Technological Forecasting & Social Change 115 (2017) 210–220

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

Use of a new patent text-mining and visualization method for identifying patenting patterns over time: Concept, method and test application

Helen Niemann, Martin G. Moehrle *, Jonas Frischkorn

University of Bremen, IPMI – Institute of Project Management and Innovation, P.O. Box 330 440, D-28344 Bremen, Germany

ARTICLE INFO

Article history:
Received 17 March 2015
Received in revised form 29 June 2016
Accepted 3 October 2016
Available online 26 October 2016

Keywords:
Patent lanes
Patent analysis
Cluster analysis
Similarity measurement
Visualization
Citation networks
Technological trajectories

ABSTRACT

Understanding the evolution of a technological field in the course of time is a key task in technology analysis. Analysts in research institutions as well as in companies need to know which topics are relevant for the respective technological field, which are the emerging topics, which traditional topics have been deepened in the course of time and which have been abandoned. For this purpose we suggest a patent lane analysis. Patent lanes can be seen as the deployment of patent clusters in the course of time. We use a method based on semantic similarities to develop patent lanes. A case study focuses on the application of carbon fibers in bicycle technology; it is used to demonstrate our method, i.e. to establish patent lanes in this case and characterize them by multiple use of a Tf idf measure. Despite some limitations, patent lanes enable deep insights into the development of patent-friendly technological fields.

1. Introduction

A lot of corporate technology managers and scientists in research and political institutions seek to understand typical evolutionary patterns of technological fields. For instance, they might wish to know which topics are relevant for the respective technological field, which are the emerging topics, which traditional topics have been deepened in the course of time and which have been abandoned? With respect to many, though by no means all, technological fields, patent analyses may help to answer such questions. They have been used successfully and extensively in many cases, and comprise different techniques, such as co-classification analysis (see as examples Choi et al., 2007; Chang et al., 2009; Dereli and Durmusoglu, 2009) or citation analysis (Tseng et al., 2011; Frietsch, 2007; Kuusi and Meyer, 2007; de Souza Carvalho et al., 2009; Lee et al., 2009, 2012).

A multitude of techniques for patent analysis makes use of the so-called meta-data of patents. Meta-data are defined by patent laws like the U.S. Code Title 35 and comprise information on applicants, inventors, classifications (international patent classification [IPC], current patent classification [CPC], in some cases national classifications like the US patent classification [USPC]), application and granting dates, cited patents and other literature (Ernst, 2003; Lee et al., 2011). Valuable answers to some of the questions mentioned above can be obtained by such analyses; techniques like activity analysis, co-classification activity analysis and citation network analysis may provide answers regarding different technical aspects and the development of topics in the technological field over time. The answers to some questions are not quite perfect yet, and there is still potential for improvement. Especially exploiting the information contained in the full-text of patents (instead or in addition to meta-data) by means of text mining technologies, as suggested by Yoon and Kim (2011) as well as by Moehrle and Gerken (2012) and Gerken and Moehrle (2012), may provide researchers with deeper insights. Text mining offers the opportunity to establish semantic similarity measures between documents and in doing so provides an alternative or an addition to the well known citation analysis.

In this paper we concentrate on these text mining technologies and suggest so-called patent lanes which we define as the deployment of patent clusters over time. The idea of patent lanes is related to the timeline visualization of the development of technological clusters (Shibata et al., 2010), but uses disaggregated information instead of clusters. The paper is organized as follows: In the next section we briefly explain how to measure semantic similarities between patents, as this constitutes the foundation of our method. In order to underpin our methodical contribution, we compare semantic similarities and citations as basic elements that establish links between patents, and show their interrelation in analyses. As patent lanes may be configured in different ways, we discuss the most important design decisions. A case study which focuses on carbon fiber reinforcements and the utilization thereof in bicycle technology, serves to illustrate the use of patent lanes and the

* Corresponding author.
E-mail addresses: helen.niemann@innovation.uni-bremen.de (H. Niemann), martin.moehrle@innovation.uni-bremen (M.G. Moehrle), jonas.frischkorn@innovation.uni-bremen (J. Frischkorn).

http://dx.doi.org/10.1016/j.techfore.2016.10.004
0040-1625/

interpretation of results. We compare our method with methods characterized by rolling clustering to identify criteria for the usefulness of its application. Some concluding remarks will highlight implications as well as limitations of our method.

2. Measurement of semantic similarities between patents

One basic feature of our method is the application of semantic similarities between patents (see Moehrle, 2010). There is one major idea behind this: We assume that similarities between the contents of patents are reflected by similarities in language, for instance by the use of similar terminology (e.g. specifically scientific terminology), explanations of similar application situations, or a focus on similar useful functions. There seems to be some evidence to support this assumption, as in a recent study Möller and Moehrle (2015) have shown that this type of background information may significantly supplement and improve traditional keyword-based patent searches.

In the available literature several methods of measuring semantic similarities can be found. Having generated a basic set of patents representing the technology under investigation by means of keyword or classification-based search, the related tasks may be summarized as a generic process in four steps (see Moehrle and Gerken, 2012, see Fig. 1), comprising (i) preliminary language processing, (ii) concept extraction and building, (iii) variable measurement, and (iv) similarity calculation.

Before semantic measurements can take place, the data should be cleaned, i.e. terms should be reduced to their word-stems, synonyms should be harmonized, and filters for non-discriminant terms should be applied.

There are different ways to extract and build concepts (in the sense of key terms) from patent documents. Yoon and Kim (2011) use subject-action-object structures (SAOs) for this purpose and make use of knowledge about the syntactical functions of the extracted concepts. Moehrle and Gerken (2012) apply n-grams to generate solitary and combined concepts and give advice on how to configure the extraction. In this paper we concentrate on the latter option.

After extracting semantic concepts one way or the other, different variables can be measured. Such variables may represent the size of a patent, the overall overlapping set, or an overlapping set measured from the perspective of a pair of patents (double single-sided, abbreviated DSS, see Moehrle, 2010).

Based on the established variables established, semantic similarities can be calculated. "Similarity is formally defined as an increasing function of commonality and decreasing function of differences among objects to be compared" (Jeong et al., 2008). Different formulas are available for this purpose (see Gower and Legendre, 1986), for instance the Jaccard index that relates an overlapping set of terms to the sum of patent related sets of terms, or the Inclusion index that relates an overlapping set of terms to the patent related set of terms of the smaller patent.

3. Similarities and differences in connections between pairs of patents based on citations and semantic similarities

Having introduced semantic similarities, we now compare semantic similarities and citation analysis to underpin our methodical contribution. For this purpose we focus on the characteristics of semantic similarities and citations; later on we show the use thereof in different analyses.

Patent citations are generally differentiated into forward and backward citations. "Forward citations are the number of citations received by a patent. Counting the forward citations of patents shows whether patented inventions are mentioned – either by examiners … or by applicants or their lawyers" (Rost, 2011). In contrast, "backward citations are made by a patent to a previously issued patent. Studies using backward citation information investigated spillovers … between technology classes … or regions" (Rost, 2011). Patent citations have been used since the 1990s for establishing the importance of patents (see the work by Jaffe et al., 1993) and more recently for analyzing knowledge flows based on complexity theory (see Sorenson et al., 1993).

Basically, both citation based and semantic similarity based approaches connect pairs of patents. In the following we will first focus on the connection between a pair of patents, and then briefly discuss superior network structures.

There are five major differences in the connections between pairs of patents based on citations or on semantic similarities. They differ (i) in the range of values of the connection, (ii) the establishment of the connection, (iii) the timely availability of the connection, (iv) the foundation of the connection, and (v) the localization of the connection in a patent's parts (see Table 1). Compared to citations, semantic similarities comprise a continuous range of values, they are caused by lawyers and inventors who formulate the patent's wording (which leads to a fuzziness of the approach that has to be taken into account), they are completely available on the issue date of a patent, the connection elements can be identified (as the set of shared terms), and the connection elements can be located in the parts of a patent such as the claims

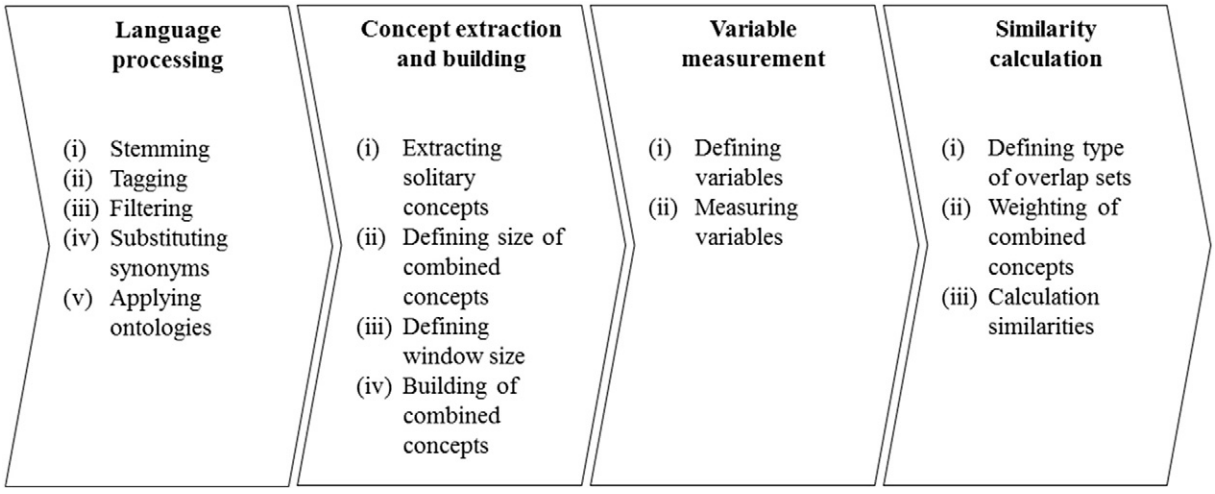

Fig. 1. Generic process for semantic similarity calculation.
Source: Moehrle and Gerken (2012), p. 807.