***A new instrument for technology monitoring****: novelty in patents measured* ***by semantic patent analysis****. Jan M.* ***Gerken*** *and Martin G. Moehrle*

Can’t access because it’s in book, or for paid journals.

- Motivation: Information overflow (lots of patents)

- Goal: Asses the novelty of a patent automatically

- Current Methods: lack depth as they fail to exploit functional relationships

- Questions: How can we even measure patent novelty? (kind of abstract, **recombinant** vs. **pioneering**)

How can semantic structures inform us of a patent’s qualities?

***The Double Edged Sword*** *of Recombination in Breakthrough Innovation. Sarah* ***Kaplan****, Keyvan* ***Vakili****.*

Two schools of thought form “the double edged sword”:

**1.) ‘Tension’ view**

-Fight against myopia, induced by deep singular domain knowledge.

-Solution is broad and diverse knowledge recombination for economically valuable novel patents.

**2.) ‘Foundational’ view**

- Only diving deeper produces breakthroughs. Deep/foundational domain knowledge is needed.

In summary: recombination is either seen as promoting or detracting from innovation.

Economic value of inventions are not always realized. (must be recognized and successfully marketed)

One rough way of measuring novelty is simply citations, which correlates noisily with economic value.

Shifts in ideas drive shifts in language, thus innovation is reflected in language change. Patents that introduce such language are, by induction, innovative.

\*idea: for each patent store a list of all unique words in that document. Speed improvements for later when looking for presence of particular word.

\*idea: can compare frequency of words in patents to the norm. Identify words used significantly more normal (suggest as defining vocabulary, or tags, or features to associate with other patents)

\*idea: create multidimensional idea space, look at breakthroughs in time, identify ‘star forming regions’ where a lot of new patents are being born, i.e. hot dense regions.

Topic modelling, to identify topic source patents and themes over large numbers of texts.

Latent Dirichlet Allocation (LDA) uses co-occurrence of words (Bayesian) outputs

1.) topics, and their constituent words weighted in a vector

2.) documents, with their constituent words weighted in a vector

Use the Stanford Topic Modeling Toolbox

Should probably constrain the number of topics to keep it interpretable

\*idea: Treat 5-year forward citations as forward returns and do a regression over possible factors. Tech space, novelty, (estimated position along foundational/tensional spectrum)

got to look out for ‘patent families’ with identical abstracts

\*idea: Pick a topic, look at top N earliest patents (first 12 months) with that topic weighted above some threshold. Can thusly identify topic-source patents.

\*concern: Isn’t topic modelling more descriptive than predictive? Given a new patent, won’t it just try to fit it into an existing space? What would be better is if we could track and transform the space over time and measure the influence of a new paper on that space.

\*idea: To identify Topic-Source Patents, for patents in a topic, plot(x=time, y=forward citations). Look at patents with abnormally high forward citations for their time. May not be the first one that really captures the ‘essence’ of the topic.

How do you measure technological distance?

**Prior art** is all information publicly available at a given date relevant to a patent’s claims of originality.

\*Startup Idea: at the end of this, could potentially have a search engine for evaluating patents. Could enter in an existing patent name, and it would find the patents closest to it in the space and display a degree of similarity (for evaluating novelty of a patent). Also helpful for suggesting related works (as a normal search engine might). Could alternately look by topics, and it would suggest canonical readings in that arena to get you off the ground (could look by most cited, originating, and then by most important recently to get you to the current point).

\*idea: would be awesome to be able to suggest to it topics and it would return a basket of patents most likely to, when combined, produce a novel, or economically valuable patent.

Combination familiarity (i.e. how much has this combination been used) is important. Could mean that there’s good technology and support for it enabling researchers to use it, could also mean that its potential has already been reached.

\*idea: [Following the Genetic Model of Recombination] look at parent and children patents. What characteristics in the parents and their environments produced strong offspring patents? How disparate were their fields, how new were their fields, how new was the field of their child? Look at their citation rate (though it would have to be the rate at the time of child patent submission), their importance or novelty ranking at that time etc. Create a model of these factors and identify current parents that when combined might yield strong offspring.

Possible packages/tools:

gensim

lda – python LDA package

Stanford NLP Toolkit

Gephi for visualization?

Vowpal\_Wabbit

Alex Mannish’s online LDA code

Matt Hoffman’s online LDA code

Number of topics vs goodness of fit. Choose a number that gives good fit for its size.

Potential Project Goals:

Descriptive-

Map tech space

Describe/measure novelty

Identify factors leading to high forward citation rates

Identify factors leading to high economic value

Comment on best methods for obtaining novelty(recombinant or foundational)

Predictive-

Given new patent, where does it fall in tech space? Vectorize it via ‘folding in’ http://stats.stackexchange.com/questions/9315/topic-prediction-using-latent-dirichlet-allocation

Given new patent, what’s likelihood of novelty?

Given new patent, what’s the likelihood of high citation rate?

Given new patent, what’s the likelihood of high economic value?

Given current tech space, what two (or more) patents when combined are likely to yield novel child patents.

Expectation maximization – spark

Beyond vanilla LDA

Author Topic Models

N-gram LDA models

ML-Lib

Issues with access to AWS. 90-100pounds a month

Other alternative is to use Hadoop or spark

Harmonized firm names to patent ID’s

Harmonized names

Assigning applicant ID’s to beuro van-dike entries. (improving match rate)

**Amadeus & patstat**

Tracking scientist interest and scientist publications

(absorbing knowledge spillover)

+ [tracking movement of a firm] single dimensional change, and magnitude.

‘Clean space’ vs ‘dirty space’ – electric vehicles are clean (make sure it’s clearly defined)

what is clean and what is dirty – international patent classification system (carbon lock in)

technological propensity to patent

septoral level spend, revenue created by patents, modeling growth rate.

Cost vs. benefit of patent – profit ratios – equilibrium profits, how it look in the long term.

Propensity to patent at a sector level, national level

Recombination of knowledge

Function of the rate of innovation in the field

In related fields

Growth models of topics

MLib = 100x mahoot

LDA source algo is on their github

Moving forward:

OECD patent statistics manual

Clarity on supervisors.

[X] Rare event modelling

[X] Causal models. More explaining why or a mechanism

Darren MIT growth model - acemoglu

Recombinant & citations

National level times series of patenting in a series

Page rank for citation data.

Technological diffusion in regards to disease models. Clusters of firms, economic shocks due to regulations. Resiliency to shocks.

Who’s grading it?

Beuro-van-dike database to get structure of firms.

List of harmonized names from OECD

Assigning patents to companies and their subsidiaries alias etc.

Use naïve bayes to go beyond simple name matching.

---

Importance/economic value can also be measured by litigation rate as a proxy

Semantic Patent Analysis – Propensity to Patent Across Topics

Use online LDA to update topic space and inform a model that yields the likelihood of a given entity to patent in various fields.

- Obtain unique entities

- Create a tech space

- what factors? (size of company, patents in this area so far, patents in other areas, patents in area in last 12-months, how active related authors are in space etc.)

- Regression using each of these facts.

- Essentially 12 month momentum across sectors.

\*would be interesting to look at this patent momentum overlayed with product release.

**What Data is Available?**

OECD Harmonized Applicant Names

Get names of entities from:

Get Patents from:

Collate patents to registered entities. Assign a unique author.

Leaning towards Implementing bei’s dynamic topic modelling code for patents from PATSTAT

Author topic would be cool, but I’m not sure there’s a version out there fast enough for our corpus.

Might take a stab at author matching if it will help. (naïve bayes? LSTM? Deep net in keras? whatevs)

Vanilla LDA

<https://pypi.python.org/pypi/lda>

<https://radimrehurek.com/gensim/models/ldamodel.html>

Possible variations on vanilla LDA

**Author Topic** Models (what authors are most likely to go with a topic or words)

(might not work for large data sets…slow + memory footprint)

<http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm#Example_Scripts>

<https://github.com/arongdari/python-topic-model>

it seems the issue really will be speed. What has he already tried? Saw his gensim thread. Think coding author topic modelling from scratch and making it faster would be biting off more than I can chew.

**Online LDA** – Stream documents, update model (good for memory footprint)

<http://mlwave.com/tutorial-online-lda-with-vowpal-wabbit/>

<https://www.cs.princeton.edu/~blei/topicmodeling.html>

**Online Hierarchical Dirichilet Process HDP** (would determine the number of topics itself!)

<https://github.com/Blei-Lab/online-hdp>

<https://radimrehurek.com/gensim/models/hdpmodel.html>

**Dynamic Topic Models** (evolution of topics over time)

**Dynamic Influence Models** (measure document influence)

<https://radimrehurek.com/gensim/models/dtmmodel.html>

<https://code.google.com/archive/p/princeton-statistical-learning/downloads>

Get a feel for required ML content.

Project Pitch So Far

External Supervisors:

* Paulo Agnolucci. BSEER, usually energy and environment econometrics.
* Christopher Grainger. PhD. Student. Energy econometrics as well. So he’s been interested in ‘clean’ vs. ‘dirty’ technology in terms of environmental impact etc.

Idea:

* Semantically analyze patents from the EPO’s PATSTAT database using Latent Dirichilet Allocation for Dynamic Topic Modelling and Dynamic Influence Modelling. Essentially we wish to map patents to a tech space and track how that tech space evolves with the addition of new patents, and subsequently asses how influential a patent is, how pivotal it was in a particular region of tech space’s development.

Data:

* Chris Grainger has 100GB database from PATSTAT. Contains patent applications with classification, patent family, priorities, citations, publications, applicants, inventors, INPADOC worldwide legal status.
* OECD [Organisation for Economic Collaboration and Discovery] also has harmonized patent applicant names, covers EPO patents. (in case I want to do author topic modelling.)

Method:

* Follow the work of David Blei of Princeton fairly closely, good topic modelling guy.
* Dynamic Topic Models: <https://radimrehurek.com/gensim/models/dtmmodel.html>
* Dynamic Influence Models: <https://code.google.com/archive/p/princeton-statistical-learning/downloads>
* Gensim (popular Python topic modelling library) offers a python wrapper for Blei’s code.
* Alternately, Author Topic Models (tracking what companies are most likely to patent in), or Hierarchical LDA (automatically finds the number of topics.)

Obstacles:

* Computation time/power
* Getting data into systems…

Past Papers and Research:  
 - Author disambiguation

-Novelty prediction

-Economic value prediction

-knowledge spillover (looking at citation networks, pagerank)

Pior Work the Team has done:

* Vanilla LDA on patent abstracts, used AWS and Hadoop. Think they leveraged some public scala code for distributed version. Additionally vowpal/wabbit version of LDA for speed. Already have well defined metrics for distance between patents.

Logistics:

* Need to register online by Friday March 4th.
* Not sure what else? Could ask him what comes up later… like paper length etc.

Questions for Professor Shawe-Taylor:

* What are things to watch out for during the process of working on the project? What has tripped students up in the past?
* What are some of the great projects you’ve seen in the past and what made them memorable?
* Since The project is with an external team, who’s research goals aren’t necessarily developing ML techniques, and are more descriptively focused. What advice do you have on strattling the divide between internal and external project requirements? Trying to keep everyone happy, myself included.
* General Advice. (maybe tools, packages, tech stacks etc.)
* Tips for what to do if I need help or get stuck.

Timeline of most immediate “to-do’s”:

* Registration by March 4th
* Get PatStat data from Chris Grainger
* Look into AWS for computation (how to get data into it?...)

Evaluation: (need to think a bit more about this…)

Perplexity/likelihood on holdout set

Model convergence rate

Try various numbers of topics and evaluate coherence and interpretability of topics.

Pick a paper, reproduce results first few weeks.

Interesting bits in the middle.

Write up time.

Tensor method,

Get small subset of data and work from there.

Police movements using LDA,

Improving topic coherence by bias connectivity between words with co-ocurrence from a prior dataset like Wikipedia.

Use patent families as labels for classification based off of.

Coherence topic modelling. (google for papers.)

March 3rd Shawe leaves.

By family first, then by 1GB,