Investigating Patterns of Technological Innovation

J. Raimbault^{1,2*}, A. Bergeaud³ and Y. Potiron⁴ *juste.raimbault@parisgeo.cnrs.fr

¹UMR CNRS 8504 Géographie-cités ²UMR-T IFSTTAR 9403 LVMT ³London School of Economics ⁴Faculty of Business and Commerce, Keio University

> CSS 2016 - Amsterdam Session CFS1 22th September 2016

> > うして ふぼう ふほう ふほう しょう

Technological innovation and patents

Patenting as a footprint of technology and its coevolution with science and culture [Bais, 2010]

(54)

(75)

(73)

(21)

(60)

(58)



X000001 July 31, 1790

 $\label{eq:second} MO many shows the first optic to get the first second secon$

Bity of New York hely at tope -

ingunum often ein foreight establishing of one is foreign a sumplification of the end of the second of the contemporary the terms of t

The United States, Felt Fisher, low Orsers shall mus busin

		US005285999B1				
United States Patent Page		(10) Patent No.: US 6,285,999 E (45) Date of Patent: Sep. 4, 200	1			
METHO LINKED	D FOR NODE RANKING IN A DATABASE	Craig Boyle "To link or not to link: An empirical comparison of Hypertext linking strategies". ACM 1992, pp. 221–231."				
Inventor:	Lawrence Page, Stanford, CA (US)	I. Katz, "A new status index derived from sociometric analysis," 1953, Psychometricka, vol. 18, pp. 39–43.				
Assignee:	The Board of Trustees of the Leland Stanford Junior University, Stanford,	C.H. Hubbell, "An input-output approach to clique identi fication sociometry," 1965, pp. 377–399.				
Notice:	CA (US) Subject to any disclaimer, the term of this pattern is extended or adjusted under 35 U.S.C. 154(b) by 0 days.	Misrochi et al., "Techniques for disagregating centrality scores in social networks," 1996, Sociological Methodology, pp. 25–48. E. Garfield, "Clattion analysis as a tool in journal evalua- ion," 1072 Science, vol. 123, pp. 471–470.				
Appl. No.: 09/064,827		 Pinski et al., "Citation influence for journal aggregates of scientific publications: Theory, with application to the It- erature of obvice," 1976, 167, Proc. And Management vol. 				
PDRO.	Jan. 9, 1998	12, pp. 297-312.				
Re Proviniceal 1997.	kited U.S. Application Data application No. 60/035,205, filed on Jan. 10	N. Geller, "On the citation influence methodology of Pinski and Narin," 1978, Inf. Proc. And Management, vol. 14, pp. 93–95.				
Int. CL ² U.S. CL Field of S 207	G06F 1720 70725; 70777; 707250 Search 7077100, 5, 7 513 13 10 104 501: 345140 362725	P. Doreian, "Measuring the relative standing of disciplin journals," 1988, Inf. Proc. And Management, vol. 24, p 45–56.	n P			
	229, 230, 231	(List continued on next page.)				

・ロト ・ 理 ト ・ ヨ ト ・ ヨ ト

Why study patents ?

 \rightarrow Applied Epistemology : particular case of the ecology and evolution of knowledge ; propagation of knowledge

 \rightarrow Economics [Griliches, 1998] : dynamics of innovation ; R&D returns ; innovation at the core of endogenous growth theories [Aghion and Howitt, 1992]

Examples :

- [Youn et al., 2015] interaction between technological fields ; combinatorial nature of inventions
- [Bruck et al., 2016] citation network analysis to detect emerging research front
- [Gerken and Moehrle, 2012] [Tseng et al., 2007] semantic analysis

Research Context

Technological classification by Offices does not anticipate future research fields ; retrospective reclassification at some times when classification become obsolete

However, information contained in patent themselves should contain some kind of endogenous structure, as emerge from microscopic activities of inventors/firms : e.g. semantic dynamics show combinations before classes are established by Offices

Research Objective : Investigate endogenous patterns of technological innovation contained in particular in semantic content of patents and metadata ; assess complementarity of other informations and potentialities for economic modeling

うして ふぼう ふほう ふほう しょう

Database Construction

Construction of a Database from US Patent and Trademark Office redbook 1976-2012 (full patent description), which provide raw data but on separate files and different formats

Data Collection Procedure

- Automatic download of raw data file
- Parsing depending on format : dat or xml (varying schema)
- Uniformisation and storing in MongoDB

 \rightarrow 4391272 patents with text data ; dated by application date (current state of knowledge)

Extraction of Semantic Features

Text-mining in python with nltk [Bird, 2006], method adapted from [Chavalarias and Cointet, 2013]

- Parsing and tokenizing / pos-tagging (word functions) / stemming done nltk built-in pos-tagger
- Selection of potential *n-grams* (with $1 \le n \le 3$) : English $\bigcap \{NN \cup VBG \cup JJ\}$
- $\bullet\,$ Database insertion for instantaneous utilisation (several days $\rightarrow\,$ 1min)
- Estimation of *n-grams* relevance, following co-occurrences statistical distribution (*termhood* score as chi-2 score)

Semantic Network

Co-occurrence network of keywords



Semantic Classes Construction

Communities in network not well defined (presence of connector words) : need for a filtering, done on technological class disparity (exogenous control) and edge weight ; pareto optimization on (network size, communities)



▲ロト ▲撮 ▶ ▲ 臣 ▶ ▲ 臣 ▶ ● 臣 ● のへで

Classes Examples (2000-2004)

- Memory devices : semiconductor memori devic ; memori cell plural ; memori cell transistor ; layer ferroelectr
- Chemical analysis : time-of-flight mass spectromet ; chromatograph column ; ion trap mass
- Particular steel : martensit ; austenit stainless steel
- Laser : emit laser beam ; vertic caviti surfac ; vcsel
- Sewing : circular knit machin ; stitch ; sew machin ; embroideri
- Lithography : lithograph mask ; project beam radiat ; heat-sensit ; planograph print plate
- Tobacco : cigarett filter ; cigarett pack ; tobacco ; tobacco rod

Results : Patent-level originalities

Patent originality $o_i = 1 - \sum_k p_{ik}^2$

Decrease (resp. increase) of technological (resp. semantic) originality \rightarrow increase of invention originality in time



◆ロト ◆昼 ▶ ◆臣 ▶ ◆臣 ▶ ○ 臣 ○ のへの

Results : Communities overlaps

Overlap between classes as a measure of specialization

Decrease of semantic overlap ; constant technological overlap \rightarrow increase of specialization in time



Discussion

Theoretical and Practical Implications

 \rightarrow Seems to contain relevant information ; full complementarity/relation with other type of information (citations, classification) to be confirmed \rightarrow Semantic analysis methodology can be applied to any network whose nodes have textual description, adding a layer to the network

Further Developments and Potentialities

- \rightarrow Bipartite Network patents/keywords analysis
- \rightarrow Machine Learning to identify innovative patents/emerging research fronts
- \rightarrow Stochastic Block modeling with citation links to statistically validate classifications
- \rightarrow Application for interactive exploration of semantic database

Conclusion

 \rightarrow Explorative insight into patterns of technological innovation ; semi-big-data treatments (previous literature always restrained to subfields)

 \rightarrow Crucial role of interdisciplinarity and integration of qualitative/quantitative approaches

ション ふゆ アメリア メリア しょうくしゃ

Code available at https://github.com/JusteRaimbault/PatentsMining

Data available on Dataverse [Raimbault, 2016], at http://dx.doi.org/10.7910/DVN/BW3ACK



Reserve slides

▲□▶ ▲圖▶ ▲臣▶ ★臣▶ = 臣 = の��

Sensitivity to Window Size

1 year window : more noise in classes



3 year/10 year windows : work in progress

イロト イポト イヨト イヨト

э

Graph clustering : alternatives

Alternative pre-clustering filtering : max degree, edge weight, min/max frequency \rightarrow more difficult sensitivity analysis but more intuitive (*used here for overlaps*)



Temporal Semantic Flows

						portion upper_1999
						bottom surfac_1999
						peripher surfac_1999
						connector electr_2010
						solf 2010
						cather 2010
						sii 2010
portion second_1998			salt_2007		11 2000	bend 2008
			golf_2007	sale 2008	gou_2009	log second 2010
r.sub.2_1998	r.sub.2_1999	formula gener_2005	elkyl 2006 radiu_2007	golf 2008	gear_2009	stator 2010
ester_1998	halid_1999		posit second 2007	Purble 2008	salt 2009	Sink 2010
dioxid_1998	copper_1999		golf 2006 month month 2007	10000	turbin_2009	calibr 1999
perman 1998	chitch 1999	alloy_2005	Turkia 2007	Pole 2008	200m 2009	three-dimension 2008
radio 1998		stator_2005	connector electr 2006	/sink_2008	distal end_2009	Curcator 1999
	africare 1999	tarbin_2005	perman. 2005	refract_2008	stator 2009	devie emit 2010
Terriger_1998	femal 1999		pickup_2007	distal end 2008	wawezuid 2009	apparatu imag process_2010
game_1998	golf 1999	first posit_2005	jet_2005 acoust_2007	inform around 2008	1+4.2009	electrolyt_2010
golf_1998	anna 1999		leg_2006 cather 2007	without 2000		inform record_2010
inflat_1998	game triffe	apparatu form imag. 2005		Childa 2008	authent_2009	financi 2010
apparatu form_1998	cannet_1999		display panel 2006	toner_2008	second signal_2009	rfsd 2010 -
refract_1998	first-portion_1999	refract 2005	graphic interfac user_2007	transact_2008	toner_2009	devic form_2010
	apparatu form_1999	cuber 2005	presup_actory cathod_2007	clock signal 2008	inform record 2009	Criticonal 2008
document 1998	refract_1999	andia 2005	cathod 2005 authent 2007	storag system 2008	conduct second 2009	unterli voltast 2010
OKUNKIK_1770	server 1999	30000 2000	data storag_2007	tulation 2005		Julya tonak 2010
digit signal_1998		at a factor of the state	cach_2006 dock signal_2007	telephon_2008	output voltag_2009	conduct type_2010
cathod_1998	suppli voltag_1999	clock signal 2003	outout voltage 2006	output voltag_2008	cach 2009	data second 2008
suppli voltag_1998	defect_1999	electrod gate 2005	clock signal 2009	layer semiconductor_2008	artibodi 2009	
cach_1998	access memori_1999	non-volatil 2005	crystal devic display liquid 2007	game 2008	1.0000000000000000000000000000000000000	annio_2010
electrod gate_1998	electrod gate_1999	MARYONIAL 2005	distinct_2006	film transistor 2008	base station_2009	nonvolatil_2010
antibodi_1998	attibodi 1999	nucleic_2005	layer settilebildbildbild 2005		game_2009	acoust 2010
			non-yolatil_2006	pharmaceut_2008	gate insul_2009	plural second_2009
			amino 2006 antibodi 2007	cultivar_2008	flash_2009	display plasma_2010
			non-volatil_2007	devic memori semiconductor_2008	cultivar_2009	cultivar_2010

▲□▶ ▲圖▶ ▲ 臣▶ ▲ 臣▶ ― 臣 … のへぐ

References I

Aghion, P. and Howitt, P. (1992). A Model of Growth through Creative Destruction. *Econometrica*, 60(2):323–51.

Bais, S. (2010).

In praise of science: curiosity, understanding, and progress. MIT Press.

i E

Bird, S. (2006).

Nltk: the natural language toolkit.

In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics.

・ロト ・ 理 ト ・ ヨ ト ・ ヨ ト ・ りゅつ

References II

- Bruck, P., Réthy, I., Szente, J., Tobochnik, J., and Érdi, P. (2016). Recognition of Emerging Technology Trends. Class-selective study of citations in the U.S. Patent Citation Network. *ArXiv e-prints*.
- Chavalarias, D. and Cointet, J.-P. (2013).

Phylomemetic patterns in science evolution—the rise and fall of scientific fields.

Plos One, 8(2):e54847.

Gerken, J. M. and Moehrle, M. G. (2012).

A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis.

Scientometrics, 91(3):645-670.

References III

🔋 Griliches, Z. (1998).

Patent statistics as economic indicators: a survey. In *R&D and productivity: the econometric evidence*, pages 287–343. University of Chicago Press.

- Raimbault, Juste; Bergeaud, A. P. Y. (2016).Uspto redbook patent data 1976-2012.
- Tseng, Y.-H., Lin, C.-J., and Lin, Y.-I. (2007). Text mining techniques for patent analysis. Information Processing & Management, 43(5):1216–1247.
- 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).