Identifying emerging technologies with NLP-based patent networks: an application to EV batteries

Antoine Mathieu Collin[a], Reinhilde Veugelers[a,b,c], Simone Tagliapietra[b,d,e]

[a] *KU Leuven, Department of Managerial Economics, Strategy and Innovation, Faculty of Business and Economics, Leuven, Belgium.*
[b] *Bruegel, Brussels, Belgium*
[c] *CEPR, London, United Kingdom*
[d] *Fondazione Eni Enrico Mattei, Milan, Italy*
[e] *The Johns Hopkins University, SAIS Europe, Bologna, Italy*

Abstract

We use EP full-text data for text analytics in combinaison with PATSTAT and Orbis, which contains financial and ownership data on corporate and public companies. (a) We first identify in PATSTAT breakthrough patents in the field of EV battery technologies, and retrive their corresponding textual claims and ownership information by merging PATSTAT with the other two databases. (b) After vectorising the text content in a vector space model (VSM), the enriched dataset is subsequently used to model the technology space in EV battery technologies as a statistical (directed) multi-graph. In this multi-graph, the patents are the nodes, while the edges are either the forward citations weighted by the cosine similarity of the textual content of their patent claims. (c) We then explore the dynamics of the evolution of this network over time to identify the fastest growing communties (emerging technologies) and (d) the factors driving the emergence of new technologies in this technological field via inference methods (R&D&I expenditure, company size, corporate vs university, etc..).

1. Introduction

1.1. Research question and main contribution

The digitalization of the patent literature over the past decades and the emergence of natural language processing (NLP) techniques now allow to use the very rich and largely unexplored information contained in the inside of patents (abstract, patent claims). While patent metadata has been largely used in the field of innovation economics and in technological forecasting, patent text data is still a largely unexplored frontier. Until the most recent years, only researchers and practitioners were able to process this textual information and to assess qualitatively how patents are interacting with each other, in a time-consuming

Email addresses: antoine.mathieucollin@kuleuven.be (Antoine Mathieu Collin), reinhilde.veugelers@kuleuven.be (Reinhilde Veugelers), simone.tagliapietra@bruegel.org (Simone Tagliapietra)

process. The use of bibliometrics mesures such as citation counts or keywords allow to link patents with one another, but largely fails to assess the nature of the interaction between a given pair of patents linked by a citation: for instance, a patent can consist in a slight incremental improvement of a previous patented invention, cite another one for the sake of completness, or import a technique from another field and be the starting point of a whole new branch of its discipline. By combining NLP techniques to compare the similarity of any pair of patents (claims) and statistical analysis of the associated patent networks, we were able to identify in the EV battery technology field emerging technologies. After enriching our network by corporate information from ORBIS about our assignees, we infer the drivers of the growth of network communities using a *regression model* [to adjust with the model most accurate to infer causality effect whithin the data produced] and identified *X, Y and Z* as key factors driving innovation in this field [to be updated].

1.2. Literature review

1.2.1. Similarity and classic metrics in patent data

[Add literature review]

1.2.2. NLP with patent data

[Add literature review]

1.2.3. NLP powered-networks

[Add literature review]

1.2.4. Community detection in evolving networks

[Add literature review]

1.3. Novelties

The originality of our approach resides in the combination of natural language processing techniques and statistical analysis of network data. This allows us to build a more realistic patent network which takes into account the actual patent claims and does not rely on patent technology classes.

2. Data

2.1. Patent metadata

Patent data has been retrieved from the PATSTAT database of the European Patent Office (EPO). PATSTAT contains bibliographical and legal status patent data from leading industrialised and developing countries. We extracted information on patent data in the EV battery technological field from the bulk EPO dabase, based on a list of CPC codes forming the EV battery technological landscape as established by the Joint Research Center (JRC) of the European Commission (European Commission - Joint Research Centre et al., 2017).

[Add summary statistics table]

2.2. Patent text data

The EPO provides users with EP publication data as character-coded full-text in the product EP full-text data.[1] This product consists of all EP publications in XML, PDF and TIFF format (where available), making it a very comprehensive data set. EP full-text data for text analytics contains all EP publications, from 1978 until the end of January 2019. The data set, which will be updated annually, comprises approximately 5.8 million EP publications and is about 210 GB in size.

[Add summary statistics table]

2.3. Financial, ownership and geographic data on public and corporate undertakings from ORBIS

Orbis has information on more than 360 million companies across the globe, notably regarding financial and ownership information. We use the ORBIS flat files, which contains historical ownership information, in order to capture the changes in corporate structures and therefore changes in patent ownership over time.

[Add summary statistics table]

3. Methodology

3.1. The NLP-based patent network model

3.1.1. Definition of the 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

[1] The EPO grants permission to use EP full-text data for text analytics data under the "Creative Commons Attribution 4.0 International Public license". https://www.epo.org/searching-for-patents/data/bulk-data- sets/data.html#tab-1

$\mu : e \to \mathbb{R}, \mu : (i,j) \mapsto e_{i,j}$ is the function a weight to each edge. In this graph, the vertices (nodes) are the N patents uniquely identified, while the vertices represent the (directed) citation links from the citing (backward citation) and the cited patent (forward citation). The edges (citations) are subsequently weighted according to the similarity of their title, abstract and claims, computed using NLP techniques described below which define μ. The vertices are decorated using patent and patentee information from PATSTAT and financial and ownership data about the patentee from Orbis.

3.1.2. Definition of the patent content similarity metric

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. The feature vector is not weighted at this stage, and words can appear multiple times. We call *corpus* the set of the N features vectors.

[Add figure]

3.1.2.1. Stemming. Natural language can use different forms of a word, such as "process", "processing" and "processed", and it exists families or words with similar meanings, like "technology" and "technologically". To cope with this grammatical complexity, we reduce every word used in the EP full-text database to its linguistic root (word stem). We perform this task using the Porter (Snowball) algorithm, which is a standard algorithm for performing stemming (Porter, 1980). 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, and we call the *stemmed corpus* the set composed of all stemmed feature vectors.

3.1.2.2. Vectorisation with TF-IDF. In patent documents, like in most real-word corpuses, meaning is most likely located in the most rare terms, while most frequent terms are less useful to compare patents among each other. For instance, one can expect any patent to comprise stopwords ("a", "about", "above", etc.) or words belonging to the lexical field of innovation ("process", "invention") with a high frequency. 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. Let W be the vocabulary of the stemmed corpus, i.e. the set of unique stemmed words it contains. We note $p = card(W)$ the length of the vocabulary. For each element $t \in W$ of the vocabulary and each patent $P_k \in V$, we compute $\text{tfidf}(t, P_k, W) = \text{tf}(t, P_k) \cdot \text{idf}(t, W)$ where tf is the (logarithmically scaled) term frequency of the term t in the stemmed feature vector and $\text{idf}(t, W)$ 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* $V_k = (\text{tfidf}(1, P_k, W), ..., \text{tfidf}(t, P_k, W))$. The bag-of-words (feature space) obtained is a $k \times t$ matrix:

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

3.1.2.3. Patent similarity. For each patent P_k, its term frequency–inverse document frequency vector V_k is a defined point in a t-dimensional space, and we can define a similarity measure between any pair of patents in the feature space. As the feature space is highly dimentional and very sparse (since each patent contains only a restricted number of words from the total voculabulary for all patents, the bag-of-words is a sparse matrix), we use the cosine distance to measure similarity between the textual content of patents. 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.

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:

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

3.1.3. Evolution of the patent network over time

The network composed of all patents filled under a given juridiction is a dynamic network which grows in size as new patents are registered. These patents citing previous work constitute in the model new nodes that attach themselves to the existing network, making its structure evolve over time. To capture network dynamics, study the drivers explaining nodes attachement to some patents rather than others and to keep track of the evolution of the communities (clusters) the patent network contains over time, we keep track of the state of the patent network over time. The network is defined in a static state at the beginning of the period of interest and denoted G_0. We add one by one, based on their application filing date, the nodes in the network. 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}. Financial and ownership information about the companies are updated at the end of each calendar year, following the availaility of the ORBIS data.

3.2. Identification of emerging technologies as fastest growing communities in the defined network

3.2.1. Community detection in weighthed graphs

Each time a new node is attached to the network and the state of the network changes, the Stabilised Louvain Method (Aynaud and Guillaume, 2010) is applied to identify the different clusters composing the network. This method is derived from the Louvain Modularity Method (Blondel et al., 2008) well suited to identify communites in large weighted networks, and optimise to ensure the stability of the detected communities in order to facilitate the sudy of their evolution over time. Indeed, classic clustering method tend to exhibit an unwanted impredictability and tend to merge, split or detect new communities easily when the topology of the network changes slighlty. On the contrary, the Stabilised Louvain Method reduces the impredictability of the community detection at the expense of a slighlty reduce accurancy in a non-dynamic setup.

3.2.2. Growth rate of detected communities

By constuction of our model, as the network G_t is a subset of the network G_{t+1} containing just one node less, the number and the internal composition of both communities *shoud* [verify this point in practice as a consistency check] should remain the same at one community of difference maximum (the inclusion of the new node can *only* let the community structures inchanged, split one community in two parts or led to the merge of two commmunities). The link between communities of both network is then done by computing the Jaccard measure. With this setup, for each community, we are able to follow its evolving size and patent composition over time.

[Add figure]

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

Moreover, we are able to 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 (emerging technology), whether it has change community over time (the patent was a pionner in a new field that has grown and detached itself from the previous main research area) or if has been a reference point for several newly formed communities. At the same time, we use a *regression model* [to adjust with the model most accurate to infer causality effect whithin the data produced] to identify the patent/firm specific features which allow a corporate groups or universities to take gain a technological advantage by creating an entirely new technological field.

[Add equation of the statistical model chosen for topology inference]

4. Case study

[Add a short introduction on why this specific field has been chosen]

4.1. Review of NLP of patent data applied to EV battery technologies

[Add literature review of natural language processing or patent network in the EV/EV battery field]

4.2. Data selection and data cleaning

[Once fixed, explain the data selection (SQL queries from PATSTAT, Orbis and patent text) and data cleaning process]

4.3. Fitting the model

[Add explanation on how the theoretical model is filled with the data]

4.4. Identification of emerging technologies and key innovators

[Add information on the technology detection and the leader companies of the fastest growing technologies]

[Add figure]

4.5. Inference results

[Add inference results]

5. Conclusions

[Add conclusions]

6. Acknowledgements

[Add Acknowledgements]

References

Aynaud, T., Guillaume, J.L., 2010. Static community detection algorithms for evolving networks, in: WiOpt 2010 - 8th Intl. Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks.

Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment. doi:10.1088/1742-5468/2008/10/P10008

European Commission - Joint Research Centre, Fiorini, A., Georgakaki, A., Pasimeni, F., Tzimas, E., 2017. Monitoring R&I in Low-Carbon Energy Technologies. doi:10.2760/434051

Porter, M., 1980. The Porter Stemming Algorithm.