Mining Patents Data through Semantic Network Analysis

Project Proposal

J. Raimbault

September 19th

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.

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]

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.

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 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

3.2 Technical details

4 Project organisation

- 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 - ETA 10h
- Empirical datamining - AB, JR - ETA 20h
- Theoretical feedback, revision of mining techniques - AB, JR, PA - ETA 10h
- Theoretical modeling, results interpretation, perspectives - AB, PA - ETA ?

References

[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.

[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*.

[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.

[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.