Overlay of science and technology patterns with unsupervised learning: Case of thermal management system

Tech Mining Conference 2015

Samira Ranaei[1], Arho Suominen[2], Mika Lohtander[1] and Tuomo Kassi[1]

Lappeenranta University of Technology[1], Finland

VTT research center[2], Finland

Background

Open your mind. LUT.
Lappeenranta University of Technology

Methods for measuring science and technology interaction: (Mayer 2000, Narin et al 1995, 1997)

1. Industrial publication
2. University patenting
3. Non-patent literature

Disagreement over the reliability of patent citation analysis to assess science and technology relationship:

- Patent citations only reference novel arts or limited output and cannot reveal the complete knowledge transfer flows of patent innovation . (Jaffe and Trajtenberg , 2002, Criscuolo & Verspagen, 2008).

- Patents Citation patterns vary significantly by firm, industry, and even country characteristics (Alcacer et al. 2009)

- Firms' citation choices can be strategic (Hegde and Sampat 2009), Meaning of citation behaviour (Bornmann & Daniel, 2008, Martyn, 1964).
- NPL citations not only scientific citations: Mixed set of other type of publications. conference proceedings, books, and many other non-scientific sources such as disclosure bulletins, abstract services, and so forth.

Research objectives

Open your mind. LUT.
Lappeenranta University of Technology

Purpose: Study the linkage between science and technology with content-based approach

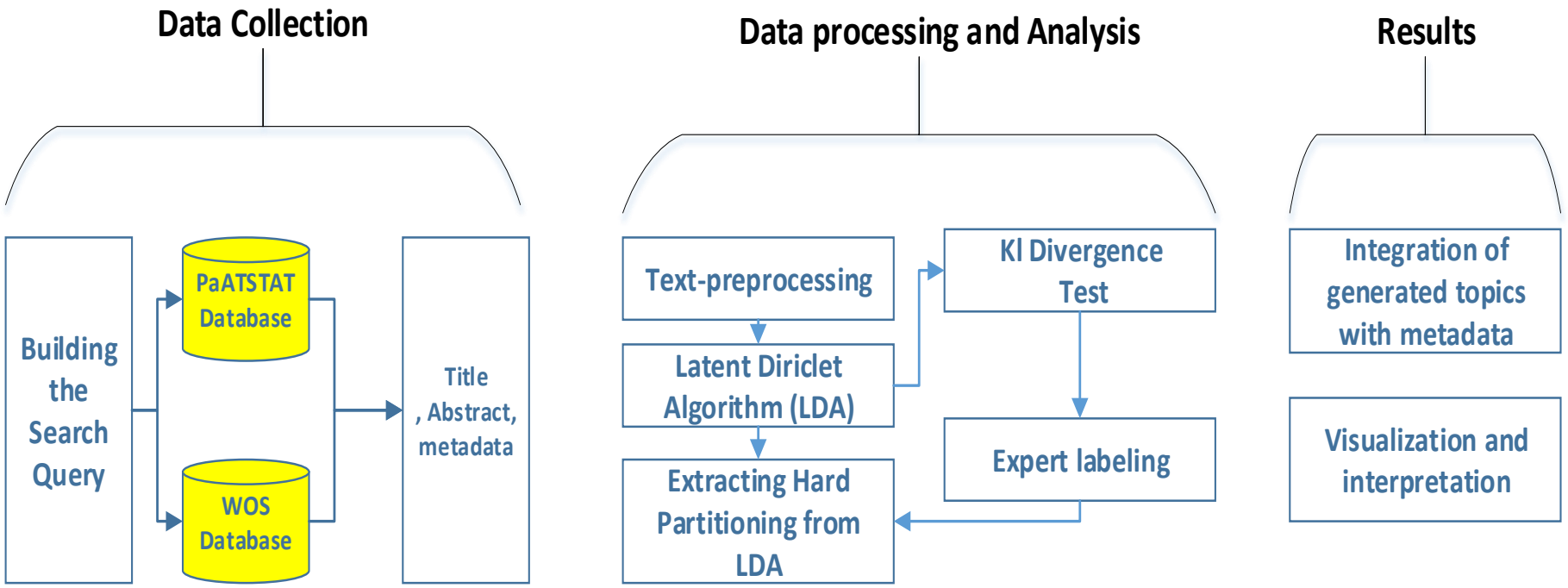
Research Question:

- How the content based indicator can be used to assess the science and technology relationship ?
- How feasible it is to apply unsupervised algorithms on textual data to measure science and technology linkage

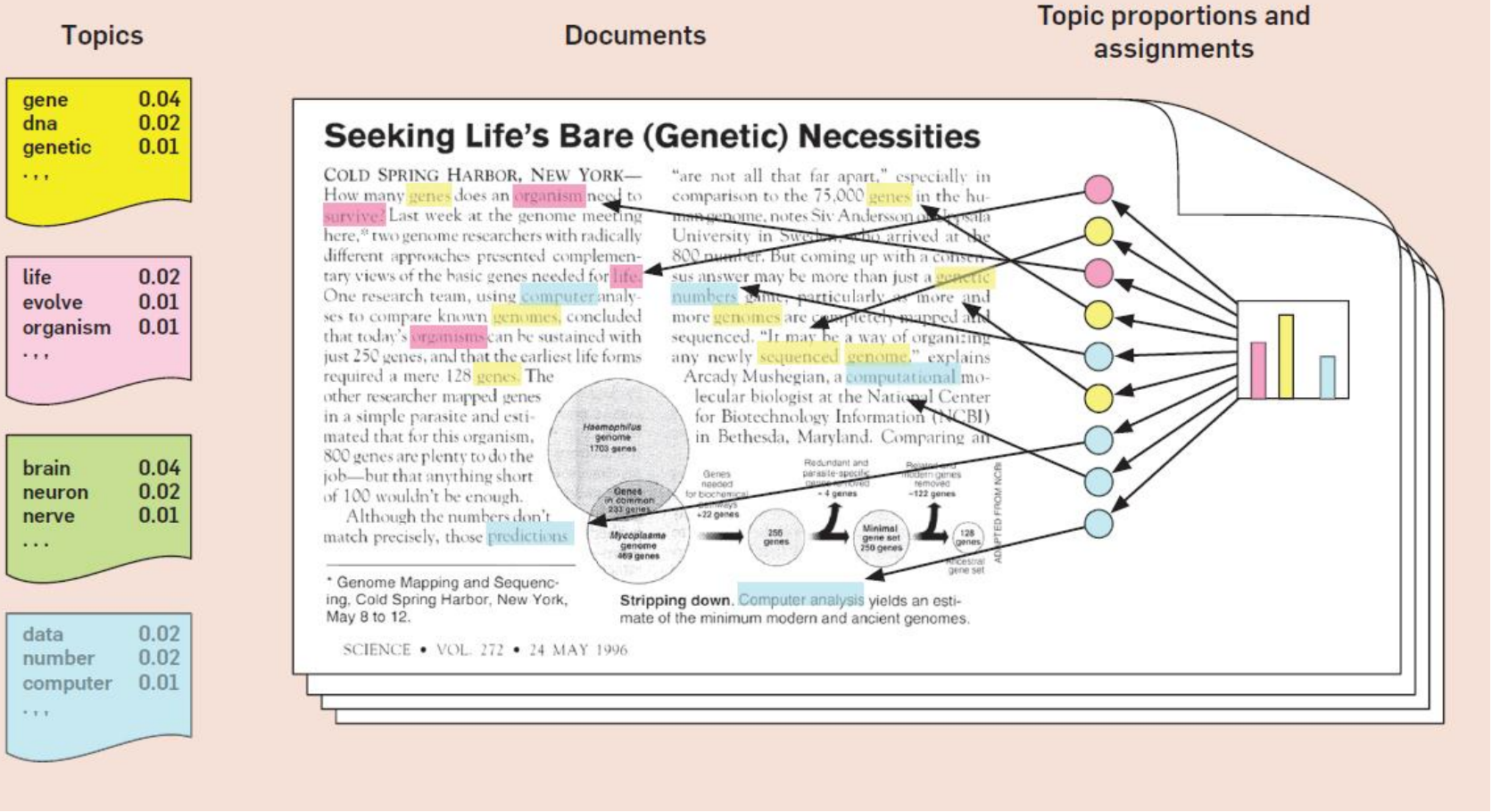
Case : Thermal management systems- cooling methods used in electronic devices

Data collection and methodology

- Data: 610 patents, 625 papers from 1980-2013
- Gensim Python library designed for topic modeling



Latent Dirichlet Algorithm



Challenges …

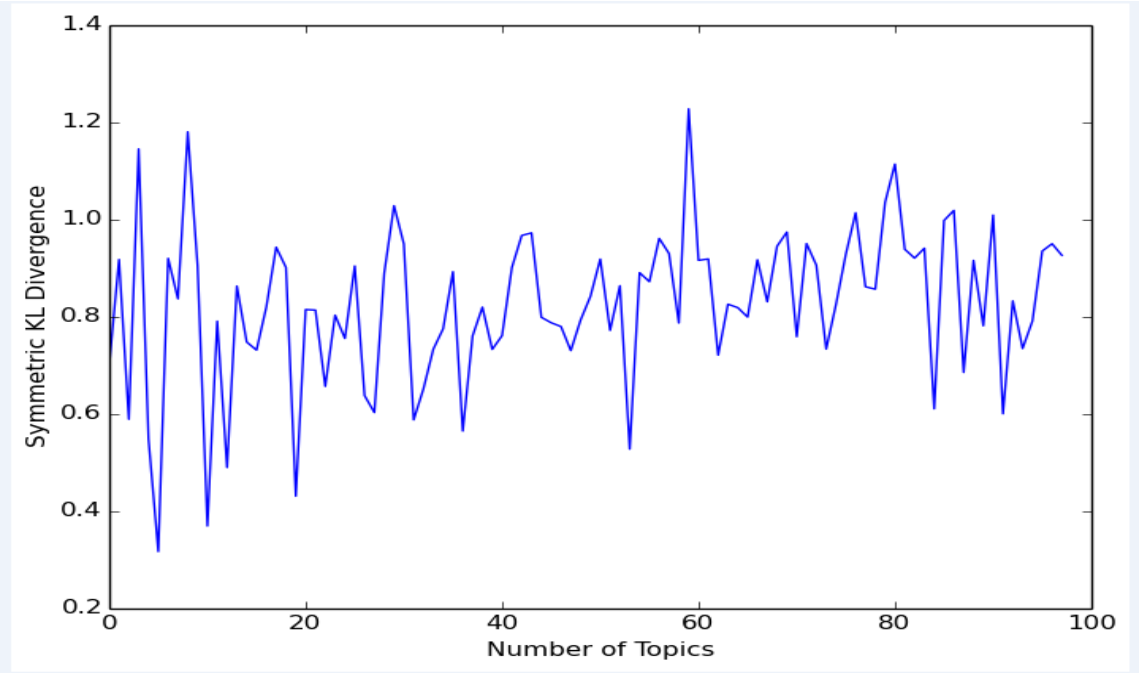
- How many topics should be learned?
- How many learned topics are useful?

- How to measure the similarity between patents and publications?
- How different would be the result if we cluster patents and publication in one data set or separately?
- Mixed corpus
- Separated corpus

Optimum number of topics

- Kullback -Leibler divergence measure
- Less than 5

- Expert review
- LDA generated 20 topics
- Expert reviews the top keywords of each cluster and proposed 5 topics



Source : Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N.(2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In Advances in Knowledge Discovery and Data Mining (pp. 391-402).

LDA on mixed data set

Open your mind. LUT.
Lappeenranta University of Technology

1. Title and abstract of 1235 docs
2. Pre-processing = tokenization, downcasing, stopword removal, removing words appearing once
3. Symmetric Dirichlet priors α=0.5, β=0.1, 1000 iteration
4. Number of topics = 5
5. Hard Clustering

6. Manual screening of the results
-10 docs from each cluster
-Only 4 documents were clasified by experts differently

Doc Type	Topic a	Topic b	Topic c	Topic d	Topic e	Top topics
Patents1	0,0259	0,7479	0,0262	0,0258	0,1743	b
Publication857	0,0204	0,0209	0,0205	0,0206	0,9177	e
Publication862	0,0279	0,8899	0,0275	0,0273	0,0274	b
Patents3	0,0309	0,0308	0,8765	0,0311	0,0307	c
Patents4	0,0338	0,0338	0,8649	0,0337	0,0337	c
Publication858	0,0203	0,0200	0,9196	0,0201	0,0201	c
Patents6	0,0251	0,9003	0,0249	0,0248	0,0249	b
Patents7	0,0323	0,0322	0,8716	0,0320	0,0320	c
Publication859	0,8300	0,1094	0,0201	0,0201	0,0203	a
Publication860	0,0235	0,9062	0,0235	0,0234	0,0234	b

Number of topics	Patents	Publication	Grand Total
Topic a	162	135	**297**
Topic b	151	95	**246**
Topic c	164	125	**289**
Topic d	79	125	**204**
Topic e	54	145	**199**
Grand Total	**610**	**625**	**1235**

Top keywords in each cluster

Open your mind. LUT.
Lappeenranta University of Technology

Topic a =Performance and efficiency	Topic b= cooling Methods variation	Topic c=componets and parts	Topic d=application of methods	Topic e= arrangementd of components
velocity	Air-cooling Two phase cooling	nozzle	Skin	duct
temperature	Spray cooling	water	laser	transfer
droplets	jet	condenser	treatment	heat
impact	Jet impingement	channel	patient	flux
evaluation	Dry cooling	duct	treat	stage
mass	laser	chamber	device	surface
experiment	Spary	plate	human	plate
	Liquid characteristics	pump	epidermal	steel
		transfer	pulse	particle

LDA on separated dataset

Open your mind. LUT.
Lappeenranta University of Technology

1. Title and abstract of 610 patents, 625 publications
2. Pre-processing = tokenization, downcasing, stopword removal, removing words appearing once
3. Symmetric Dirichlet priors α=0.5, β=0.1, 1000 iteration
4. Number of topics = 5 for each set
5. Hard Clustering
6. Cosine similarity measure

$$\text{Cosine similarity} = \frac{T_1 \cdot T_2}{\|T_1\| \cdot \|T_2\|} = \frac{\sum_{i=1}^{n} T_{1i} \cdot T_{2i}}{\sqrt{\sum_{i=1}^{n} T_{1i}^2} \cdot \sqrt{\sum_{i=1}^{n} T_{2i}^2}}$$

		Patent Topics				
		0	1	2	3	4
Paper Topics	0	0,8180	0,8460	0,8841	**0,9821**	0,9911
	1	0,7847	0,6030	**0,9879**	**0,9730**	0,8788
	2	0,6841	0,8060	0,9779	0,7700	**0,9102**
	3	0,5211	**0,9916**	0,8060	0,7073	0,8102
	4	**0,9911**	0,7567	0,9017	0,8516	0,8522

Source: Lee, H. et al., 2015. Coherence analysis of research and education using topic modeling. *Scientometrics*, 102(2), pp.1119–1137.

Discussion

- Patents and papers are different but share similar features (topics). They can be compared based on their similarities
- Both approaches show that semantic relationship exists between patents and publications of cooling systems. Therefore it is possible to measure S&T relationship using unsupervised algorithm.

Limitations: In the case of cooling methods, we learnt that patent and publications are utilizing almost similar vocabularies. The cosine similarity signals this issue. Therefore, more cases should be studied in terms of generalization.

Future work:

- Study the relationship of documents in each cluster.
- Combination of meta data with generated topics

Open your mind. LUT.
Lappeenranta University of Technology

Thank you for your attention

ABSTRACT. The analysis of citation networks of patents or papers has been extensively used to define the knowledge structure or linkage between science and technology(S&T). However, citation approach is limited dues to the time lag, data coverage to cited or citing documents, and may under-represent the possible knowledge flow between S&T data sources. In this paper, it is assumed that the linguistic pattern of patents and publications illustrate their topical overlaps and would spot the potential growing fields in research or practice. The novelty of our approach is the utilization of topic modeling and expert opinion, in order to cluster patents and articles based on their content rather than citations. Applicability and accuracy of our method is tested on a corpus of documents in field of thermal management system.