ELSEVIER

World Patent Information 30 (2008) 21–33

WORLD
PATENT
INFORMATION

www.elsevier.com/locate/worpatin

Towards content-oriented patent document processing

Leo Wanner [a,c,*], Ricardo Baeza-Yates [a,c], Sören Brügmann [b], Joan Codina [c], Barrou Diallo [d], Enric Escorsa [e], Mark Giereth [f], Yiannis Kompatsiaris [g], Symeon Papadopoulos [g], Emanuele Pianta [h], Gemma Piella [c], Ingo Puhlmann [i], Gautam Rao [e], Martin Rotard [f], Pia Schoester [i], Luciano Serafini [h], Vasiliki Zervaki [g]

[a] *Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain*
[b] *Brügmann Software, Bokeler Straße 18, 26871 Papenburg, Germany*
[c] *Pompeu Fabra University, Passeig de Circumval.lació, 8, 08003 Barcelona, Spain*
[d] *European Patent Office, Postbus 5818, 2280 HV Rijswijk, The Netherlands*
[e] *IALE Inc., Balmes, 48, 2° 1ª, 08007 Barcelona, Spain*
[f] *University of Stuttgart, Universitätsstr. 38, 70569 Stuttgart, Germany*
[g] *Informatics and Telematics Institute, 1st Km Thermi-Panorama Road, Thermi-Thessaloniki 57001, Greece*
[h] *Istituto di Cultura de Trentino, Via Sommarive, 18, 38050 Povo-Trento, Italy*
[i] *Fraunhofer Gesellschaft, Leonrodstraße 68, 80636 Munich, Germany*

Abstract

In this article, we present ongoing work on an advanced patent processing service PATExpert. The central assumption underlying PATExpert is that in order to meet the needs of the users of patent processing services, recourse must be made to the content of patent material. We introduce a content representation schema for patent documentation and sketch the design of techniques that facilitate the integration of this schema into the patent processing cycle. Two types of techniques are discussed. Techniques of the first type facilitate the access to the content of patent documentation provided in a textual format – be it by the human reader or by the machine – in that they rephrase and summarize the documentation and map it onto a formal semantic representation. Techniques of the second type operate on the content representation. At this stage, PATExpert is explored in two technology areas – optical recording devices and machine tools. The work is being carried out in the framework of an R&D-project partially funded by the European Commission.

Keywords: Patent content representation; Patent retrieval; Content extraction; Paraphrasing; Summarization; Visualization; Navigation; Valuing; PATExpert; Classification; Translation; Documentation ontologies; Knowledge base

1. Introduction

Currently, patent material is maintained in a textual format (be it in electronic or paper form). In order to retrieve, classify, interpret or assess it, the user must hypothesize how surface textual clues reflect the content. This is costly and the positive outcome is less than guaranteed. An alternative would be to specify the content representation of patent material explicitly in terms of a formal and unambiguous semantic representation. The advantages of this alternative are obvious. On the one hand, such a representation would make the examination and invalidation (by both machine and humans) much more straightforward and, on the other hand, it would facilitate retrieval, classification and interpretation of patent material. As a consequence, the patent processing techniques would be

* Corresponding author. Address: Institució Catalana de Recerca i Estudis Avançats (ICREA), Pompeu Fabra University, Passeig de Circumval.lació, 8, 08003 Barcelona, Spain. Tel.: +34 935422963; fax: +34 93542517.
E-mail address: leo.wanner@upf.edu (L. Wanner).

0172-2190/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.wpi.2007.03.008

semantics-driven, which would imply a change of the paradigm in patent processing from textual (viewing patents as text blocks enriched by "canned" picture material, or sequences of morpho-syntactic tokens) to semantic (viewing patents as multimedia knowledge objects). The recent advances in semantic web technologies [1] and the determination of the steering institutions to formalize the input and processing formats of patent documentation[1] speak for the implementation of the paradigm change. We are convinced that in the long run, this change will culminate in the compilation of *patent knowledge bases* (instead of or along with patent data bases).

Two strategies can be pursued to obtain a patent knowledge base: (1) extraction of the content from patent material rendered in text format and its subsequent mapping onto the content representation; (2) explicit representation of patent material in terms of a content representation (such that patent applications are already submitted as formal semantic descriptions).[2] The second strategy is more straightforward and more reliable. However, given the vast amount of patent material available in text format and taking into account that the text format continues to be the unique format of patent documentation, the first strategy is for the time being more practical. Unfortunately, a closer look at the state of the art techniques reveals that they do not fully account for the implementation of this strategy. Even recent initiatives that stress the importance of semantics and seek to develop techniques that extract the content of patent documentation for further use fall short of obtaining a true semantic representation since they rely exclusively upon surface-oriented criteria such as term frequency, term co-occurrence, and morpho-syntactic categories of the terms (i.e., noun, verb, adjective, etc.). In other words, the use of semantic web oriented notations for the resulting representation does not automatically imply that this representation is indeed a semantic (=content) representation. In order to obtain the representation of the content of a given document, "deep" analysis is required, and, in order to be able to make proper use of the content representation, knowledge-oriented techniques that operate on content rather than on the text surface are required.

PATExpert[3] addresses the problem of meaning representation and processing of patent documentation. The goal of PATExpert is twofold: (i) to push forward the adoption of the semantic paradigm for patent processing; (ii) to provide the user techniques for better access to the content of textual patent documentation. To achieve this goal, PATExpert focuses on the following four topics:

- content representation that is suitable for the description of inventions in several technology areas,
- semantics-based techniques that operate on the content representation of patent documentation,
- techniques that facilitate the mapping of the existing textual patent documentation to its content representation,
- techniques that facilitate a better access to the content of textual patent documentation.

In this article, we present PATExpert's general approach to these four topics.[4] The feasibility of this approach will be demonstrated within the life time of the still ongoing project for two technology areas: optical recording devices and machine tools.

The remainder of the article is structured as follows. We assume that the representation required for encoding the content of patent documentation must depend on the techniques that make use of it, and the techniques, in their turn, must reflect the needs of the users. Therefore, we start with the analysis of the needs of the users and an assessment of the consequences of these needs for the definition of the semantic representation (Section 2). Section 3 provides a sketch of the content representation framework in PATExpert. In Section 4, first the architecture of the PATExpert-service is presented and then the individual modules that realize the whole range of techniques offered by the service are discussed. Section 5, finally, contains a short summary and an outline of the future work plan within PATExpert.

2. Patent content representation from the user's point of view

The available commercial and experimental patent processing services can be assumed to reflect the central needs of the users[5] – although, obviously, only to the extent to which the state of the art allows for the implementation of a technique that meets a specific need of the user. Thus an attempt to meet some of the user needs requires a workaround. In this case, a deeper analysis of the service is required to identify the real need of the user behind the implemented technique.

In this section, we first examine the central services offered so far and draw then conclusions for the definition of an adequate content representation framework.

[1] ST36, which defines the XML-based format of patent(s) (applications), provides evidence for this determination.

[2] Obviously, an editor supporting the authoring of such semantic descriptions would be needed.

[3] PATExpert [2] is partially funded by the European Commission in its Sixth Framework Programme (FP6 028116).

[4] Note, however, that the development and implementation of the individual techniques within the showcase that will demonstrate the viability of PATExpert's approach has not yet been terminated.

[5] When we speak of users, we primarily mean professional examiners of patent applications, inventors and patent offices of research and industrial institutions and patent lawyer's agencies. These user profiles are represented either by members of the PATExpert-Consortium or by clients of members of the PATExpert-Consortium, which have been interviewed to obtain information on their needs. All user requirements are summarized in an internal working document of the project.