Bibliometrics, Time Series Analysis and SAO-based Approach for Technology Roadmapping of Emerging Technologies

Introduction

This work proposes an approach that combines a set of quantitative methods to generate technology roadmaps, which draw on Science, Technology & Innovation data. The roadmap is designed to show the whole evolution of the core concepts of the technology, from initial researches to current issues which are being approached by scholars and practitioners. Additionally, short-term forecasting information is also integrated, based on the observed evolution. The approach is designed to be applied to emerging technologies and its outcomes can be considered as inputs for competitive technical intelligence activities. It facilitates the identification of the fields that have had more importance in the development of the technology, as well as those that may have it in the future. The approach was applied to the combined field of additive manufacturing in aeronautics; the results allow to obtain a clear idea of which are the main subfields of work and research of the technology, both the already developed ones as well as those that are currently on the cutting edge.

Methodology

The approach comprises five integrated methods within the tech mining field. Each of the methods is used aiming the generation of valuable information, which is aggregated to depict the technology comprehensively. Firstly bibliometrics is applied for the retrieving of scientific publications and patents, which are directly related to the analyzed technology. These data constitute the raw data to be mined. Afterwards, text mining methods are used, in terms of term-clumping and subject-action-object (SAO) analysis, for topical analysis. The whole analysis is based on the study of the text of the publications, specifically on the analysis of the titles, keywords, abstracts and claims. Term clumping allows a consistent grouping of the most important terms, and combined with principal component analysis (PCA) it is used to identify the main subfields of the technology. These subfields are constituted by those terms that present a high correlation in terms of appearance in the publications. On the other hand, SAO semantic analysis is a fact-oriented modeling technique based upon the theory of Inventive Problem Solving (Russian acronym: TRIZ) (Wang et al. 2017). S and O denote the components, and A denotes the effect or relationship between components in the invention. Once the topical analysis is completed, the approach integrates the hierarchical clustering method to identify the structure of the technology. By means of this step, both the vertical structure of the technology and its main areas are identified.

Time series (TS) analysis is used to get a quantitative measure of the evolution and forecast of the technology. In order to do so, the subfields that have been obtained from the term-clumping are characterized through TSs of monthly frequency. These series are generated by means of a novel method, which is based on the frequency of occurrence, within the publications, of the terms contained in each subfield. The TSs are then modelled with Structural Time Series Models (STSMs), presented by Harvey (1989). These models belong to the class of Unobserved Components models, which decompose the series into components such as trends, cycles and seasonalities. The advantage of STSMs is that each component is specified stochastically so they can change over time following changes in the data. Once the model is validated, it can be used to estimate and project the trend component, which is used as a proxy for the behavior of each subfield over the period considered. The value of the trends over the years analyzed provide a good measure of the intensity of the activity related to each subfield, and their projection provides a good vision of the potential future evolution. Even though any forecasting attempt should be analyzed with caution, the utility of the present approach lies in considering a credible proxy for the trend and in projecting it on a statistically sound basis.

Finally, all the generated information is integrated into a single visual element, the technology roadmap (TRM). Regarding the structure of the TRM, it has two main layers: the technology layer that is generated based on the information contained in the scientific publications (articles); and the application layer that is based on the information contained in the patents. These layers, in turn, are divided into sublayers, which directly correspond with those main areas obtained in the clustering task. As regards its contents, a comprehensive depiction of the technology is obtained by means of the graphic representation of the main subfields, the terms and SAO elements contained within them and the projection of all the elements into the future.

Data, results and discussion

The approach was applied to the additive manufacturing in aeronautics field. The data was retrieved from Patseer (1892 patents), WoS (891 articles), and Scopus (2337) databases, and the time range was 1997-2016. The term-clumping process was done for patents and articles separately by means of ClusterSuite software (O'Brien et al. 2013), integrated in VantagePoint tech mining tool. The outcome was a set of 1097 and 340 core terms for patents and articles respectively. Principal component analysis (PCA) was then applied, with which 33 and 25 factors were identified respectively. As mentioned, these factors were considered the main subfields of the technology. Examples of these subfields are silicon rubber or superphobic surface for patents, and blisk or deterioration for articles. Once the main subfields were identified, the hierarchical clustering was carried out. The selected clustering algorithm was agnes (Kaufman and Rousseeuw 2009), with ward clustering method. The resulting structure provided the main areas of the technology: material, cladding, additive manufacturing process, resistance properties, deposition, aircraft parts and fuel for application layer; alloy & composites, polymers, products and processes for technology layer. In parallel, the data were mined using an NLP tool named ReVerb (Fader et al. 2011), for the identification of SAO components. Thus, in the initial years of the scientific publications there can be encountered SAOs of the type of 'the prototypes were manufactured using stereolithography'. On the other hand, the last patents contain elements like 'the textiles are selected from polyester-polyurethane copolymers', and similar ones. Besides, the SAOs were linked with their corresponding subfield, observing for that to which factor belonged the S or the O of each one of them.

In order analyze the evolution of the subfields, individual TSs were generated and modelled for all of them. The resulting model for all the series is known as smooth trend model, which is formulated as follows: $\log(Y_t) = \mu_t + \varepsilon_t, \varepsilon_t \sim NID(0, \sigma_\varepsilon^2)$. Where the trend can be divided into two elements, the level μ_t, and the slope, β_t, by $\mu_t = \mu_{t-1} + \beta_{t-1}$ and $\beta_t = \beta_{t-1} + \zeta_t, \zeta_t \sim NID\left(0, \sigma_\zeta^2\right)$. Based on the information provided by these models the state of each subfield was classified as one of these four situations: marginal, if it presents a level and slope lower than the average; emerging, if it presents a lower level but higher slope; established, if it presents a higher level but lower slope; and core, if it presents both higher level and slope. This analysis was performed dividing the time range into four equal periods, which allows depicting the evolution of each subfield over the lifetime of the technology. Different cases were found, such as laser cladding, which has been a core subfield for all the periods; or copolyestercarbonates, which has gone through the phases of emerging, established and marginal, respectively. The projection was then generated as follows: $m_{T+h} = m_T + b_T * h, h = 1, 2, \ldots$. Where m_{T+h} is the forecast of the trend, h periods ahead and m_T and b_T are the estimates of the level and slope of the trend at the end of the sample. These forecasts along with their RMSE are computed using STAMP 8.2 software (Koopman et al. 2009). The projections reveal new products to be available in the market, such as 3D printed blisk blades and 3D printed acoustic liners. Finally, the TRM was drawn based on the previous information. The main two layers were divided into the mentioned main areas, the SAOs were placed taking into account their characteristics, and short-term future period was completed taking the projections into account.

Future lines of research should consider the integration of webscraping for the identification of SAO structures in technology marketers' webpages, which would be integrated in a market layer of the TRM. Thus, the TRM would contain the whole itinerary of technology, from basic research to commercialization.

References

Fader A, Soderland S, Etzioni O (2011). Identifying relations for open information extraction. In Proceedings of the Conference on Empirical Methods in Natural Language Processing . 1535-1545. *Association for Computational Linguistics.*

Harvey AC (1989). Forecasting, Structural Time Series and the Kalman Filter, Cambridge: *Cambridge University Press*.

Kaufman L, Rousseeuw PJ (2009). Finding groups in data: an introduction to cluster analysis. Vol. 344. *John Wiley & Sons.*

Koopman SJ, Harvey AC, Doornik JA, Shephard N (2009). STAMP 8.2: Structural Time Series Analyser, Modeler, and Predictor. *London: Timberlake Consultants*.

O'Brien JJ, Carley S, Porter AL (2013). Keyword field cleaning through ClusterSuite: a term-clumping tool for VantagePoint software. *Poster presented at Global Tech Mining Conference, Atlanta, GA*.

Wang X, Ma P, Huang Y, Guo J, Zhu D, Porter AL, Wang Z. (2017). Combining SAO semantic analysis and morphology analysis to identify technology opportunities. *Scientometrics*, 111(1), 3-24.