Technological Forecasting & Social Change 94 (2015) 236–250

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

Topic based classification and pattern identification in patents

Subhashini Venugopalan [a,*], Varun Rai [b,c]

[a] Department of Computer Science, The University of Texas at Austin, Austin, USA
[b] LBJ School of Public Affairs, The University of Texas at Austin, Austin, USA
[c] Department of Mechanical Engineering, The University of Texas at Austin, Austin, USA

ARTICLE INFO

Article history:
Received 15 January 2014
Received in revised form 26 May 2014
Accepted 11 October 2014
Available online 7 November 2014

Keywords:
Patents
Topic modeling
Document classification
Technology convergence
Solar photovoltaic
Knowledge flows

ABSTRACT

Patent classification systems and citation networks are used extensively in innovation studies. However, non-unique mapping of classification codes onto specific products/markets and the difficulties in accurately capturing knowledge flows based just on citation linkages present limitations to these conventional patent analysis approaches. We present a natural language processing based hierarchical technique that enables the automatic identification and classification of patent datasets into technology areas and sub-areas. The key novelty of our technique is to use topic modeling to map patents to probability distributions over real world categories/topics. Accuracy and usefulness of our technique are tested on a dataset of 10,201 patents in solar photovoltaics filed in the United States Patent and Trademark Office (USPTO) between 2002 and 2013. We show that linguistic features from topic models can be used to effectively identify the main technology area that a patent's invention applies to. Our computational experiments support the view that the topic distribution of a patent offers a reduced-form representation of the knowledge content in a patent. Accordingly, we suggest that this hidden thematic structure in patents can be useful in studies of the policy–innovation–geography nexus. To that end, we also demonstrate an application of our technique for identifying patterns in technological convergence.

1. Introduction

The primary goal of this paper is to develop methods that exploit language similarities for clustering and classifying patents, potentially allowing for the discovery of knowledge-related temporal and spatial connections between groups of patents. Meaningful quantification of such 'connections' can be useful in the identification of patterns of inventive activity in specific technologies and markets, how those patterns evolve, and what factors drive that evolutionary process. The methods we present here provide a promising alternative approach for studying technology evolution and the role of markets therein as compared to the dominant approach based on the use of patent classification system(s), which has several limitations. In particular, our methods can be useful in studying questions in the economics and geography of innovation, such as knowledge spillovers and the impact of demand-pull policies on innovation.

1.1. Limitations of patent classification systems

Two examples will help clarify the above point. *First*, while it is known that demand-pull (market creation) policies are important drivers of innovation (Schmookler, 1966; Nemet, 2009; Popp et al., 2010; Weyant, 2011), the impact of such policies on the location of innovative activity (domestic vs. foreign) is not clear, and there are differing views on this issue in the literature (Peters et al., 2012). Especially, in the context of low-carbon technologies in recent years there has been an apparent 'race to the top': several developed and developing countries have provided strong policy support for such technologies in order to spur innovation and economic activity led by *domestic* firms. But if domestic policies incentivize

* Corresponding author.
E-mail addresses: vsub@cs.utexas.edu (S. Venugopalan), raivarun@utexas.edu (V. Rai).

http://dx.doi.org/10.1016/j.techfore.2014.10.006
0040-1625/

significant levels of innovation by foreign firms as well, then the competitiveness argument in favor of such policies may be weakened. Yet, the empirical literature in addressing the impact of demand-pull policies on domestic vs. foreign innovation is relatively thin (Peters et al., 2012; Popp, 2006; Dechezleprêtre and Glachant, 2012; Lanjouw and Mody, 1996).

To take the example of low-carbon technologies further, the industry/market-specific (i.e., targeted) nature of several key policies – such as the production tax credit (PTC) for wind generation in the U.S. and the California Solar Initiative (CSI) – necessitates focusing on particular markets in order to make the policy–innovation–location link. But the common approach employed by many researchers for studying inventive and innovative activity through analyzing patent data tagged into different technology categories and sub-categories through 'classification codes' is difficult for market-oriented studies for two reasons. First, because patent classification is somewhat subjective due to variations in judgements of the patent examiners and due to possible strategic patenting by firms, data based on classification codes may be both inaccurate and noisy (Allison and Lemley, 2002; Dahlin and Behrens, 2005; Nemet, 2012). Second, and perhaps more importantly, patent selection based on classification codes is not conducive for studying inventive activity in specific product or market areas, as the classifications are oriented towards the relevant technical stream of the invention, not the product or market area (Jaffe, 1986; Altwies and Nemet, 2013; Nemet, 2012): "… classification system … is *not* an alternative product or industry classification … the mapping from classes to industries is not unique in either direction" (Jaffe, 1986). To circumvent some of these limitations associated with using patent classification codes, some recent studies have used a combination of classification codes and keywords to study inventive activity in specific technologies and markets (Popp and Newell, 2009; De La Tour et al., 2011; Altwies and Nemet, 2013).

Second, patents are also an appealing way for studying knowledge spillovers, an important mechanism in the process of innovation whereby knowledge generated in one domain/industry positively impacts knowledge creation in other domains/industries (Rosenberg, 1994; Mowery and Rosenberg, 1999; Arthur, 2007). While the concept of knowledge spillover is intuitively easy to understand, there are significant challenges in efforts to operationalize and quantitatively measure and track this concept (Dahlin and Behrens, 2005; Choi et al., 2007; Thorleuchter et al., 2010; Lee et al., 2011; Nemet, 2012; Roach and Cohen, 2013). Traditionally, knowledge spillovers in patents are studied using citation linkages (Harhoff et al., 1999; Jaffe et al., 2000; Nagaoka et al., 2010; Nemet, 2012; Érdi et al., 2013). However, this approach is fraught with difficulties. For example, Alcacer et al. (2009) show that citation patterns vary significantly by firm, industry, and even country characteristics. They also make the insightful observation that, "… a substantial number of firms won patents without listing a single applicant citation". Hegde and Sampat (2009) speculatively suggest that firms' citation choices are likely strategic — firms may forgo citing prior art if those citations might potentially block the claims in their patent, but preferentially cite patents that support their own claims. In a recent study, using matched patent and survey-based data Roach and Cohen (2013) show that citations systematically under-represent knowledge flows from public research – "errors of omission" – as well as over-represent factors not representative of knowledge flows — "errors of commission". Overall, the scale and scope of citations contained in patents depend heavily on firm-level practices, rendering citations a tricky tool for studying intra- and inter-industry knowledge flows. Further, when analyzing knowledge flows using patent citations it is common to assume that every cited patent contributes an equal "amount" of knowledge (for example, see Nemet, 2012), which is a potentially troublesome assumption (we discuss this further in Section 6.3). More accurate measurements of knowledge spillover from patent citations need to go beyond the assumption that all backward citations are equally important.

1.2. Contributions of this paper

In view of the above inherent limitations posed by patent classification systems, exploitation of the textual content of patents to develop new techniques for classifying patent sets has been suggested (Cascini et al., 2004; Bonino et al., 2010). In this paper we propose a hierarchical technique based on natural language processing (NLP) for classifying patents and for identifying linkages between them.

Our hypothesis is that textual description of patents and linguistic patterns are powerful indicators of the knowledge content that is not necessarily apparent in the context of citations. In this work, we develop a two-stage, hierarchical technique to exploit this insight. In the first stage, we use document term frequency to identify patents relevant to a particular set of technologies – solar photovoltaics (PV) – and then in the second stage we use topic modeling (Lafferty and Blei, n.d.; Blei and Lafferty, n.d.) on the relevant patents to capture similarities based on linguistic patterns. Our two-stage method begins with the collection of abstract and claim text of issued patents and patent applications that are returned from a simple keyword-based search of the technology area of interest (in this case, PV) in the U.S. Patent and Trademark Office (USPTO) database. Prior to automating the classification into relevant and irrelevant patents (Stage 1), we annotate the dataset by categorizing patents into specific technology areas with the help of experts. We use 70% of the annotated data (training dataset) to learn our classifiers and hold out 30% for final evaluation purposes. Using the training dataset, in the first stage, we build a supervised classifier based on document term frequency to filter out the irrelevant patents. In the second stage we run a topic modeling algorithm to identify key topics in the patents. This step generates a probability distribution of the patents over the set of topics (see Section 4.1 for more details). Using the topic distribution as features, we build classifiers that learn to categorize the patents into their technology sub-areas (Stage 2). We then compare and verify the effectiveness of our approach on the held-out annotated data. We finally augment the classification results with the published dates (years) of the patents (issued or applications) to observe changes and patterns in time. Based on our computational experiments we show that:

- Our classifier based on document term frequency can separate relevant and irrelevant patents with an accuracy of 98.2%. This classifier correctly retrieves (recall) 88.7% of the relevant patents.