An Hypernetwork Approach to Measure Technological Innovation - Submission to PLOS Journals

Antonin Bergeaud[1,☯], Yoann Potiron[2,☯], Juste Raimbault[3,☯]

1 Department of Economics, London School of Economics, London, UK
2 Department of Statistics, University of Chicago, Chicago, US
3 UMR 8504 Géographie-cités, Université Paris VII, Paris, France

☯These authors contributed equally to this work.
*** le.corresponding@polytechnique.edu**

Abstract

Introduction

The study of innovation through the lens of technological patents is not a novel idea [1] but the recent rise of new methods and computational abilities, including data-mining and network analysis [2] has shed a new light on the approach. With methods relatively close to applied epistemology studies such as citation dynamics modeling [3] or co-autorships networks analysis [4], recent works have studied patents citation network to understand the processes of technological innovation. As in science, where reflexivity is crucial and is becoming a mandatory step to build future research agendas, as e.g. in the recent analysis on 20th century physics [5], the structure of technology and particularly of technological innovation should show special patterns which understanding must have positive feedback on the economy.

A consequent amount of research already proposed to use semantic networks to study technological domains. One of the first works to enhance the approach was [6], where the idea of visualizing keywords network was introduced and illustrated on a small technological domain. Semantic analysis has already proved its efficiency in various fields, such as [7,8] for technology studies, or [9] in political science for example. We will also be interested in measures based on technological classification, as in [10] where the study of the distribution of classes within patents leads to confirm the combinatory nature of the innovation process. Advanced citation-based measures have been proposed in order to gain an order of information extracted from the citation network [11], but we will stay to rather simple indicators concerning this network as we will cross it with other type of data. Concerning dynamic analysis, models of citation processes have been proposed and fit to data such as in [12], and depending on temporal patterns we observe on features, we may propose to use dynamic models for network evolution.

Materials and Methods

Raw Data

Semantic Network Construction

We first assign to a patent $p \in \mathcal{P}$ a set of significant keywords $K(p) \in \bigcup_{n\in\mathbb{N}} \mathcal{A}^{*n}$, that are precisely extracted following a similar procedure to the one detailed in [13] :

- Text parsing and tokenizing.
- Part-of-speech tagging, normalization.
- Stem extraction and multi-stems constructions.
- Relevant multi-stems filtering.

The semantic network is then constructed by co-occurrences analysis : nodes are keywords, i.e. $V = \bigcup_{p \in \mathcal{P}} K(p)$, and edges represent co-occurrences : $E = \{(k, k') | \exists p \text{ s.t. } k, k' \in K(p)\}$. This specification aims to capture semantic communities that do necessarily have a thematic meaning and that should correspond to a specific domain, field, technology or technique.

Text processing operations will be implemented in `python` in order to use the `nltk` library [14] which is highly ergonomic and supports most advanced state-of-the-art natural language processing operations. Source code is openly available on the repository of the project[1]. Steps one to three are directly done using built-in functions of the library. Step four needs a particular treatment that we propose as an extension of the original method for large corpuses, which is detailed in the following.

Bootstrap on random sub-corpuses for relevance estimation Once multi-stems have been extracted, one scores them by *unithood*, defined for the multi-stem i by $u_i = f_i \cdot \log(1 + l_i)$ where f_i is the number of apparitions of the multi-stem over the whole corpus and l_i its length in words. Let K_w the maximal number of relevant keywords per patent. If N is the total number of relevant keywords extracted, that we consider as a parameter, the heuristic described in [13] proposes a first filtration of $k \cdot N$ keywords on the whole corpus (and take a fixed value $k = 4$), and then a filtration on a secondary score called *termhood*, computed as a chi-squared score on the distribution of cooccurrences of the stem, compared to an uniform distribution within the whole corpus. More precisely, one computes the co-occurence matrix (M_{ij}), defined as the number of patents where stems i and j appear together, what allows to define the *termhood* score as

$$t_i = \sum_{j \neq i} \frac{\left(M_{ij} - \sum_k M_{ik} \sum_k M_{jk}\right)^2}{\sum_k M_{ik} \sum_k M_{jk}}$$

One issue arising when working with large corpuses is the square complexity for determining cooccurrences, that can be simplified at best to the sum of squared sizes of pre-relevant keywords for each patent.

Results

Semantic Communities Dynamics

Layer Structure Comparison

[15]

Unsupervised Data Mining

Features

[1] at `url`

- Citation relative centrality regarding technological or semantic classes : given a partition of $\mathcal{P}$ represented by a classification function C (constructed for example by clustering or community detection within technological or semantic networks), a vector feature is for patent i

$$\left(\frac{\sum_{j \in c} Cit^{out}(i,j)}{\sum_{j \in c} Cit^{in}(i,j)} \right)_{c \in C(\mathcal{P})}$$

- Dynamic evolution of classification vector : if C_t is stratified over successive time periods indexed by time t, either $\Delta \vec{C}_t(i)$ if $\vec{C}_t$ is a vector of probabilities to belong to each class (in case of a Bayesian approach), or $\Delta(C_t(j))_{Cit(i,j,t) \neq 0}$ in case of a deterministic approach, could both be interesting features.

- Deviation from the expected classes given position in other layers of the hypernetwork (it would need explorations if these conditional probabilities first can be well estimated, then if they contain relevant information).

Discussion

References

1. Basberg BL. Patents and the measurement of technological change: a survey of the literature. Research policy. 1987;16(2):131–141.
2. Newman M. Networks: an introduction. Oxford University Press; 2010.
3. Newman MEJ. Prediction of highly cited papers. ArXiv e-prints. 2013 Oct;.
4. Sarigöl E, Pfitzner R, Scholtes I, Garas A, Schweitzer F. Predicting Scientific Success Based on Coauthorship Networks. ArXiv e-prints. 2014 Feb;.
5. Sinatra R, Deville P, Szell M, Wang D, Barabasi AL. A century of physics. Nat Phys. 2015 10;11(10):791–796. Available from: http://dx.doi.org/10.1038/nphys3494.
6. Yoon B, Park Y. A text-mining-based patent network: Analytical tool for high-technology trend. The Journal of High Technology Management Research. 2004;15(1):37–50.
7. Choi J, Hwang YS. Patent keyword network analysis for improving technology development efficiency. Technological Forecasting and Social Change. 2014;83:170–182.
8. Fattori M, Pedrazzi G, Turra R. Text mining applied to patent mapping: a practical business case. World Patent Information. 2003;25(4):335–342.
9. Gurciullo S, Smallegan M, Pereda M, Battiston F, Patania A, Poledna S, et al. Complex Politics: A Quantitative Semantic and Topological Analysis of UK House of Commons Debates. ArXiv e-prints. 2015 Oct;.
10. Youn H, Strumsky D, Bettencourt LMA, Lobo J. Invention as a combinatorial process: evidence from US patents. Journal of The Royal Society Interface. 2015;12(106).
11. Alstott J, Triulzi G, Yan B, Luo J. Mapping Technology Space by Normalizing Technology Relatedness Networks. ArXiv e-prints. 2015 Sep;.

12. Valverde S, Solé RV, Bedau MA, Packard N. Topology and evolution of technology innovation networks. Physical Review E. 2007;76(5):056118.

13. Chavalarias D, Cointet JP. Phylomemetic patterns in science evolution—the rise and fall of scientific fields. Plos One. 2013;8(2):e54847.

14. NLTK. Natural Language Toolkit, Stanford Univeristy; 2015.

15. Iacovacci J, Wu Z, Bianconi G. Mesoscopic Structures Reveal the Network Between the Layers of Multiplex Datasets. arXiv preprint arXiv:150503824. 2015;.