Nowcasting & Placecasting of Patent Quality around the Globe

– A Temporal Semantic Similarity Approach to Patent Impact Prediction –

May 22, 2019

Daniel S. Hain

Assistant Professor, Ph.D.
Data Science, Computational Social Science, Innovation Studies
IKE, SDS, Aalborg University, DK
E-mail: dsh@business.aau.dk
Github: github.com/daniel-hain

AALBORG UNIVERSITY
DENMARK

IKE
Innovation, Knowledge and Economic dynamics

SDS

Patents & innovation performance: What we know...

- Common measure of inventive/innovative activity & performance [Griliches, 1990].
- Technological & economic significance of patents varies broadly [Basberg, 1987].
- Consequently, the quality rather than number of patents more informative.

Patent Quality: What's been done so far

- Number/composition of IPC assignments [Lerner, 1994].
- Backward [Shane, 2001] & forward [Trajtenberg et al., 1997] citations.
- Lately, first attempts to introduce text (keyword) based indicators [Arts et al., 2017].

What we do instead...

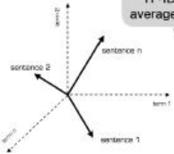
1. Exploit rich textual information with semantic embedding techniques to capture technological signatures.
2. *Relational mapping* of similarity structures between patents (network analysis).
3. Temporal mapping of technological similarity between patents (lead-lag analysis).
4. Prediction of *ex-post* quality indicators with deep learning (*nowcasting*).
5. Provide interactive visualization with high granularity (*placecasting*).

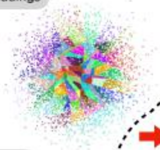
⇒ AKA: What (and where) will be “Europe's Next Super Patent”?

Methods: Pipeline Overview

Our Approach in a Nutshell

Custom word embedding model	Phrase embedding representation	Similarity Matrix construction	Lead-Lag	Prediction
approx. 42m patent titles & abstract	approx. 10m patent titles & abstract transformed TF-IDF weighted average embeddings	**Annoy** Approximate Nearest Neighbours (Oh Yeah) matching	Analysing temporal similarity structures	Auto-encoder anomaly detection approach



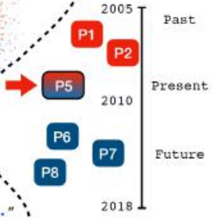


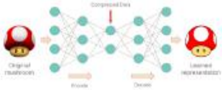
300-dimensional vector representations

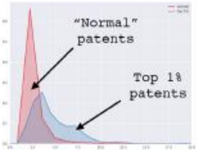
```
target = "Cats are beautiful animals."
test = "Felines are gorgeous creatures."
test2 ="Dolphins are swimming mammals."

print(target.similarity(test))
print(target.similarity(test2))

Out: 0.9522696443950177
Out: 0.7822956256736615
```







+

For starters: Why to look at text?

US 20190080260A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2019/0080260 A1**

Acuna Agost et al. (43) **Pub. Date: Mar. 14, 2019**

(54) **MACHINE LEARNING METHODS AND SYSTEMS FOR PREDICTING ONLINE USER INTERACTIONS**

(71) Applicant: **Amadeus S.A.S.**, Biot (FR)

(72) Inventors: **Rodrigo Acuna Agost**, Golfe Juan (FR); **Alejandro Ricardo Mottini D'Oliveira**, Nice (FR); **David Renaudie**, Valbonne (FR)

(21) Appl. No.: **15/704,320**

(22) Filed: **Sep. 14, 2017**

Publication Classification

(51) **Int. Cl.**

G06N 99/00	(2006.01)
G06N 5/02	(2006.01)
G06Q 30/02	(2006.01)

(52) **U.S. Cl.**

CPC ***G06N 99/005*** (2013.01); ***G06Q 30/0275*** (2013.01); ***G06Q 30/0242*** (2013.01); ***G06N 5/022*** (2013.01)

(57) **ABSTRACT**

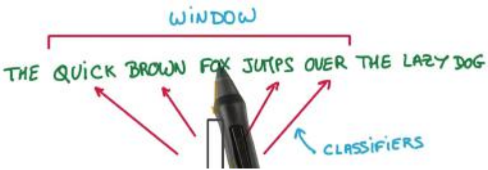
Methods and computing apparatus for retrieving records relating to content placement events and records relating to user interaction events. A set of enriched training feature vectors is computed from raw feature values, and used with interaction event tags to train a machine learning model. Information is received relating to an online content placement slot and information is received relating to a user to whom content within the online content placement slot will be displayed. An enriched estimation feature vector is computed based upon a content item selected for placement within the online content placement slot, the information relating to the user, and the information relating to the online content placement slot. A machine learning model is executed to determine an estimate of likelihood of the user interacting with the selected content item, based upon the enriched estimation feature vector.

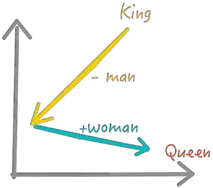
Numbers vs. Text

Methods: 1 - Patent-to-Vector

Creating and validating

- Simple intuition: Counting keyword appearance → But what about synonyms, antonyms, analogies etc.?
- We instead use word embedding: Natural-language-processing technique that represents words as high dimensional vectors according to the context in which they tend to appear.





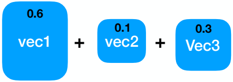
a) Learns Analogy

b) Similar Words have same angles

Patent embedding & Technological distance

- We use a TF-IDF weighted average word embedding representations.
- Result: 300-dimensional patent embedding vector ⇒ technological signature.
- Embeddings subject to vector algebra ⇒ Distance between two patent embeddings = technological distance.

TF-IDF - weighted - embeddings



electrical_connector characterised by a **receptacle** containing a plurality of **female_contacts** having **redundant_contact** portions and **wiping_capabilities** with respect to **male_pins**

```
target = "Cats are beautiful animals."
test = "Felines are gorgeous creatures."
test2 ="Dolphins are swimming mammals."

print(target.similarity(test))
print(target.similarity(test2))

Out: 0.9522696443950177
Out: 0.7822956256736615
```

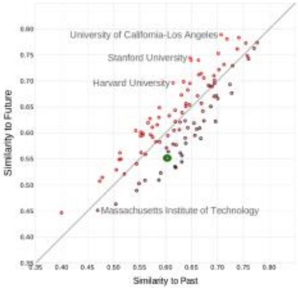
First validation exercise

- Patent embeddings predict IPC3 Classification with 83% multi-class prediction accuracy. (out-of-sample).
- Patents which cite each others, are from the same applicant, inventor, patent family etc. have significantly lower technological distance.

Methods: 3 - Similarity-to-Quality

Temporal similarity: Intuition

- Semantic similarity independent of time.
- Temporal similarity distribution can be exploited
- Inspired by the lead-lag approach of Ramage et al. [2010]; Shi et al. [2010].



Temporal similarity: Types

Similarity to past: Novelty

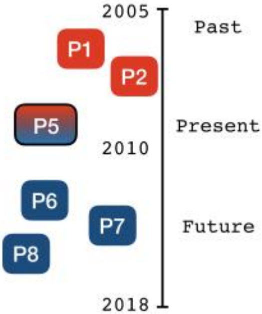
- Exploitation of existing knowledge.
- High values might indicate backward orientation, low values indicate novelty.

Similarity to present: Popularity

- "Riding the wave", indicates activity in a trending area.

Similarity to future: Impact

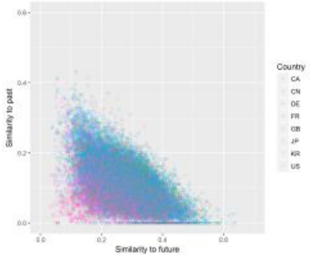
- Shaping the agenda, indicator of future impact.
- Also: Indicator of "Window-of-Opportunity", high growth technological field.

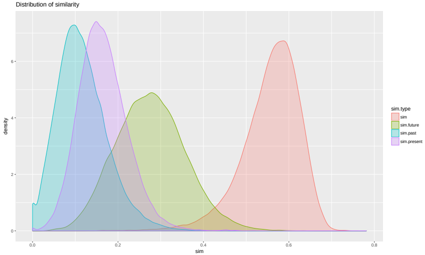


Use-Case: Electromobility Technologies

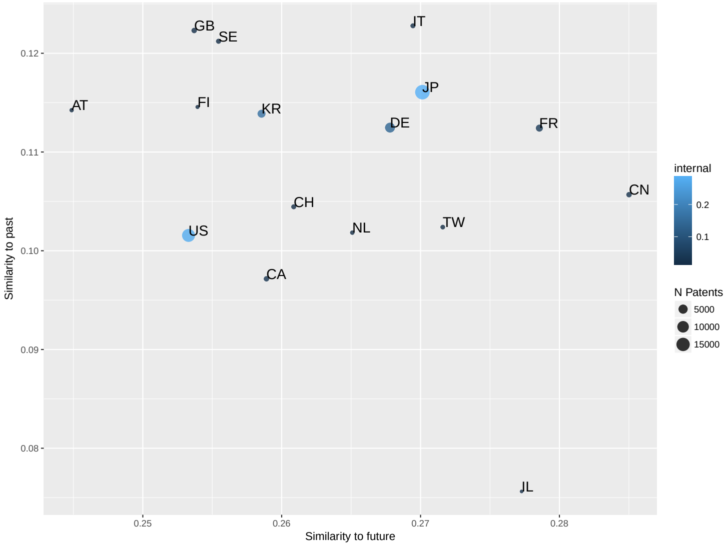
- Electromobility related patents based on expert-adviced IPC class selection.
- Further, all patents cited by “seed” also included (ca. 13k).

IPC class	Level	Description
B60L 11/00	0	Electric propulsion
B60L 11/02	1	using engine-driven generators
B60L 11/04	2	using dc generators and motors
B60L 11/06	2	using ac generators and dc motors
B60L 11/08	2	using ac generators and motors
B60L 11/10	2	using dc generators and ac motors
B60L 11/12	2	with additional electric power supply
B60L 11/14	2	with provision for direct propulsion
B60L 11/16	1	using power stored mechanically
B60L 11/18	1	using power from primary cells





Aggregate Picture: Where is Novelty and Impact created?



Use-Case: Electromobility Technologies

(Less) Aggregate Picture

- Different levels of aggregation deliver different insights.
- Enables nuanced and dis aggregated analysis where, by whom, and when novelty and impact is produced.

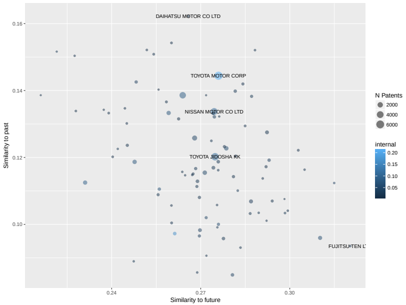


Figure: Firm Level

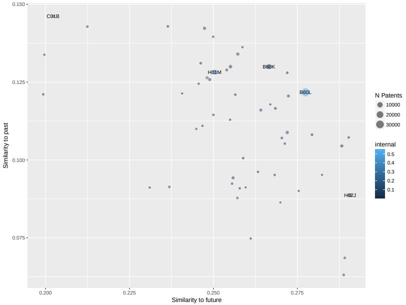
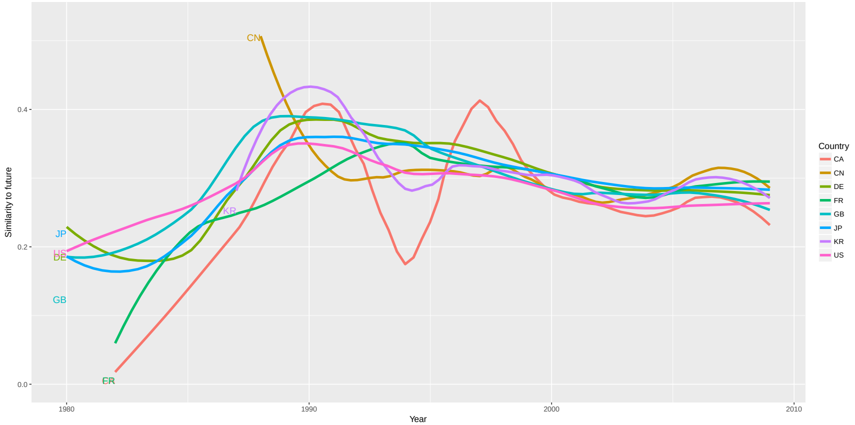


Figure: Technology Level

Use-Case: Electromobility Technologies

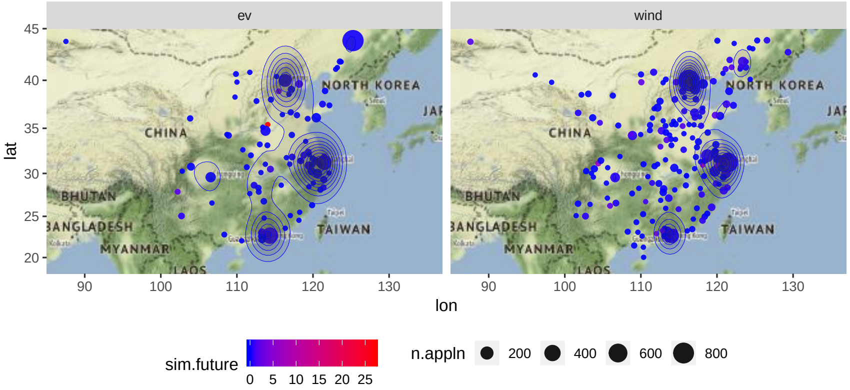
Dynamics over time: Capturing Technology Life-Cycles

- Reveals global technology life-cycles and “windows of opportunity”.
- Highlights different entry strategies and catching-up dynamics by latecomers.



Geography of Inventive Activity

- Providing granular insights in quality of inventive activity across regions.
- Facilitating smart specialization policies.



Methods: 4 - Nowcasting

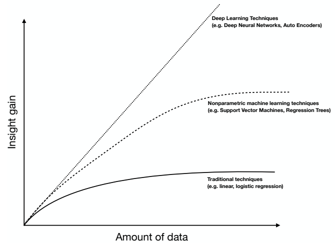
Who would care so far?

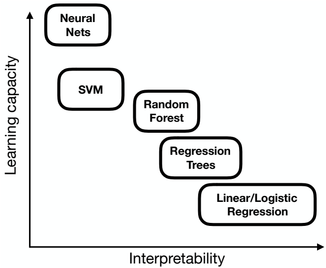
- Academia: Interesting for historical and theoretical analysis.
- Policy: Not really. The state 5 years ago not so helpful for actions today...

$\Rightarrow$ Need for nowcasting (prediction)

A Note on Predictive Modelling

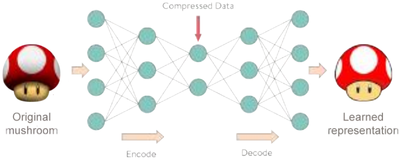
- Econometric modelling: Given a set of carefully selected variables of interest, how to identify causal effects on an outcome of interest?.
- Predictive modelling (aka machine learning): Given all available information, what is the best possible prediction of an outcome of interest ($\hat{y}$ rather than $\hat{\beta}$).





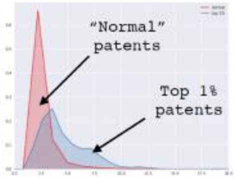
(Ex-Ante) Predicting Patent Quality

- Forecasting of patent quality measures (novelty, impact, citations etc.) with modern ML rather easy.
- More interesting: Significant rare events → Europes Next Superpatent?
- Task: Identifying breakthrough patents (top-1%) [Ahuja and Lampert, 2001]



Rare event (anomaly) prediction

- Deep Neural Autoencoder: self-supervised model that aims at reproduction of its inputs
- Train on “boring normality” (non-breakthrough patents)
- High reproduction-error when facing anomal inputs → “something is wrong”.
- Results so far: Very nice AUC (>0.8), high accuracy (0.87) and sensitivity (0.81) out of sample.



Methods: 5 - Placecasting: The Global Patent Exlorer

The power of visualization and data-narratives

- So far so good, but after all we just produce numbers.
- Complex data pipelines are of little use without producing a narrative.

⇒ We went a step further, and provide interactive visualizations of geolocations, granular geographical networks of knowledge flows, ad further indicators.[1]

GLOBAL PATENT EXPLORER
EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE
OECD

www.gpxp.org

[1] As a goodie, many traditional patent measures [cf. Squicciarini et al., 2013].

Some central questions remain...

1. How to understand and trust predictions?

2. How to evaluate and improve predictions?

Modern predictive models (eg., deep learning) are incredibly complex and nuanced.
Result often:

What's next? 1: Explain Models

Challenge 1: Explain model prediction

- Global model mechanics often to complicated for human annotation.
- Local model decision criteria can be approximated.
- One approach: "Local Interpretable Model-Agnostic Explanations" (LIME) [Ribeiro et al., 2016].
- Enables questioning and correcting model decisions.
- Can be used to increase fairness of models, and our trust in them.

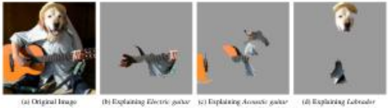


Figure 4: Explaining an image classification prediction made by Google's Inception network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

Explaining predictions with LIME


DJ U.S. Coal Exports Plunge 31% in April

Most of the trouble can be traced to slower demand growth in China, the world's biggest coal consumer, and increased production as multiyear capital projects come online at miners including BHP Billiton, the world's No. 1 producer of metallurgical, or coking, coal. "You have too many tons chasing a smaller market."

What's next? 2: Improve Models

Challenge 2: Improved and more nuanced predictions

- While our text-based method for technological similarity replicates commonly used approaches well, believes in superior performance are yet mainly technical.
- Even more prevalent when moving beyond similarity towards functional relationship mapping (eg., complements, substitutes, enabler, platforms).
- Ground truth still Human Intelligence.
- Computer Science approach to such hard problems: Produce a large annotated benchmark dataset → community challenges to puhs state-of-the-art.
- Example Computer Vision: IMAGENET - Enormeous (ca 1.2M train, 100k test) human annotated (1k classes) image dataset, annual competition. In 2010 unthinkable task → 2015: Solved (96.4% classification accuracy).
- Our (first) approach: Establish benchmark dataset of patent-similarity, joint effort of many cooperating POs.

Some (Optimistic) Take-Away's

1. Natural language processing (particularly: embedding) techniques are powerful tools to map and understand relationships in large bodies of text data.
2. Use-Cases are by far not limited to patent descriptions (eg., Policy reports, media debates, H2020 project descriptions, reports and meeting summaries).
3. Predictive modeling (ML) techniques have high potential to improve timely, granular, and precise forecasts of outcomes of interest (nowcasting & placecasting), and rare events (eg., breakthrough patents, unicorn start-ups).

Some more (Critical) Take-Away's

1. ML models crucially depend on data (amount & details), and corresponding labels.
2. ML model mechanics tend to be opaque, but there are promising developments to change that.
3. Need of modern means of outcome-communication which are interactive (facilitates own insight generation), engaging (create data narratives), and selective (more not always better).
4. Collaborative effort needed to establish benchmarks, scrutinize and validate.
5. Open method and data workflows and requirements crucial for progress.
6. Cross-disciplinary efforts of Computer & Social Science necessary.

Fin.

References I

To whom it may concern

Ahuja, G. and Lampert, C. (2001). Entrepreneurship in the large corporation: A longitudinal study of how established firms create breakthrough inventions. *Strategic management journal*, 22(6-7):521–543.

Basberg, B. L. (1987). Patents and the measurement of technological change: a survey of the literature. *Research policy*, 16(2-4):131–141.

Griliches, Z. (1990). Patent statistics as economic indicators: A survey. *Journal of Economic Literature*, 28(4):1661–1707.

Lerner, J. (1994). The importance of patent scope: an empirical analysis. *The RAND Journal of Economics*, pages 319–333.

Ramage, D., Manning, C. D., and McFarland, D. A. (2010). Which universities lead and lag? toward university rankings based on scholarly output. In *Proc. of NIPS Workshop on Computational Social Science and the Wisdom of the Crowds*. Citeseer.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM.

Shane, S. (2001). Technological opportunities and new firm creation. *Management science*, 47(2):205–220.

Shi, X., Nallapati, R., Leskovec, J., McFarland, D., and Jurafsky, D. (2010). Who leads whom: Topical lead-lag analysis across corpora. In *NIPS Workshop*.

Squicciarini, M., Dernis, H., and C, C. (2013). Measuring patent quality: Indicators of technological and economic value.

Trajtenberg, M., Henderson, R., and Jaffe, A. (1997). University versus corporate patents: A window on the basicness of invention. *Economics of Innovation and new technology*, 5(1):19–50.