, pp. 216–221, 1997.

[153] P. Parapatics and M. Dittenbach, "Patent claim decomposition for improved information extraction," in *Proceedings of Workshop on Patent Information Retrieval*, ACM, pp. 33–36, 2009.

[154] J. Perez-Iglesias, A. Rodrigo, and V. Fresno, "Using bm25f and kld for patent retrieval," in *Cross-Language Evaluation Forum (Notebook Papers/ LABs/Workshops)*, 2010.

[155] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *IEEE Conference on Computer Vision and Pattern Recognition, 2007 (CVPR '07)*, pp. 1–8, 2007.

[156] F. Piroi, M. Lupu, A. Hanbury, A. Sexton, W. Magdy, and I. Filippov, "Clef-ip 2012: Retrieval experiments in the intellectual property domain," in *Cross-Language Evaluation Forum (Notebook papers/Labs/Workshop)*, 2012.

[157] F. Piroi, M. Lupu, A. Hanbury, and V. Zenz, "Clef-ip 2011: Retrieval in the intellectual property domain," in *Cross-Language Evaluation Forum (Notebook Papers/Labs/Workshop)*, 2011.

[158] F. Piroi and J. Tait, "Clef-ip 2010: Retrieval experiments in the intellectual property domain," in *Cross-Language Evaluation Forum (Notebook papers/Labs/Workshop)*, 2010.

[159] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Foundations and Trends in Information Retrieval*, vol. 3, 2009.

[160] S. E. Robertson, "The probability ranking principle in ir," *Journal of Documentation*, vol. 33, 1977.

[161] G. Roda, J. Tait, F. Piroi, and V. Zenz, "Clef-ip 2009: Retrieval experiments in the intellectual property domain," in *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, Springer, pp. 385–409, 2010.

[162] J. Rousu, C. Saunders, S. Szedmák, and J. Shawe-Taylor, "Kernel-based learning of hierarchical multilabel classification models," *Journal of Machine Learning Research*, vol. 7, pp. 1601–1626, 2006.

[163] J. Ryley, "Latent semantic indexing for patent information," http://cogprints.org/5710/1/ryley.html, 2007. Last accessed: September, 4, 2012.

[164] M. Sahlgren, P. Hansen, and J. Karlgren, "English-japanese cross-lingual query expansion using random indexing of aligned bilingual text data," in *Proceedings of NII Test Collection for IR Systems*, 2002.

[165] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing and Management*, vol. 24, pp. 513–523, 1988.

[166] M. Sanderson, "Test collection based evaluation of information retrieval systems," *Foundations and Trends in Information Retrieval*, vol. 4, 2010.

[167] M. Sanderson, M. L. Paramita, P. Clough, and E. Kanoulas, "Do user preferences and evaluation measures line up?," in *Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2010.

[168] I. Schellner, "Japanese File Index classification and f-terms," *World Patent Information*, vol. 24, pp. 197–201, 2002.

[169] M. W. Seeger, "Cross-validation optimization for large scale hierarchical classification kernel methods," in *Proceedings of Neural Information Processing Systems (NIPS)*, pp. 1233–1240, 2007.

[170] S. Sheremetyeva, "Natural language analysis of patent claims," in *Proceedings of ACL Workshop on Patent Corpus Processing*, 2003.

[171] S. Sheremetyeva and S. Nirenburg, "Knowledge elecitation for authoring patent claims," *IEEE Computer*, vol. 29, pp. 57–63, 1996.

[172] S. Sheremetyeva, S. Nirenburg, and I. Nirenburg, "Generating patent claims from interactive input," in *Proceedings of International Workshop on Natural Language Generation (INLG'96)*, pp. 61–70, 1996.

[173] A. Shinmori, M. Okumura, Y. Marukawa, and M. Iwayama, "Patent claim processing for readability — structure analysis and term explanation," in *Proceedings of The ACL Workshop on Patent Corpus Processing*, 2003.

[174] P. Sidiropoulos, S. Vrochidis, and I. Kompatsiaris, "Content-based binary image retrieval using the adaptive hierarchical density histogram," *Pattern Recognition*, vol. 44, pp. 739–750, 2011.

[175] A. Singh, S. Hallihosur, and L. Rangan, "Changing landscape in biotechnology patenting," *World Patent Information*, vol. 31, pp. 219–225, 2009.

[176] A. Smeaton, "Nlp and ir," in *European Summer School on Information Retrieval*, 1995.

[177] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1349–1380, 2000.

[178] H. Smith, "Automation of patent classification," *World Patent Information*, vol. 24, pp. 269–271, 2002.

[179] S. Taduri, G. T. Lau, K. H. Law, H. Yu, and J. P. Kesan, "Developing an ontology for the U.S. patent system," in *Proceedings of Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times*, pp. 157–166, 2011.

[180] S. Taduri, G. T. Lau, K. H. Law, H. Yu, and J. P. Kesan, "An ontology-based interactive tool to search documents in the u.s. patent system," in *Proceedings of the Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times*, New York, NY, USA, pp. 329–330, 2011.

[181] S. Taduri, H. Yu, G. Lau, K. Law, and J. Kesan, "Developing a comprehensive patent related information retrieval tool," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 6, pp. 1–16, August 2011.

[182] T. Takaki, A. Fuji, and T. Ishikawa, "Associative document retrieval by query subtopic analysis and its application to invalidity patent search," in *Proceedings of Conference on Information and Knowledge Management*, 2004.

[183] T. Tang, N. Craswell, D. Hawking, K. Griffiths, and H. Christensen, "Quality and relevance of domain-specific search: A case study in mental health," *Information Retrieval*, vol. 9, pp. 207–225, 2006.

[184] J. Tangelder and R. Veltkamp, "A survey of content based 3D shape retrieval methods," *Multimedia Tools and Applications*, vol. 39, pp. 441–471, 2008.

[185] W. Thielemann, "Ocr errors in patent full-text documents," in *Proceedings of the IRF Symposium*, 2007.

[186] D. Tikk, G. Biró, and A. Törcsvári, "A hierarchical online classifier for patent categorization," in *Emerging Technologies of Text Mining: Techniques and Applications*, (H. A. do Prado and E. Ferneda, eds.), IGI Global, pp. 244–267, 2007.

[187] A. Tiwari and V. Bansal, "Patseek: Content based image retrieval system for patent database," in *Proceedings of International Conference on Electronic Business (ICEB)*, pp. 1167–1171, 2004.

[188] K. Tombre, "Analysis of engineering drawings: State of the art and challenges," in *Graphics Recognition — Algorithms and Systems*, Springer, pp. 257–264, 1998.

[189] A. Trippe and I. Ruthven, "Evaluating real patent retrieval effectiveness," in *Current Challenges in Patent Information Retrieval*, (M. Lupu, K. Mayer, J. Tait, and A. J. Trippe, eds.), Springer, 2011.

[190] Y.-H. Tseng, C.-J. Lin, and Y.-I. Lin, "Text mining techniques for patent analysis," *Information Processing and Management*, vol. 43, pp. 1216–1247, September 2007.

[191] Y.-H. Tseng and Y.-J. Wu, "A study of search tactics for patentability search: A case study on patent engineers," in *Proceedings of Workshop on Patent Information Retrieval*, pp. 33–36, 2008.

[192] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun, "Support vector machine learning for interdependent and structured output spaces," in *Proceedings of International Conference on Machine Learning*, pp. 104–111, 2004.

[193] J. Urbain and O. Frieder, "Trec chemical ir track 2009: A distributed dimensional indexing model for chemical patent search," in *Proceedings of Text Retrieval Conference*, 2009.

[194] USPTO, "Acceptance, processing, use and dissemination of chemical and three-dimensional biological structural data in electronic format," *Biotechnology Law Report*, vol. 24, pp. 638–644, 2005.

[195] A. T. Valko and A. P. Johnson, "Clide pro: The latest generation of CLiDE, a tool for optical chemical structure recognition," *Journal of Chemical Information and Modeling*, vol. 49, pp. 780–787, 2009.

[196] S. Verberne and E. D'hondt, "Prior art retrieval using the claims section as a bag of words," in *Proceedings of the Cross-language Evaluation Forum Conference on Multilingual Information Access Evaluation: Text Retrieval Experiments*, Springer, pp. 497–501, 2009.

[197] S. Verberne, E. D'hondt, N. Oostdijk, and C. Koster, "Quantifying the challenges in parsing patent claims," in *Proceedings of Workshop on Advances in Patent Information Retrieval (AsPIRe)*, 2010.

[198] S. Verberne, M. Vogel, and E. D'hondt, "Patent classification experiments with the linguistic classification system lcs," in *Working Notes of the CLEF Labs*, 2010.

[199] M. Verma and V. Varma, "Applying key phrase extraction to aid invalidity search," in *Proceedings of International Conference on Artificial Intelligence and Law*, pp. 249–255, 2011.

[200] S. V. Vishwanathan, N. N. Schraudolph, and A. J. Smola, "Step size adaptation in reproducing kernel Hilbert space," *Journal of Machine Learning Research*, vol. 7, pp. 1107–1133, 2006.

[201] S. Vrochidis, S. Papadopoulos, A. Moumtzidou, P. Sidiropoulos, E. Pianta, and I. Kompatsiaris, "Towards content-based patent image retrieval: A framework perspective," *World Patent Information*, vol. 32, pp. 94–106, 2010.

[202] M. Z. Wanagiri and M. Adriani, "Prior art retrieval using various patent document fields contents," in *Cross-Language Evaluation Forum (Notebook Papers/LABs/Workshops)*, 2010.

[203] X. Wang, X. Zhang, and S. Xu, "Patent co-citation networks of fortune 500 companies," *Scientometrics*, vol. 88, pp. 761–770, 2011.

[204] L. Wanner, R. Baeza-Yates, S. Brügmann, J. Codina, B. Diallo, E. Escorsa, M. Giereth, Y. Kompatsiaris, S. Papadopoulos, E. Pianta, G. Piella, I. Puhlmann, G. Rao, M. Rotard, P. Schoester, L. Serafini, and V. Zervaki, "Towards content-oriented patent document processing," *World Patent Information*, vol. 30, pp. 21–33, 2008.

[205] H. Wongel, "The reform of the IPC — consequences for the users," *World Patent Information*, vol. 27, pp. 227–231, 2005.

[206] World Intellectual Property Organisation, "Glossary of terms concerning industrial property information and documentation, appendix iii to part 10, of the wipo handbook on industrial property information and documentation," WIPO Publication No. CD208, 2003.

[207] World Intellectual Property Organisation, "Wipo handbook on intellectual property, chapter 5," http://www.wipo.int/export/sites/www/about-ip/en/iprm/pdf/ch5.pdf, last visited August, 2012.

[208] X. Xue and W. B. Croft, "Automatic query generation for patent search," in *Proceedings of Conference on Information and Knowledge Management*, pp. 2037–2040, 2009.

[209] X. Xue and W. B. Croft, "Transforming patents into prior-art queries," in *Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 808–809, 2009.

[210] M. Yang, G. Qiu, J. Huang, and D. Elliman, "Near-duplicate image recognition and content-based image retrieval using adaptive hierarchical geometric centroids," in *International Conference on Pattern Recognition, 2006 (ICPR 2006)*, pp. 958–961, 2006.

[211] S.-Y. Yang and V.-W. Soo, "Comparing the conceptual graphs extracted from patent claims," in *IEEE International Conference on Sensor Networks, Ubiquitous and Trustworthy Computing (SUTC)*, pp. 394–399, 2008.

[212] Y. Yang, L. Akers, T. Klose, and C. Barcelonyang, "Text mining and visualization tools — impressions of emerging capabilities," *World Patent Information*, vol. 30, pp. 280–293, December 2008.

[213] F. Zaccá and M. Krier, "Automatic categorisation applications at the European Patent Office," *World Patent Information*, vol. 24, pp. 187–196, 2002.

[214] V. Zenz, S. Wurzer, M. Dittenbach, and E. Ambrosi, "On the effects of indexing and retrieval models in patent search and the potential of result set merging," in *Proceedings of Workshop on Advances in Patent Information Retrieval (AsPIRe)*, 2010.

[215] L. Zhao and J. Callan, "Formulating simple structured queries using temporal and distributional cues in patents," in *Proceedings of Text Retrieval Conference*, 2009.

[216] L. Zhao and J. Callan, "How to make manual conjunctive normal form queries work in patents search," in *Proceedings of Text Retrieval Conference*, 2011.

[217] Z. Zhiyuan, Z. Juan, and X. Bin, "An outward-appearance patent-image retrieval approach based on the contour-description matrix," in *Proceedings of Japan-China Joint Workshop on Frontier of Computer Science and Technology (FCST)*, IEEE Computer Society, pp. 86–89, 2007.

[218] Y. Zhou and C. L. Tan, "Chart analysis and recognition in document images," in *Proceedings of International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1055–1058, 2001.

[219] G. Zhu, X. Yu, Y. Li, and D. Doermann, "Learning visual shape lexicon for document image content recognition," in *Proceedings of European Conference on Computer Vision (ECCV)*, Springer, pp. 745–758, 2008.