World Patent Information 54 (2018) 28–38

Contents lists available at ScienceDirect

World Patent Information

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

Automatic users extraction from patents

Filippo Chiarello[a,*], Andrea Cimino[b], Gualtiero Fantoni[c], Felice Dell'Orletta[b]

[a] Department of Energy, Systems, Territory and Construction Engineering, University of Pisa, Largo Lucio Lazzarino, 2, 56126, Pisa, Italy
[b] Institute for Computational Linguistics of the Italian National Research Council (ILC-CNR), Via G. Moruzzi,1, Pisa, Italy
[c] Department of Mechanical, Nuclear and Production Engineering, University of Pisa, Largo Lucio Lazzarino, 2, 56126, Pisa, Italy

ARTICLE INFO

Keywords:
Patent analysis
Deep learning
Text mining
User of an invention

ABSTRACT

Patents contain a large quantity of information which is usually neglected. This information is hidden beneath technical and juridical jargon and therefore so many potential readers cannot take advantage of it. State of the art natural language processing tools and in particular named entity recognition tools, could be used to detect valuable concepts in patent documents.

The purpose of the present research is to design a method capable of automatically detecting and extracting one of the multiple entities hidden in patents: the users of the invention.

The method is based on a new approach tailored for users extraction by integrating state-of-the-art computational linguistics tools with a large knowledge base. Furthermore the paper shows a comparison among different machine learning algorithms with the twofold aim of achieving the highest recall and evaluating the performance in terms of precision and computational effort.

Finally, a case study on two patent sets has been conducted to evaluate the effectiveness and the output of the entire tool-chain.

1. Introduction

Nowadays patent data can be used for planning technological strategy [1]. The focus on the technological usefulness of patent data is certainly a great advantage, but this huge research area could hide other useful application for patents. For example, in Ref. [2] the authors consider on one side patents as a source to collect information about technologies and products, and on the other side manuals, handbooks and market reports to collect market information. Since patents are only technological documents many potential patent reader (e.g. designers, marketers) could be taken aside. Despite this problem, some researchers [3] affirm that there is an increasing variety of readers: not only technician and researchers but also marketers and designers who have grown an interest in patent analysis. Nevertheless, to our knowledge there are no researches that aim at facilitating information extraction for non-technological focused patent readers.

The bias that patent are only tech-oriented documents is due to two main reasons:

- Patents are produced to disclose and protect an invention, their content is mainly technical and legal.
- 80% of technical information is not available elsewhere [4,5], so patents are one of the most comprehensive resources for technical analysis.

Focusing on the second point, our hypothesis is that also a fraction of all the other kinds of information (e.g. marketing and sociological information) is not contained elsewhere and it will appear in public documents (e.g. manual handbooks and market reports) in 6–18 months [6].

Unfortunately there are four aspects reducing the non-tech readers' ability to analyze patents efficiently. First of all, an increasingly high number of patent filings generates a massive information overflow [7]; secondly, analyzing patents takes a long time and requires skilled personnel [8]; the quality of patent assessment process is decreasing [9,10] because of the reduced assessment time available for patent examiners; finally, activities like patent hiding, proliferation and bombing, contribute to the generation of confusion and to the loss of time in research and analysis phases [11]. These problems affect non-tech oriented patent readers as well as typical readers, even though the impact may be stronger on the firsts.

The main difference between typical and non-tech patent readers is the information they focus on. *Patent attorneys* and *Intellectual Property (IP) managers* are interested in reading patents for legal reasons to orient the IP direction. Analyzing patents is the core of their work, so they are experts in finding the information they need. Furthermore,

[*] Corresponding author.
E-mail address: filippo.chiarello@destec.unipi.it (F. Chiarello).

https://doi.org/10.1016/j.wpi.2018.07.006
Received 13 February 2018; Received in revised form 22 May 2018; Accepted 18 July 2018

they can spend most of their work-time on the activity. On the other hand, usually *marketers and designers* (taken as example of non-tech oriented readers) search users' behavioral changes and needs, market trends, designers' vision, R&D trends and competitors' strategies. In addition, they rarely work with patents, so they do not know what and how to search. Lastly, they have short time to spend on the activity, and they waste most of this time understanding the legal and technical jargon used in patents.

In this paper we present a *system for entity extraction from patents*, aimed at putting the focus on valuable non-technical information, making patent content more accessible and usable by non-technical readers like marketers and designers. Marketing and design business functions are always searching for new users and market niches where to experiment new products or to capture new trends, therefore an automatic system able to scan and explore a huge patent set could be useful. Moreover the extracted information could be used as keywords for a standard search in Facebook or Twitter in order to crosscheck such evidences coming from patents.

Entity extraction means to find textual elements which are unique identifiers of entities belonging to predefined classes [12]. To build a suitable system for this task we can not leave aside two aspects:

1. *effectiveness*: selection of useful entities contained in patents, in particular for non-technical readers.
2. *efficiency*: deep knowledge of state of the art techniques for data analysis required for the extraction of these entities from large corpora of documents.

In the present work we focus on the extraction of one of the possible entities hidden in patents: the users of the invention. The concept of user is of great interest to many disciplines, in particular to marketing and design.

Recent Natural Language Processing (NLP) tools and techniques have been proved to be adequate for the extraction of specific information from texts [13]. Unfortunately, NLP tools suffer from a dramatic drop of accuracy when tested on domain specific texts [14] such as patents. Consequently, tools must be adapted to the patent domain [15].

The paper is then organized as follows. In section 2 we present the concept of *user of the invention* and discuss why we want to extract this entity. In the same section we also provide some examples of the state-of-the-art patent text analysis systems. Section 3 presents the process we used to solve the problem. In Section 4 the process is applied to a case study and the outcomes are discussed. Finally, section 5 concludes with an overview of future developments.

2. Users: a key information hidden in patents

In this section we explains the concept of *user of the invention*. Section 2.1 gives a definition of users and presents the way that this concept is exploited in different knowledge fields. Section 2.2 briefly discusses the state of the art of automatic processing of patents, focusing on entity extraction systems.

2.1. Users: definition and usage

Patents are documents that must provide a detailed public disclosure of an invention [16]. An *invention* is a new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof.[1]

The notion of usefulness implies that the invention must have some value and not necessarily for a human entity. In fact, patents usually describe processes, machines or composition of matter which are useful for another process, machine or composition of matter.

Therefore, we distinguish between stakeholders and users, considering the definitions given by the authors in Ref. [17].

Definition 1. Stakeholder: *Stakeholders are entities on which the invention has or will have a positive or negative effect in order to show usefulness.*

This definition covers all possible entities that engage an active or passive relation with the invention. Given the logical condition of usefulness of patents, all patents must have stakeholder information. If a patent has not got any stakeholder information in it the patent application should be rejected.

Definition 2. User: *Users are animated or previously animated entities (human or animal, alive or dead), on which the invention has a positive or negative effect at an unspecified moment.*

Given definition 2, it is clear that every user is a stakeholder while non-users stakeholders include artifacts, machines, manufacturing or operational processes.

Corollary 1. *Multiple roles*: *Identities may have multiple roles as users.*

Our idea of users describes roles, not identities. Animated entities have an identity, as it happens for a specific person. A person has many roles as a user. For example, a working mother starts her day taking on the role of a mom, in which she is expected to feed her children and get them ready for school. At the office she shifts to the role of project manager, so she oversees projects in a timely and professional manner. Working mother, mom or project manager can be considered user roles attributed to the same person, or identity. From this definition it is clear that users are close to what social sciences define as social roles (a focus on the way that social sciences interpret social roles is shown in section 2.1.1).

Afterward, we can outline knowledge fields using the concept of user. Sections 2.1.1, 2.1.2, 2.1.3 and 2.1.4 have a twofold aim: help the reader to understand how the concept of users is interpreted in different knowledge fields; explain the background of the methodology described in section 3.1).

2.1.1. Social sciences: social roles as users

In social sciences *social roles* are comparable with our definition of user. As defined in the psychological dictionary Alleydog,[2] ”*social roles refer to the expectations, responsibilities and behaviors we adopt in certain situations.*”. The example of the working mother shown in section 2.1 is the case of social roles.

The field of social sciences is the only one in which an attempt of automatic extraction of users has been done. In Ref. [18] the authors extracted social roles from Twitter using heuristic methods. The authors looked for all the words preceded by constructions like ”I'm a” and similar variations. This search resulted in 63.858 unique roles identified, 44.260 of which appeared only once. The result of the extraction process is noisy and only a low percentage of the extracted words are social roles. Despite of this noisy extraction, some entities are consistent with our definition of user, e.g. *doctor, teacher*, (*mother*) or *christian*.

Another work [19] tries to identify social roles on Twitter exploiting a set of assumptions. The authors take into account roles, each one with a set of related verbs: if someone uses verbs from a set, that person may cover that particular social role. To sanitize the collection of positively identified users, the authors crowd-sourced a manual verification procedure, using the Mechanical Turk platform.[3] Also here some interesting extractions are performed, obtaining users like *artist, athlete, blogger, cheerleader, christian, DJ*, or *filmmaker*. These two works differ from the present study for what concerns the analyzed texts and the

[1] http://www.uspto.gov.

[2] http://alleydog.com.

[3] https://www.mturk.com/mturk/.