Identifying emerging technologies with NLP-powered networks

Antoine MATHIEU COLLIN[1]

Promotor - Reinhilde VEUGELERS[1]
Co-promotor - Sien MOENS[2]

8/06/2020

[1]KU Leuven, FEB, Department of Management, Strategy and Innovation
[2]KU Leuven, FirW, Department of Computer Science

Emerging technologies

- Identifying emerging technologies and understanding how they appear is a priority concern for policy-makers, since they can have *“a revolutionary impact on the economy and society”* (Martin 1995).
- European Union: Battery Alliance (2017) and a pan-European research and innovation project in all segments of the battery value chain (2019);
- France and Germany: initiative for a cloud computing ecosystem called Gaia-X (2020).

Research questions

Three main research questions in innovation economics I will contribute to answer are:

- (1) How and where does emerging technologies arise?
- (2) How do emerging technologies evolve, compete, and spread over time?
- (3) What are the drivers of emerging technologies?

Definitions

There is no consensus in the economic literature about what defines an emerging technology:

- An important technological impact that can create or reshape an entire industry (Day and Schoemaker 2000);
- A high economic impact on the following 15 years (Porter et al. 2002).

An encompassing definition has been proposed in Rotolo, Hicks, and Martin (2015)'s literature review. An emerging technology has 5 characteristics:

1. Radical novelty;
2. Relatively fast-growth;
3. Coherence;
4. Prominent impact;
5. Uncertainty and ambiguity.

Methodologies to identify emerging technologies I

The major empirical approaches used are (Rotolo, Hicks, and Martin 2015):

- **Indicators based on patent data and trends**
  - Cohesion between documents (Watts and Porter 2003)
  - Model the growth of citations counts as S-curves (Porter and Detampel 1995)
  - Epidemic model to describe the number of authors in a given field (Bettencourt et al. 2008)
  - Normalised searching traffing on Google (Jun, Yeom, and Son 2014)
- **Citation analysis**
  - Seminal paper (Garfield 1955)
  - Clusters of publications (Kajikawa et al. 2008)
  - Network topology indicators (Iwami et al. 2014)
  - Co-citations (Small 1973) (Boyack, Small, and Klavans 2013)
  - Bibliographic coupling (Kessler 1963)

Methodologies to identify emerging technologies II

- **Co-word analysis**
  - Seminal paper (Callon et al. 1983)
  - Clustering (Lee 2008)
  - Session of conferences (Furukawa et al. 2015)
- **Overlay mapping** (Rotolo et al. 2017)
  - On geographical maps
  - Co-authorship
  - Intellectual space
- **Hybrid approaches** (Glänzel and Thijs 2012)

Methodologies to identify emerging technologies III

And since 2015, some other techniques have shown up, mainly for technological forecasting:

- **NLP** (Han et al. 2017)
- **Machine learning** (Aristodemou and Tietze 2018) (Krallinger et al. 2015)
  - Random Forest
  - K-means
  - Support Vector Machines
  - Artificial Neural Networks (Lee et al. 2018)
  - Deep Learning

The Contribution of this Paper I

What is the expected contribution of this research project to the literature?

- Build a robust method to identify emerging technologies using full-text data of patents;
- Identify the key factors that make a technology emerge;
- Apply the methodology to various technological fields to gain sectorial knowledge.

The Contribution of this Paper II

The methodology follows a four steps process:

- (1) Identification of links between technological items and assessment of their quality with NLP techniques;
- (2) Modelling of the technological knowledge as large-scale networks;
- (3) Tracking of the emergence and diffusion of technological ideas with community detection algorithms;
- (4) Statistical analysis to identify the the factors driving the emergence of new technologies.

Data

For doing this research, three datasets are available:

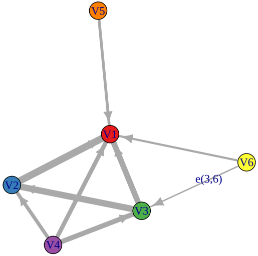
- Patent data: *PATSTAT*;
- Patent text: *EP full-text data for text analytics*;
- Financial and ownership information about the patentees: *ORBIS*.

Modelling

The NLP-powered patent network

We model the state of the patent network at a moment t as a directed graph $G_t = (V, e, \mu)$, where:

- $V = V_1, ..., V_N$ is the set of vertices of the graph G_t;
- $e = (i, j) \in V^2$ denotes the set of pairs of vertices forming the edges of G; and
- $\mu : e \to \mathbb{R}, \mu : (i, j) \mapsto e_{i,j}$ is a weighting function for the edges edge.



Modelling

Text preprocessing

- For each patent $P_k \in V$, we assign a feature vector $P_k = (p_{k1}, ..., p_{km})$ composed of all words p contains in its title, abstract and claims, with m the total number of words contained in the patent textual information;
- Let ϕ be the function which for any word return its stem word. A stemmed feature vector $\phi(P_k) = (\phi(p_{k1}), ..., \phi(p_{km}))$ is now assigned to each patent;
- We subsquently compare the stemmed feature vectors of patents using the bag-of-words (BOF) model widely used in text mining, and encoding patents according to their term frequency–inverse document frequency (TF-IDF) statistic.

Modelling

Vectorisation

- For each element $t \in W$ of the vocabulary and each patent $P_k \in V$, we compute $\mathrm{tfidf}(t, P_k, W) = \mathrm{tf}(t, P_k) \cdot \mathrm{idf}(t, W)$ where:
  - tf is the (logarithmically scaled) term frequency of the term t in the stemmed feature vector and;
  - $\mathrm{idf}(t, W)$ the (logarithmically scaled) inverse document frequency of the term t in the stemmed corpus.
- For each patent P_k we define the *term frequency–inverse document frequency vector*:

$$V_k = (\mathrm{tfidf}(1, P_k, W), ..., \mathrm{tfidf}(t, P_k, W))$$

- Which yields the bag-of-words (feature space) obtained is a $k \times t$ matrix:

$$BOW = \begin{pmatrix} V_1 \\ \vdots \\ V_N \end{pmatrix} = \begin{pmatrix} \mathrm{tfidf}(1, P_1, W) & \dots & \mathrm{tfidf}(t, P_1, W) \\ \vdots & \ddots & \vdots \\ \mathrm{tfidf}(1, P_N, W) & \dots & \mathrm{tfidf}(t, P_N, W) \end{pmatrix}$$

Modelling

Definition of the patent content similarity metric

- For each patent P_k, its term frequency–inverse document frequency vector V_k is a defined point in a t-dimensional space;
- For any pairs of patents P_i and P_j, we define:

$$\text{similarity}(P_i, P_j) = \cos(\theta) = \frac{\mathbf{V_i} \cdot \mathbf{V_j}}{\|\mathbf{V_i}\| \|\mathbf{V_j}\|} = \frac{\sum_{k=1}^{n} V_{ik} V_{jk}}{\sqrt{\sum_{k=1}^{n} V_{ik}^2} \sqrt{\sum_{k=1}^{n} V_{jk}^2}}$$

where θ is the angle between the vectors V_i and V_j.

Modelling

Weighting the edges of the network

- We use the similarity measure to link the patents in the network. We use direct and indirect citation links:
  - Direct backwards citation (at the patent family level);
  - Co-citations (CC);
  - Biographic coupling (BC);
  - Longitudinal coupling (LC).
- For any pairs of patents P_i and P_j, the presence or not of a citation link $C_{i,j}$ is denoted by the indicator function:

$$\mathbb{1}_{C_{i,j}} = \begin{cases} 1 & \text{if } P_i \text{ cites } P_j \text{ or conversely} \\ 0 & \text{else} \end{cases}$$

and the edge weighting function is defined by:

$$\mu(P_i, P_j) = \begin{cases} \text{similarity}(P_i, P_j) & \text{if } \mathbb{1}_{C_{i,j}} = 1 \\ 0 & \text{else} \end{cases}$$

Modelling

Evolution of the patent network over time

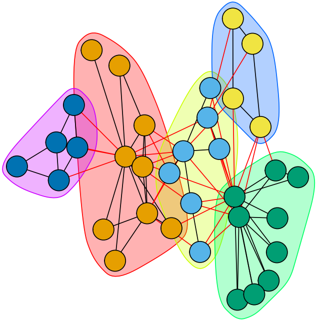
- The network composed of all patents filled under a given jurisdiction is a dynamic network which grows in size as new patents are registered.
- Formally, the network at time t is G_t, and for all dates t_1 and t_2 with $t_2 > t_1$ in the time period of interest, G_{t_1} is a subgraph of G_{t_2}.
- Not only the patents in the network change, but also their ownership and all the variables associated to their patentee (a start-up can be acquired by a large company some years later)!

Modelling

Community detection in weighted dynamic networks

In this setup, identifying clusters is a complex task:

- Large scale weighted networks;
- Dynamics: how to keep track of communities over time?

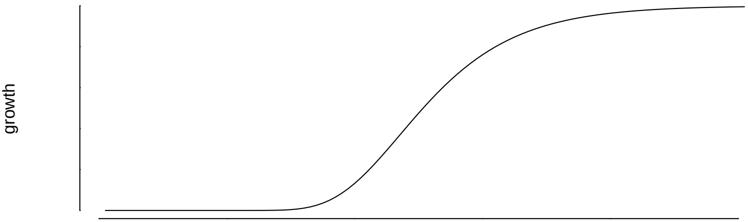


- Stabilised Louvain Method (Aynaud and Guillaume 2010) (derived from the Louvain Modularity Method (Blondel et al. 2008)).

Modelling

Growth rate of detected communities

- We track the communities over time using the Jaccard distance.
- For each community, we are able to follow its growth rate and patent composition over time.
- Emerging technologies trajectories typically follow S-curves *à la* Gompertz (Pezzoni et al. 2019), $f(t) = a\mathrm{e}^{-b\mathrm{e}^{-ct}}$. For each technology, once we identify the parameters of the curves, we know its trajectory.



Implementation

- Python: the core of the model;
- SQL for fast retrieving of the data;
- R for statistical analysis.

Applications

Drivers of innovation

- Using patent and firm-level information for network topology inference.

Micro focus

- Study the preferential attachement for each patent and to track in the following months and years whether this patent has played a key role in a growing community.

Case study

- Electric Vehicles Batteries (with Simone Tagliapietra).

Case study - First results I

Objectives of the exercise

- Assess the quality of the data and identify difficulties
- Identify the best performing NLP techniques for patent classification
  - Vectorisation method
  - Classification algorithm
  - Distance metric
- Assess the quality of the measure
  - Silhouette plot
  - Visualisation of the clusters
  - Correlation of text clustering with CPC classes

Case study - First results II

- K-means clustering using cosine distance;
- 5 clusters identified (power supply, battery pack, hybrid vehicules, power transmission and battery composition).

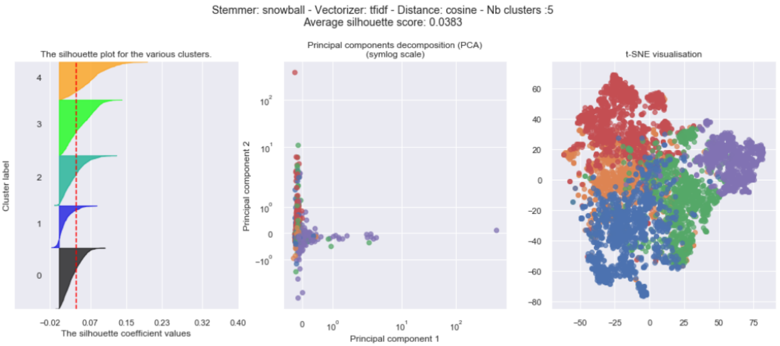


Figure 1: Clustering of breakthought patents in the Electric Vehicle Battery field

Next steps

- Implementation of the model on the EV battery field;
- Testing the methodology on past data (e.g. before 2010) to see of the model successfully predicts the emergence of the new technologies (e.g. post 2020);
- Enrich the network with corporate and ownership data from ORBIS to identify the characteristics of firms at the origin of emerging technologies.

References I

Aristodemou, Leonidas, and Frank Tietze. 2018. "The State-of-the-Art on Intellectual Property Analytics (Ipa): A Literature Review on Artificial Intelligence, Machine Learning and Deep Learning Methods for Analysing Intellectual Property (Ip) Data." *World Patent Information* 55: 37–51.

Aynaud, Thomas, and Jean-Loup Guillaume. 2010. "Static Community Detection Algorithms for Evolving Networks." In *8th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks*, 513–19. IEEE.

Bettencourt, Lus, David Kaiser, Jasleen Kaur, Carlos Castillo-Chavez, and David Wojick. 2008. "Population Modeling of the Emergence and Development of Scientific Fields." *Scientometrics* 75 (3): 495–518.

References II

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. "Fast Unfolding of Communities in Large Networks." *Journal of Statistical Mechanics: Theory and Experiment* 2008 (10): P10008.

Boyack, Kevin W, Henry Small, and Richard Klavans. 2013. "Improving the Accuracy of Co-Citation Clustering Using Full Text." *Journal of the American Society for Information Science and Technology* 64 (9): 1759–67.

Callon, Michel, Jean-Pierre Courtial, William A Turner, and Serge Bauin. 1983. "From Translations to Problematic Networks: An Introduction to Co-Word Analysis." *Information (International Social Science Council)* 22 (2): 191–235.

Day, G. S., and P. J. H. Schoemaker. 2000. "Avoiding the Pitfalls of Emerging Technologies." *California Management Review*, no. 2: 8–33. https://doi.org/10.2307/41166030.

References III

Furukawa, T., K. Mori, K. Arino, K. Hayashi, and N. Shirakawa. 2015. “Identifying the Evolutionary Process of Emerging Technologies: A Chronological Network Analysis of World Wide Web Conference Sessions.” *Technological Forecasting and Social Change* 91: 280–94. https://doi.org/10.1016/j.techfore.2014.03.013.

Garfield, Eugene. 1955. “Citation Indexes for Science.” *Science* 122 (3159): 108–11.

Glänzel, Wolfgang, and Bart Thijs. 2012. “Using ‘Core Documents’ for Detecting and Labelling New Emerging Topics.” *Scientometrics* 91 (2): 399–416.

Han, Qi, Florian Heimerl, Joan Codina-Filba, Steffen Lohmann, Leo Wanner, and Thomas Ertl. 2017. “Visual Patent Trend Analysis for Informed Decision Making in Technology Management.” *World Patent Information* 49: 34–42.

References IV

Iwami, Shino, Junichiro Mori, Ichiro Sakata, and Yuya Kajikawa. 2014. "Detection Method of Emerging Leading Papers Using Time Transition." *Scientometrics* 101 (2): 1515–33.

Jun, Seung-Pyo, Jaeho Yeom, and Jong-Ku Son. 2014. "A Study of the Method Using Search Traffic to Analyze New Technology Adoption." *Technological Forecasting and Social Change* 81: 82–95.

Kajikawa, Yuya, Junta Yoshikawa, Yoshiyuki Takeda, and Katsumori Matsushima. 2008. "Tracking Emerging Technologies in Energy Research: Toward a Roadmap for Sustainable Energy." *Technological Forecasting and Social Change* 75 (6): 771–82.

Kessler, Maxwell Mirton. 1963. "Bibliographic Coupling Between Scientific Papers." *American Documentation* 14 (1): 10–25.

References V

Krallinger, Martin, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, et al. 2015. "The Chemdner Corpus of Chemicals and Drugs and Its Annotation Principles." *Journal of Cheminformatics* 7 (1): 1–17.

Lee, Changyong, Ohjin Kwon, Myeongjung Kim, and Daeil Kwon. 2018. "Early Identification of Emerging Technologies: A Machine Learning Approach Using Multiple Patent Indicators." *Technological Forecasting and Social Change* 127: 291–303.

Lee, Woo. 2008. "How to Identify Emerging Research Fields Using Scientometrics: An Example in the Field of Information Security." *Scientometrics* 76 (3): 503–25.

Martin, Ben R. 1995. "Foresight in Science and Technology." *Technology Analysis & Strategic Management* 7 (2): 139–68.

References VI

Pezzoni, Michele, Reinhilde Veugelers, Fabiana Visentin, and others. 2019. "How Fast Is This Novel Technology Going to Be a Hit?" CEPR Discussion Papers.

Porter, Alan L, and Michael J Detampel. 1995. "Technology Opportunities Analysis." *Technological Forecasting and Social Change* 49 (3): 237–55.

Porter, A. L., J. D. Roessner, X.-Y. Jin, and N. C. Newman. 2002. "Measuring National 'Emerging Technology' Capabilities." *Science and Public Policy* 29 (3): 189–200. https://doi.org/10.3152/147154302781781001.

Rotolo, Daniele, Diana Hicks, and Ben R Martin. 2015. "What Is an Emerging Technology?" *Research Policy* 44 (10): 1827–43.

References VII

Rotolo, Daniele, Ismael Rafols, Michael M Hopkins, and Loet Leydesdorff. 2017. "Strategic Intelligence on Emerging Technologies: Scientometric Overlay Mapping." *Journal of the Association for Information Science and Technology* 68 (1): 214–33.

Small, Henry. 1973. "Co-Citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents." *Journal of the American Society for Information Science* 24 (4): 265–69.

Watts, Robert J, and Alan L Porter. 2003. "R&D Cluster Quality Measures and Technology Maturity." *Technological Forecasting and Social Change* 70 (8): 735–58.