Evaluating the effect of time and journal quality to topics: Structural topic models of scientific publications

Recent literature has extensively used Topic models to analyze scientific and patent document. Several studies have shown the applicability of topic models to classify textual data sourced from scientific publications (Suominen & Toivanen 2015) or patents (Suominen, Toivanen, & Seppänen 2016). Yau et al. (2014) utilized a corpus of scientific publications based on expert opinion to evaluate a number of topic models, and showing the algorithms capabilities of distinguishing between classes. Topic models, such as Latent Dirichlet Allocation, are challenged by the fact that the model does not incorporate any metadata, which could potentially have an impact to how latent themes should be interpreted. Recent works have developed multiple targeted topic models, such as the author topic model (Rosen-Zvi et al. 2004) or dynamic topic model (Blei & Lafferty 2006), that allow for controlling one metadata attribute. Structural Topic Model (Roberts et al 2013) adds to the more targeted topic models by discovering topics but also estimate their relationship to multiple metadata variables.

Taking advantage of this possibility, this study focuses on the impact of time and journal quality on topics, evaluating 1) if topics are impacted by the emergence of new research themes or 2) by different type of research being presented in higher quality journals. The study is conducted using a 75 479 ISI WoS records on fuel cell technology ranging from 1990 to 2014. The data has been retrieved using the terms “fuel cell” or “fuel cells” in the topic field of WoS search.

The data was preprocessed using Python programming language. First, the full data was reduced to the fields AB, TI and PY. In addition, one field was calculated using the ISBN number of each journal and the grading system for publications in the Finnish rating system for publications. Each publication was matched to the Finnish rating system to see if the publication is either a “basic or unknown” research outlet or a “leading or top” publication outlet. This resulted in a dichotomous variable, JUFO, being added for each publication, which classified if the document was published in a top-ranking outlet or not. The AB and TI fields where concatenated and further preprocessed by part-of-speech tagging and keeping tags 'NN','JJ', and 'NNP'. Fuel cell types, such as Solid Oxide Fuel Cell, where transformed to single token abbreviations. Punctuations and number where removed. Finally, each of the concatenated and pre-processed text was expected to have over 500 characters. This limited the final sample to 68 637 documents. After preprocessing, the data was analyzed using R package “stm”. Through trial and error, we selected to produce 30 topics.

Topic 1: water, transport, liquid, gdl, diffus	Topic 16: electrochem, current, electrod, imped, polar
Topic 2: chemic, materi, metal, fuel, high	Topic 17: adsorpt, sulfur, adsorb, poison, remov
Topic 3: hydrogen, reform, product, steam, gas	Topic 18: membran, electrolyt, fuel, polym, catalyst
Topic 4: anod, nickel, nio, niysz, microstructur	Topic 19: conduct, oxygen, phase, temperatur, electr
Topic 5: sofc, solid, fuel, oxid, electrolyt	Topic 20: technolog, fuel, industri, commerci, process
Topic 6: flow, channel, gas, numer, mass	Topic 21: degrad, durabl, test, loss, cycl
Topic 7: reduct, oxygen, activ, reaction, orr	Topic 22: control, power, system, voltag, convert
Topic 8: process, deposit, film, powder, porous	Topic 23: catalyst, catalyt, activ, oxid, reaction
Topic 9: fuel, system, stack, design, power	Topic 24: pemfc, pem, fuel, temperatur, oper
Topic 10: oxid, activ, electrocatalyt, catalyst, electrooxid	Topic 25: carbon, electrod, electrochem, graphit, surfac
Topic 11: proton, acid, conduct, poli, sulfon	Topic 26: mfc, mfcs, power, cathod, electr
Topic 12: energi, hydrogen, electr, storag, fuel	Topic 27: electron, xray, spectroscopi, structur, microscopi
Topic 13: cathod, electrolyt, composit, temperatur, fuel	Topic 28: power, system, effici, gas, generat
Topic 14: surfac, oxid, alloy, metal, corros	Topic 29: membran, composit, conduct, nafion, exchang
Topic 15: methanol, dmfc, dmfcs, concentr, crossov	Topic 30: high, fuel, low, complex, combin

Figure 1 The highest probability words in topics.

The most probable words of each topics are seen in Figure 1. The topics clearly highlight separate technologies, such as Solid Oxide Fuel Cell technology in Topic 5 or PEM fuel cells in Topic 24. The topics also distinguish with significant technological issues relating to fuel cells, such as durability and degradation in Topic 21. For each topic, we calculated the effect of the dichotomous variable of journal quality and a stepwise variable for time starting from 1990. Seen below, for Topic 29. As with the majority of the topics, the variable time did not have a significant impact the topic. However, the variable JUFO, relating to journal quality did have a statistically significant impact. This impact is illustrated across the topics in Figure 3, which highlights how topic are impacted by journal quality.

Coefficients:					
Estimate Std. Error t value Pr(>|t|)					
(Intercept)	0.0198147	0.0096495	2.053	0.0400	*
JUFO2	0.0057500	0.0008371	6.869	6.52e-12	***
s(PY)1	-0.0210755	0.0182803	-1.153	0.2490	
s(PY)2	0.0034581	0.0083451	0.414	0.6786	
s(PY)3	0.0077393	0.0105535	0.733	0.4634	
s(PY)4	0.0105397	0.0096010	1.098	0.2723	
s(PY)5	0.0088766	0.0104084	0.853	0.3938	
s(PY)6	0.0108956	0.0107499	1.014	0.3108	
s(PY)7	0.0136161	0.0146565	0.929	0.3529	
s(PY)8	-0.0068462	0.0405122	-0.169	0.8658	
s(PY)9	0.0649827	0.0885411	0.734	0.4630	
s(PY)10	0.0170800	0.0099676	1.714	0.0866	.

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					

Figure 2 Effect of metadata variables to Topic 29.

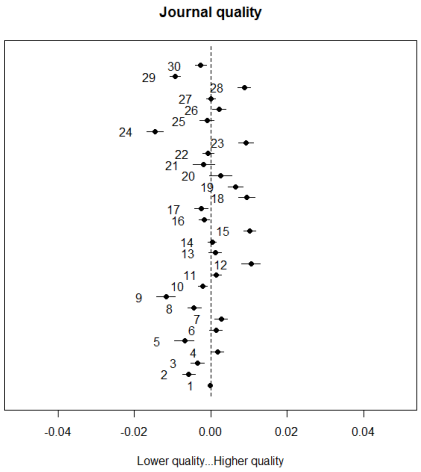

Figure 3 Journal quality across topics.

References

Blei, D. M., & Lafferty, J. D. (2006, June). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning* (pp. 113-120). ACM.

Roberts, M. E., Stewart, B. M., Tingley, D., & Airoldi, E. M. (2013, December). The structural topic model and applied social science. In *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation* (pp. 1-20).

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004, July). The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence* (pp. 487-494). AUAI Press.

Suominen, A., & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology*, *67*(10), 2464-2476.

Suominen, A., Toivanen, H., & Seppänen, M. (2017). Firms' knowledge profiles: Mapping patent data with unsupervised learning. *Technological Forecasting and Social Change*, *115*, 131-142.

Yau, C. K., Porter, A., Newman, N., & Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics*, *100*(3), 767-786.