An Hypernetwork Approach to Accurately Measure Technological Innovation

Project Proposal

J. Raimbault[1,2] and A. Bergeaud[3]

[1] UMR CNRS 8504 Géographie-cités
[2] UMR-T IFSTTAR 9403 LVMT
[3] London School of Economics

Abstract

Patents are a central proxy in the study of the economy of innovation. Indeed, the information contained in relations between patents reflects the underlying structure of the socio-technological system of research and development in innovative companies. Whereas recent focus was mainly on the study of patterns in the inter-patent citation network, taking technological fields as externally fixed, we propose a novel approach based on semantic analysis of patents textual contents. Indeed, measures such as co-occurences and repetitions of keywords should contain a slightly different information than the one extracted from the citation network, as for example links between domains, or informal domains that could appear as communities in the semantic network. The aim of this project is to investigate the nature and extent of this information, by the mining of regular patterns in features that should be established. We expect to test various features crossing measures extracted from both dynamic citation and semantic network, by unsupervised, and supervised if needed, datamining techniques. Expected results are to unveil the potentialities of such approach, and in the case of significant information, to apply it to classify patents and be able to systematically differentiate innovatives from imitating patents.

1 Introduction

The study of innovation through the lens of technological patents is not a novel idea [Basberg, 1987] but the recent rise of new methods and computational abilities, including datamining and network analysis [Newman, 2010] has shed a new light on the approach. With methods relatively close to applied epistemology studies such as citation dynamics modeling [Newman, 2013] or co-autorships networks analysis [Sarigöl et al., 2014], 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 [Sinatra et al., 2015], the structure of technology and particularly of technological innovation should show special patterns which understanding must have positive feedback on the economy. This project aims thus to fullfill the objective of exploring the potentialities of a precise measure of innovation through the mining of different aspects of patents data.

2 Research objectives

2.1 Research question

The identification of so-called *emerging research fronts* was done in the case of scientific publications in [Shibata et al., 2008]. In the same spirit, our guiding research objective is to identify such fronts for

patents, i.e. to be able to classify patents in order to distinguish the "real innovations" from imitations of existing technologies. That is relatively close to the issue of defining a relevant measure of innovation [Archibugi, 1988]. The goal of the study is therefore to build a new database based on USPTO patent applications from 1975 to 2014 where each patent would be link with an index between 0 and 1, 1 being a patent corresponding to a new product and 0 being a patent corresponding to an improvment on an existing product. This question is of major importance in modern growth theories (see Klenow et al. (forthcoming) especially to measure growth induced by incumbent improving on their own innovations [Klette and Kortum, 2004].

2.2 Link with previous work

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 [Yoon and Park, 2004], 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 [Choi and Hwang, 2014, Fattori et al., 2003] for technology studies, or [Gurciullo et al., 2015] in political science for example. We will also be interested in measures based on technological classification, as in [Youn et al., 2015] 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 [Alstott et al., 2015], 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 [Valverde et al., 2007], and depending on temporal patterns we observe on features, we may propose to use dynamic models for network evolution.

The novelty of our project relies on various points, including the systematic unsupervized research of patterns, the possibility to build hybrid features from both semantic and citation networks, the combination of various techniques from datamining to multi-layer network analysis.

3 Proposed approach

3.1 General description

This project is based on the assumption that semantic relations between patents contents must include some information on the underlying technological innovation processes. The idea is strongly inspired from the semantic science mapping proposed in [Chavalarias and Cointet, 2013], where dynamics of scientific fields was reconstructed by keywords co-occurence networks analysis. The assets of such an approach include the fact to be not dependant on predefined fields or domains, what allows specifically to uncover effective domains by community detection. Extracting significant keywords (see technical details) from titles and abstracts allows to construct semantic networks between patents or keywords, in which community detection e.g. should reveal patterns of "effective" technological fields.

Every patent is associated with many details that have been made available by national and international patent offices. First, each patent is associated with one or several technological fields which are chosen from a list by patent office specialists after careful review of the content of the patent (see IPC class). Second, every patent has to make a list of citations to all the patents used in the process, this naturally define a network between entities that have some technologies in common. From these two features innovation economists have created various indices to measure the quality of a patent (number of citations, scope of technological field, etc.). Two other indicators are of particular interest for us: originality and radicalness, their definition and expression being detailed in the next section.

Our strategy will rely on comparisons between our network based on semantic comparison and the two existing network defined above. Simple correlations and diachronic dynamic analysis for time-dependant networks, combined with unsupervised datamining for rough data exploration, should be conducted first,

and then completed by more refined datamining, using supervised learning if needed, i.e. if no significant pattern has emerged in the features constructed. The underlying assumption driving our work is that highly innovative new product are not well classified by the technological class space and should struggle to cite previous related patents. Hence, we expect such patent to have a large distance in terms of semantic analysis with patents it cites and patents in the same technological class.

3.2 Technical details

Originality Measures The originality measure is defined by Hall *et al.* (2001) [Hall et al., 2001] as

$$O_i = 1 - \sum_{j=1}^{n_i} c_{i,j}^2$$

where $c_{i,j}$is the percentage of citations made by patent i to a patent in class j out of n_i technological classes to which patent i belongs. If the scope of technologies which the patent uses and cites is large, then the originality measure will be high. Radicalness is more difficult to define. It is constructed in the same way as the originality index but here we only consider the technology classes of patents cited by patent i but to which patent i does not belong. These two indicators are good proxies and great start to estimate if a patent is protecting a new product that can hardly be classified into the official technological field space.

Citation Network We define a binary relationship between each pair of patents $Cit(i, j) = 1$ if j cites i or i cites j, otherwise $Cit(i, j) = 0$.

Technological Class Network For each patent i, let B_i be the set of technological class of i. We then define a relationship between each pair i and j as 2 times the number of technological class in common divided by the total number of class of i and j.

$$Class(i, j) = 2\frac{|B_i \cap B_j|}{|B_i| + |B_j|}$$

Thus, if two patents have no class in common, $Class(i, j) = 0$ while if the two patents are exactly identical in terms of their sets of technologial class $Class(i, j) = 1$.

Semantic Network 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 [Chavalarias and Cointet, 2013] :

- 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 [NLTK, 2015] which is highly ergonomic and supports most advanced state-of-the-art natural language processing operations.

Possible Features Content of the hypernetwork analysis is rather open at this stage of the project, as techniques used will depend on patterns found in the first explorations of each network alone and of first order structural correlations. From an unsupervised datamining perspective, relevant features to distinguish real innovation may for example include :

- 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).

4 Project organisation

Proposed research agenda :

- Project definition, litterature review - AB, JR, PA - ETA 10h
- Thematic framing (possible features, precise objectives) - AB, JR - ETA 4h
- Technical aspects (database management, text-mining implementation : JR ; citation network, technological class network : AB) - AB, JR - ETA 20h
- Empirical datamining - AB, JR - ETA 25h
- Theoretical feedback, revision of mining techniques - AB, JR, PA - ETA 10h
- Theoretical modeling, results interpretation, perspectives - AB, PA - ETA ?

References

[Alstott et al., 2015] Alstott, J., Triulzi, G., Yan, B., and Luo, J. (2015). Mapping Technology Space by Normalizing Technology Relatedness Networks. *ArXiv e-prints*.

[Archibugi, 1988] Archibugi, D. (1988). In search of a useful measure of technological innovation (to make economists happy without) discontenting technologists). *Technological Forecasting and Social Change*, 34(3):253–277.

[Basberg, 1987] Basberg, B. L. (1987). Patents and the measurement of technological change: a survey of the literature. *Research policy*, 16(2):131–141.

[Chavalarias and Cointet, 2013] Chavalarias, D. and Cointet, J.-P. (2013). Phylomemetic patterns in science evolution—the rise and fall of scientific fields. *Plos One*, 8(2):e54847.

[Choi and Hwang, 2014] Choi, J. and Hwang, Y.-S. (2014). Patent keyword network analysis for improving technology development efficiency. *Technological Forecasting and Social Change*, 83:170–182.

[Fattori et al., 2003] Fattori, M., Pedrazzi, G., and Turra, R. (2003). Text mining applied to patent mapping: a practical business case. *World Patent Information*, 25(4):335–342.

[Gurciullo et al., 2015] Gurciullo, S., Smallegan, M., Pereda, M., Battiston, F., Patania, A., Poledna, S., Hedblom, D., Tolga Oztan, B., Herzog, A., John, P., and Mikhaylov, S. (2015). Complex Politics: A Quantitative Semantic and Topological Analysis of UK House of Commons Debates. *ArXiv e-prints*.

[Hall et al., 2001] Hall, B., Jaffe, A., and Trajtenberg, M. (2001). The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools. Papers 2001-29, Tel Aviv.

[Klette and Kortum, 2004] Klette, T. J. and Kortum, S. (2004). Innovating Firms and Aggregate Innovation. *Journal of Political Economy*, 112(5):986–1018.

[Newman, 2010] Newman, M. (2010). *Networks: an introduction*. Oxford University Press.

[Newman, 2013] Newman, M. E. J. (2013). Prediction of highly cited papers. *ArXiv e-prints*.

[NLTK, 2015] NLTK (2015). Natural language toolkit, stanford univeristy.

[Sarigöl et al., 2014] Sarigöl, E., Pfitzner, R., Scholtes, I., Garas, A., and Schweitzer, F. (2014). Predicting Scientific Success Based on Coauthorship Networks. *ArXiv e-prints*.

[Shibata et al., 2008] Shibata, N., Kajikawa, Y., Takeda, Y., and Matsushima, K. (2008). Detecting emerging research fronts based on topological measures in citation networks of scientific publications. *Technovation*, 28(11):758–775.

[Sinatra et al., 2015] Sinatra, R., Deville, P., Szell, M., Wang, D., and Barabasi, A.-L. (2015). A century of physics. *Nat Phys*, 11(10):791–796.

[Valverde et al., 2007] Valverde, S., Solé, R. V., Bedau, M. A., and Packard, N. (2007). Topology and evolution of technology innovation networks. *Physical Review E*, 76(5):056118.

[Yoon and Park, 2004] Yoon, B. and Park, Y. (2004). A text-mining-based patent network: Analytical tool for high-technology trend. *The Journal of High Technology Management Research*, 15(1):37–50.

[Youn et al., 2015] Youn, H., Strumsky, D., Bettencourt, L. M. A., and Lobo, J. (2015). Invention as a combinatorial process: evidence from us patents. *Journal of The Royal Society Interface*, 12(106).