Computers in Industry 72 (2015) 1–13

Contents lists available at ScienceDirect

Computers in Industry

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

Managing technological knowledge of patents: HCOntology, a semantic approach

Maria Bermudez-Edo [a,*], María V. Hurtado [b], Manuel Noguera [b], Nuria Hurtado-Torres [c]

[a] University of Surrey, Institute for Communication Systems (ICS), Guildford, United Kingdom
[b] University of Granada, Department of Software Engineering, E.T.S.I.I., c/Saucedo Aranda s/n, 18071 Granada, Spain
[c] University of Granada, Department of Management, Business School, Campus Cartuja s/n, 18071 Granada, Spain

ARTICLE INFO

Article history:
Received 13 October 2014
Received in revised form 9 March 2015
Accepted 13 March 2015
Available online 15 May 2015

Keywords:
Ontology
OWL
Patent
IPC classification

ABSTRACT

Patent data provide technological information essential to define strategies and decisions in the context of firm innovative processes. At present, information regarding patents is usually represented and stored in large databases. Information from these databases is commonly retrieved in the form of files with a CSV- or XML-based codification but with little semantics that enable the inference of further relationships among patents. In these databases, each patent is associated with a technological field by a code. Although the codes assume a hierarchical classification approach, inclusion/subsumption relationships are not explicitly specified such that computers can process them automatically. In recent years, ontologies have been proven to facilitate the exchange of information between people and systems. In this context, the Web Ontology Language (OWL), whose formal semantics are based on description logics, has become the most widely used language for the representation of ontologies. Certain patent ontologies have already been developed in OWL to benefit from the semantics of patent information. However, none have fully exploited the information that can be derived from the formal representation of patent code classification hierarchies through description-logics-based reasoning. This paper presents an approach to automatically translate the hierarchies found in the patent classification codes into concept hierarchies. This proposal also enables the automatic inference of implicit knowledge based on reclassification techniques and relationships between different application domains without changing the applications that make use of patent information. Several examples are presented to illustrate the applicability of the proposal and how it can assist firms in patent information management.

1. Introduction

Innovation is an important driver of both firm success and national economic growth. In the context of a growing necessity for investments that develop innovation, patents are relevant tools for protection of the firm's innovative ideas. Patents are legal documents that protect the rights of the inventor of an industrial property. A patent document provides different information about the innovation: the applicant (or patenting firm), geographic location, date, the technological field, and some information about the other patents that it cites, among others. All of these data that describe or are related to patent documents are called patent metadata [1], and used with other data they are particularly useful to analyze dimensions of innovation that are of policy interest, such as the relationship between patents and economic performance, entrepreneurship, etc. [2,3].

Patent documents are usually stored in large databases of patent offices (e.g. those maintained by the European Patent Office (EPO) and those under supervision of the United States Patent and Trademark Office (USPTO)). These databases often use different data structures that make it difficult to interoperate or automatically and efficiently process the information contained therein. This difficulty is even more important in the context of the different objectives (informative and infringement searches, etc.) of the agents using these databases (e.g. patent lawyers, managers of start-up firms, scientists, and managers of competitors in the industry) [4].

In this context, ontologies have proved to be useful for sharing information by providing formal, uniform and shareable representations about a domain. Several pieces of work have proposed

* Corresponding author. Tel.: +44 1483 689760.
E-mail address: m.bermudez@surrey.ec.uk (M. Bermudez-Edo).

http://dx.doi.org/10.1016/j.compind.2015.03.010
0166-3615/

the use of an ontology-based approach to represent patent metadata using the Web Ontology Language (OWL) (e.g. [1,5]). The main objective of these works is to provide a semantically, well-defined and homogeneous representation of the main types of patent metadata. The use of ontologies enables the representation of knowledge and the identification of context and dependency information more easily than using database-centric structures and interfaces [1]. Likewise, ontologies are increasingly being used in several domains related to innovation and patent registration (Ahmet et al., 2011), including economic and management business areas such as logistics [6], knowledge-commerce [7], e-commerce [8], team organization [9], manufacturing [10] or business processes [11–13]. Furthermore, the power of reasoning in ontologies allows different domains, represented through ontologies, to be integrated, such as patents associated with economics and management studies [14].

The technological field of a patent document represents relevant metadata information about the patent. Patent databases use codes for technological fields that assume a hierarchical classification that delimits the categories a patent may pertain to or be associated with. This field is widely used in searches in databases to determine the field(s) in which a firm may infringe upon another company's industrial rights or where there exists a gap in the technology in which a company could innovate. However, in the patent databases, this hierarchy is not explicitly described and cannot be automatically processed by computers.

Previous patent ontologies have also represented technological fields by using codes. However, they do not fully exploit the formal representation of patent classification hierarchies of these technological fields, and basically mirror the technological patent codes of patent offices databases without leveraging further reasoning capabilities.

In summary, the motivations of this paper in relation to patent information and ontologies are:

- Technological patent codes are defined in patent databases according to hierarchical classifications of technological patent fields, but these hierarchies are not explicitly specified for automatic processing by computers.
- Previous pieces of work on patent metadata ontologies do not represent the hierarchies of the technological patent codes either.
- In current patent ontologies it is not possible to infer transversal relationships between the different metadata that describe a patent document through the hierarchies of concepts of technological patent codes.

This paper draws on the well-known Noy and McGuinness's methodology [15] to develop the proposed ontology by automating some of the steps and focusing on hierarchical codes. Specifically, this paper proposes a method to automatically build and populate patent metadata ontologies by indexing hierarchical codes, which can be retrieved from different patent repositories, by defining ontological categories which enrich the information retrieval process with new relationships, properties and enable the inference of new knowledge. In particular, the paper first studies the characteristics of the hierarchical codes, such as the structure they follow. Second, we propose a representation of these hierarchical codes in OWL. And third, the hierarchy of the technological patent codes is translated from XML into the formal Web Ontology Language OWL 2 using XSLT (eXtensible Stylesheets Transformation Language) [16], according to the characteristics of the hierarchical codes, and the representation proposed.

Additionally, we show the power of reasoning that OWL offers in combination with the hierarchy of concepts (OWL classes) that we have created to infer new information in different fields through transversal relations among concepts from different (but related) domains. The aim is to benefit from linking patent information with external information by providing several methods to infer knowledge in different fields and to connect different knowledge sources. We use OWL-DL, which is the sublanguage of OWL that support maximum expressiveness without losing computational completeness. We use the methodology proposed in Bermudez-Edo et al. [17] for inferring new knowledge with existing reasoning tools.

To illustrate the applicability of our proposal, this paper shows how new information can be inferred from the hierarchy of patent metadata concepts. In particular, the first case study demonstrates how a suitable ontological representation of patent metadata enables the automatic reclassification of patents when a new technological patent code appears. In the second case, a proposal is introduced to link two different knowledge domains by specifying new relationships between the representation of patent classifications and an external classification, the classification of industrial sectors provided by the United Nations Statistics Division (UNSD),[1] which can help in the elaboration of economics studies that evaluate the innovation level of firms based on patent indicators.

The remainder of the paper is organized as follows. Section 2 describes the related work. Section 3 explains our proposal on how to translate the hierarchical technological patent codes into hierarchies of concepts and accomplish the population of the ontology. Section 4 presents motivating examples of the reasoning and the inference of new information from the hierarchical codes within the patent domain and in multiple domains. Section 5 discusses the contributions of the research. Finally, Section 6 concludes with a short summary and proposals for future research.

2. Patent ontology related works

Several patent ontologies have been proposed so far for a semantically well-defined and homogeneous representation for the major types of patent metadata. The most prominent examples are the ontology created within the European *Patexpert* project [1,18] and the *PatentOntology* from Stanford University [5].

Patexpert was created to homogeneously represent different patent information from several EPO databases and to provide it with semantic meaning. However, Patexpert does not merge information retrieved from different patent offices. The patent metadata ontology has been populated by XSLT stylesheets. Unfortunately, the public version of this ontology[2] is not populated, but to the best of our knowledge this ontology does not automatically represent the semantics of the hierarchy of technological patent codes.

PatentOntology was developed to avoid the limitations of Patexpert when integrating heterogeneous domains [19]. PatentOntology merges information from USPTO patent documents retrieved from the USPTO database with information from patent courts of USPTO from the LexisNexis database [20]. Although this ontology has been populated with a parser, it does not automatically retrieve the semantics of the hierarchy of technological patent codes and does not merge information from different patent offices. The technological patent codes classify innovations into fields of activity to facilitate the searches of interested agents. This technological field is widely used to delimit the scope of the searches and is one of the most used items of patent metadata [4].

There are also patent ontologies based only on keywords found in patents [14,21,22]. These ontologies allow retrieving information from the text of the patent documents, however they do not

[1] http://unstats.un.org/unsd/cr/registry/regcst.asp?Cl=27.
[2] http://mklab.iti.gr/project/patexpert.