Advanced Engineering Informatics 27 (2013) 335–345

Contents lists available at SciVerse ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei

Searching in Cooperative Patent Classification: Comparison between keyword and concept-based search

Tiziano Montecchi [a], Davide Russo [a,*], Ying Liu [b]

[a] Department of Industrial Engineering, University of Bergamo, Italy
[b] Department of Mechanical Engineering, National University of Singapore, Singapore

ARTICLE INFO

Article history:
Available online 2 April 2013

Keywords:
Concept-based search
Patent classification
Patent mining

ABSTRACT

International patent corpus is a gigantic source containing today about 80 million of documents. Every patent is manually analyzed by patent officers and then classified by a specific code called Patent Class (PC). Cooperative Patent Classification CPC is the new classification system introduced since January 2013 in order to standardize the classification systems of all major patent offices. Like keywords for papers, PCs point to the core of the invention, describing concisely what they contain inside. Most of patents strategies are based on PC as filter for results therefore the selection of relevant PCs is often a primary and crucial activity. This task is considered particularly challenging and only few tools have been specially developed for this purpose. The most efficient tools are provided by patent offices of EPO and WIPO.

This paper analyzes their PCs search strategy (mainly based on keyword-based engines) in order to identify main limitations in terms of missing relevant PCs (recall) and non-relevant results (precision). Patents have been processed by KOM, a semantic patent search tool developed by the authors. Unlike all other PC search tools, KOM uses semantic parser and many knowledge bases for carrying out a conceptual patent search. Its functioning is described step by step through a detailed analysis pointing out the benefits of a concept-based search vis-à-vis a keyword-based search. An exemplary case is proposed dealing with CPCs describing the sterilization of contact lenses. Comparison could be likewise conducted on other PCs such as International (IPC), European (ECLA) or United States (USPC) patent classification codes.

1. Introduction

Patent database is a strategic source for knowledge management activities because the huge number of technical information contained and their uniqueness can be used for many different design activities: new product development [4,26,15,19], forecasting [6,20], technology transfer [14], problem solving [3,22,27], and many others.

Patent literature is a gigantic source containing many millions of documents (e.g. Espacenet, the most comprehensive database in the world, contains almost 80 million of patents[1]) and it is expanding every year. Thus, the management of this gigantic source is a big challenge especially because most patent searches requests to be exhaustive. In other words, some patent searches (such as patentability, freedom to operate, and validity search) can be invalidated even if one document is missing. Patent researchers agree that no patents search can be considered 100% complete ([2]). The most widely used method for patent searching is the keyword-based search, even though this method has several drawbacks. These drawbacks are reported in the following:

- *Different detail levels of patent descriptions*. Every inventor has his/her own style, and the same concept could be expressed at different detail levels (at abstract level by means of a generic language, while at specific level by technical terms and a very precise language). The main reasons are two, on one side people coming from different areas (technical field, geographical, cultural background, etc.) use different expressions and so different words to express the same concept. On the other side, patentees follow different strategies for writing patents, and sometimes they purposely use very vague or inconsistency terminology for hiding patent content or extending claims validity.
- *Inaccurate terminology*. Patentees often give words a different meaning from their ordinary dictionary definition, using non-precise or wrong words and in some cases they even create new terms to describe their inventions. Moreover, lack of

* Corresponding author.
E-mail addresses: tiziano.montecchi@unibg.it (T. Montecchi), davide.russo@unibg.it (D. Russo), mpeliuy@nus.edu.sg (Y. Liu).

[1] EPO Patent Information Conference 2012 (6–8 November 2012, Hamburg, Germany). http://www.epo.org/news-issues/press/speeches/20121109.html

1474-0346/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.aei.2013.02.002

Table 1
Comparison between the major patent classifications present in literature [9].

Info	CPC	IPC	ECLA/ICO	USPC	FI	F-term
No. of entries	≅250,000	≅70,000	≅245,000	≅160,000	≅190,000	≅375,000
Coverage of patent offices	PCT minimum and US	More than 100 countries	PCT minimum	US only	JP only	JP only
Coverage of patent docs	EPO and US docs	More than 37 million	More than 28 million	Around 15 million	More than 35 million	More than 35 million
Coverage of technical fields	All fields	All fields	All fields	All fields	All fields	Approx. 70% of all fields
Official languages	English	English, French	English, French	English	Japanese, English[a]	Japanese, partially in English
Issuing office	EPO and USPTO	WIPO	EPO	USPTO	JPO	JPO

[a] Not all PCs are translated in English.

standard names for developing technologies, devices or machines lead to use many different terms. Finally, in case that a function is obtained by a logical sequence of actions, patent writers could totally or partially omit them, just using one instead of all, or simply citing the most general. For example, "a laser beam lighting a surface to generate both an overheating and a chemical decomposition, so causing a localized ablation" can be otherwise expressed just by "laser cutting".

- *Different official languages*. Many languages can be used for writing patent documents. While, many documents can be machine translated but their translations could be incorrect.

This list of shortcomings is not complete, but it allows us to understand why a patent search based on keywords cannot be considered exhaustive. These drawbacks lead us to consider how to search patents in an alternative way. A more effective strategy to overcome these drawbacks is based on the exploitation of a tool that is characteristic of patent literature only: the patent classification [1,24]. Different patent offices have developed different patent classification systems and many studies aim to give the background information for their use ([7,17,8]). The major classifications are the International Patent Classification (IPC) provided by the World Intellectual Property Organization (WIPO), the European Classification system (ECLA) and In Computer Only (ICO) by European Patent Office (EPO) both are derived from IPC, the United States Patent Classification (USPC) by United States Patent and Trademark Office (USPTO), F-Index (derived from IPC) and F-term by Japanese Patent Office (JPO). Patent offices consider classifications very strategic tools for patent activities, but these classifications have some differences, they cover different documents and patent offices (see Table 1). This is the reason why EPO and USPTO decided to join together for developing a common system: the Cooperative Patent Classification (CPC). This new classification (in force since January 2013) covers all EPO and US classified documents. CPC system is based on IPC structure and includes three classifications: ECLA, ICO and USPC. This classification contains 250,000 classes, the highest number of subdivisions, thus it is the most granular and precise classification among those in English version.

The purpose of these classifications is to briefly describe the invention granted in each patent and they are used to classify and search documents. In particular, each patent is marked by patent officers with one or more appropriate PC codes. Each PC is defined by a description (or title) and identified by a precise code (or symbol), see Table 2. Each PC can be allotted to patent documents (patent applications, specifications of granted patents, utility models, etc.) according to the technical fields the documents pertain to. This classification is arranged into a hierarchy consisting of multiple levels, from the most general to the most specific level, shown in Table 2.

PC descriptions form a controlled vocabulary that patent search experts can use for searching an invention of interest by the selection of the nearest PC definition. PCs are used as filters to limit the research in a precise patent space [1,25,24], due to the fact that patent classifications are language-independent and allow us to search using concepts instead of words [5]. The main limits in the use of patent classifications are given by the high number of PCs, their complex and heterogeneous definitions and the difficulty to find all the relevant PCs for our research. The manual search is a way for finding relevant PCs through the hierarchical browsing of PC descriptions,[2] but it can be very time consuming, tedious and strongly dependent on user's ability and experience. This is the reason why automatic tools for selecting the relevant PCs are needed. Unfortunately, nowadays only very few methods for finding PCs are present in literature. Vijvers [23] and White [24] proposed two methods that comprise a keyword-based patent search and then to study the classifications of the patents obtained, while Valkonen and Nykänen [21] support the user to navigate through the classification answering questions till the right PC is found (using an inference engine). This is possible due to a conceptual pre-processing of the IPC classification, unfortunately only a small part of IPC is covered. However, the most widespread tools are based on keyword search and they are provided by the patent offices of EPO and WIPO. It can be demonstrated (see Section 4) that patent search tools based on keywords have a low recall in finding relevant PCs. For this purpose, the authors propose to use KOM [12,13,14], a concept-based and semantic tool for searching patents in order to find relevant PCs and compare the effectiveness of existing tools. The next section presents a review on the existing keyword-based systems for searching PCs. The algorithm for the extraction of PCs based on the conceptual patent search by KOM is described in Section 3 and a case study supported by a results comparison is shown in Section 4. Section 5 concludes.

2. Prior art on keyword-based search tools for finding PCs

The most known tools for finding PCs are provided by two patent offices: EPO and WIPO. These tools are keyword-based and they can be divided in two main groups according to which source of information they use for searching: PC description or patent text (see Fig. 1).

2.1. Tools for searching on PC descriptions

Some tools working on PC description are present at the state of the art, such as *Term Search*[3] of WIPO (see Fig. 2) and *IPC search*[4] by Deutsches Patent und Markenamt. They work with the same general functioning so for the purpose of the paper we take into account only the Term Search of WIPO as the representative tool. The Term Search

[2] IPC descriptions. Example of PCs schema to be used for manual search: http://web2.wipo.int/ipcpub/#refresh=page

[3] http://web2.wipo.int/ipcpub/fulltextsearch/#version=20120101&lang=en

[4] http://depatisnet.dpma.de/ipc/language.do?lang=EN