Technological Forecasting & Social Change 115 (2017) 131–142

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

Firms' knowledge profiles: Mapping patent data with unsupervised learning

Arho Suominen [a,b,*], Hannes Toivanen [a,b,c], Marko Seppänen [d]

[a] VTT Technical Research Centre of Finland, PL 1000, Espoo, Finland
[b] Teqmine Analytics Ltd, Pasilanraitio 5, 00240 Helsinki, Finland
[c] Lappeenranta University of Technology, School of Business, Lappeenranta, Finland
[d] Tampere University of Technology, Department of Pori/Industrial Management, PL 300, 28101 Pori, Finland

ARTICLE INFO

Article history:
Received 17 February 2016
Received in revised form 28 September 2016
Accepted 28 September 2016
Available online 6 October 2016

Keywords:
Technology management
Patent analysis
Unsupervised learning
Topic modelling
Telecommunication industry

ABSTRACT

Patent data has been an obvious choice for analysis leading to strategic technology intelligence, yet, the recent proliferation of machine learning text analysis methods is changing the status of traditional patent data analysis methods and approaches. This article discusses the benefits and constraints of machine learning approaches in industry level patent analysis, and to this end offers a demonstration of unsupervised learning based analysis of the leading telecommunication firms between 2001 and 2014 based on about 160,000 USPTO full-text patents. Data were classified using full-text descriptions with Latent Dirichlet Allocation, and latent patterns emerging through the unsupervised learning process were modelled by company and year to create an overall view of patenting within the industry, and to forecast future trends. Our results demonstrate company-specific differences in their knowledge profiles, as well as show the evolution of the knowledge profiles of industry leaders from hardware to software focussed technology strategies. The results cast also light on the dynamics of emerging and declining knowledge areas in the telecommunication industry. Our results prompt a consideration of the current status of established approaches to patent landscaping, such as key-word or technology classifications and other approaches relying on semantic labelling, in the context of novel machine learning approaches. Finally, we discuss implications for policy makers, and, in particular, for strategic management in firms.

1. Introduction

Operationalising a company's knowledge base in terms of its depth and breadth and creating trajectories to the future is challenging (Zhang and Baden-Fuller, 2010). The intensified complexity of emerging technologies (Breschi et al., 2003; Garcia-Vega, 2006) requires improved understanding of the nature and effect of cross-disciplinary activities in innovation processes (Wang and von Tunzelmann, 2000). Increasingly, companies must rely on broad knowledge bases covering diverse technology areas, while simultaneously having significant depth in their core competence. This creates a new type of tension for the management of technology and innovation. This is particularly problematic in highly dynamic industries. We examine the effects and potential of big data approaches in managing this increased complexity of company knowledge bases with a study on the telecommunication industry, and develop perspectives to exploit big data foresight approaches in support of strategic planning.

Previous studies on the depth and breadth of knowledge and technological trajectories have used patent information. As Moorthy and Polley (2010) point out, patents are the most feasible approach for analysing the breadth and depth of knowledge within a company as the data provides an insight to its competences. The simplest approach to quantifying the knowledge base is to use the patent classification scheme provided in the patent archive as a basis for evaluation – breadth correlating with the diversity in patent classifications and depth with the concentration of patent classifications in a company patent portfolio. This approach was used for example by Zhang and Baden-Fuller (2010) to analyse technology collaboration. Moorthy and Polley (2010) and SubbaNarasimha et al. (2003) use the approach to analyse the impact of breadth and depth of knowledge to company performance. Wu and Shanley (2009) operationalise the role of exploration in company knowledge stock by means of patent metrics.

Analysing classification metadata, in addition to citations, can be regarded as the de facto standard of utilizing patent metrics (e.g. Huang et al. (2015) as a case in point). This approach in analysing breadth and depth is not without limitations. Connecting patent classifications directly to industry sectors poses a challenge (Schmoch, 2008). Different patent classification systems have struggled to establish a tool to clearly distinguish industries into specific classes, limiting the

* Corresponding author at: VTT Technical Research Centre of Finland, PL 1000, Espoo, Finland.
E-mail addresses: arho.suominen@vtt.fi (A. Suominen), hannes.toivanen@teqmine.com (H. Toivanen), marko.seppanen@tut.fi (M. Seppänen).

http://dx.doi.org/10.1016/j.techfore.2016.09.028

applicability of classifications for sectoral analysis. Classifications are also of limited value in directing inventive effort (Loh et al., 2006), which is understandable due to the information retrieval nature of patent classifications. Patent classifications are a tool for the patent process, and the human process related to assigning classes is valuable in the intellectual process, even to the extent that automated classifications fall short of providing similar results (Richter and MacFarlane, 2005).

The subjectiveness of the classification process of patents remains a major limitation for the usefulness patent data (Venugopalan and Rai, 2015), making it an inadequate measure to satisfy the needs of corporate planning (Archibugi and Planta, 1996; Lai and Wu, 2005). Nakamura et al. (2015) review the managerial challenges of analysing patent data, pointing to the need for frequent updates (Herrero et al., 2010) and cost of data collection (Nakamura et al., 2015, Kajikawa et al., 2006) with a limited success rate in producing practical results. They conclude, through expert interviews, that even though patent data is a relevant decision making tool for practitioners its usefulness is hindered by the significant limitations embedded in the patent classification based metric. From this perspective, machine learning provides a valuable approach for strategic foresight and technology management (see e.g. Ventura et al., 2015). Machine learning opens the possibility for cost effective analysis of full text patent data, which can mitigate the limitations of de facto standard metadata based approaches.

By employing big data approaches to manage technology intelligence, companies can foster new forms of adaptive learning in innovation and strategy. Such approaches require the augmentation of human judgement in the categorisation and analysis of knowledge with machine learning methods, prompting serious challenges to the existing corporate foresight traditions. Leveraging these efforts within companies requires their systematic integration to existing strategic foresight processes. Using an automated continuous monitoring based on topic modelling with Latent Dirichlet Allocation and network analysis, we will show how a semantic analysis leads to the identification of opportunities for learning and innovation in complex environments. From a total dataset of 157,718 full-text telecommunication patents from USPTO, we have monitored and detected changes in the knowledge patterns of companies, e.g. how semantic analysis shows the change from a hardware-focused knowledge domain in telecommunication towards software-dominated knowledge foci. In this paper, we explore the latent knowledge dimensions of patents in global telecommunication companies, focusing on two questions: 1) Can we identify topical knowledge foci of different companies with unsupervised learning, and if so, 2) What are the dynamics of knowledge domains among the companies?

2. Background

Informetric analysis focuses on operationalising developments in the science and technology system. Informetrics can focus on a science, technologies or companies creating insight on the historical developments and forecast future trajectories. At a company level, Porter and Newman (2011) write about competitive technical information (CTI), the information companies need to survive in the dynamic marketplace. Suominen (2013) reviews the established metrics used to create quantitative insights, highlighting that metrics used to profile developments need to be objective and reproducible, while responding to Ayres (1989) call for accurate decision-making tools.

Much of the current informetrics analyses have focused on the metadata level (cf. Suominen, 2013) creating measures of activity, linkage or impact (Moed et al., 1995). Text analytics have most commonly been limited to keyword or abstract analysis. New open datasets and increases in computational efficiency have made full-text analysis possible (for example Glenisson et al., 2005). Tseng et al. (2007) have reviewed text mining techniques for patent analysis highlighting different methodological options and steps, such as text segmentation, summary extraction, feature selection, term association, cluster generation, topic identification, and information mapping. Tseng et al. (2007) describe the approaches to filter irrelevant content and retrieving the core features of the patent. Kang et al. (2007) have reviewed different clustering approached to summarizing patents, although much of this work is based on utilizing the international patent classification systems. Kim and Choi (2007) on the other hand analyse patents using the semantic structure of the patent as a starting point. Patent text mining studies often either rely on filtering text based on established knowledge on the structure of patent text or show trends of classification or aggregated technology areas.

2.1. Depth and breadth of knowledge

The increased complexity of technologies has changed the dynamics of innovation in that there is an increased need for cross-disciplinary activities (Wang and von Tunzelmann, 2000; Subramaniam and Youndt, 2005). Studies have shown that technologically diverse knowledge systems are a dominant feature in companies, as multiple fields of knowledge are integrated in the innovation process (Mendonça, 2006). To analyse change in knowledge resources we are forced to understand the multi-dimensional knowledge base of an industry (Kauffman et al., 2000).

Knowledge depth can be defined as the level of expertise within a confined technological area (George et al., 2008; Zhang et al., 2007) described by Wang and von Tunzelmann (2000) as "analytical sophistication". In contrast, breadth of knowledge refers to the number of adjacent technologies in the relevant multi-dimensional knowledge space of a company (Wang and von Tunzelmann, 2000; Zhang et al., 2007). Moorthy and Polley (2010) showed that rather than the stock of knowledge, the breadth and depth of knowledge in fact represent more important variables to explain a firm's performance. Companies are required to have a minimum depth of knowledge in a specific area and breadth enables them to cope with rapid technological change. An evaluation of these two variables at a company level allows us to follow strategic trajectories.

The breadth and depth of knowledge in companies is in parts visible outside the company in codified information such as patents, where patent classifications provide a tool for analyses. Codification has enabled easy access to analysing the knowledge structure through a posteriori labels given to new information. With patents, this metadata is in fields such as application data, patent classification, and assignee, which codify the actual information to make it more accessible.

Patent classifications have remained as the most practical approach in understanding the structure of the information. There are, however, significant caveats to this approach. Patent classifications are subjective in nature, prone to classifications errors and overall noisiness (Dahlin and Behrens, 2005; Nemet, 2009). The classifications are by definition an information retrieval system, which scholars and practitioners use as a proxy metric for example analysing the breadth and depth of knowledge. At the same time we are acutely aware of several significant limitations in the proxy we are using. We know that the implicit notions and underlying taxonomy of patents are often misunderstood (McNamee, 2013). There are clear challenges to link patent classifications to either industry (Schmoch, 2008) or market sectors (Jaffe, 1986). Classifications are also of limited value in directing inventive effort (Loh et al., 2006). Using a priori determined classification, new topics pose a challenge. Classification based on historical knowledge lacks the ability to adapt to new knowledge (for discussion on approaches, see van Merkerk and van Lente, 2005; Kuusi and Meyer, 2007).

There is a clear need for a more adaptive approach to analysing patent data, suggesting that automated classification drawn from the actual text could be a better approach for showing the actual breadth and depth of the knowledge base.

2.2. Unsupervised learning and topic modelling in patent data

Unsupervised learning produces an outcome based on an input while not receiving any feedback from the environment. As an automated