HUMAN ASSIGNED VS. MACHINE CREATED: LINKS BETWEEN PATENTS AND SCHOLARLY PUBLICATIONS

Introduction

To measure knowledge flows between scholarly literature and patents, studies have used a several approaches, such as keyword search based count of science publications and patenting in a technology with an expectation of linearity of innovation (Watts & Porter, 1997), patents citing publications (Meyer, 2000) or vice versa (Glänzel & Meyer, 2003) or author-inventor co-occurrences (Meyer, 2006). Patent citing scientific publications is a measure of linkages between two codified sets of knowledge (Narin, Hamilton, & Olivastro, 1997). This linkage measure has often been extended to serve as a proxy of for the "science-dependence" or "science-base" of a technology, although this has been critiqued as an over simplification (Meyer, 2000). For an analysis of interplay between science and technology, citations from patents to publications offer a narrow window. This is most significantly due to the fact that roughly one third of patents have a human assigned non-patent references (Callaert, Looy, Verbeek, & Debackere, 2006) of which only a part are citations to scientific publications.

Another avenue to classify patents and publications would be to rely on machine learning, specifically unsupervised learning. For example, LDA, a unsupervised learning method, draws out latent patterns from semantic text. LDA has been shown to be able to "…extract surprisingly interpretable and useful structure without any explicit "understanding" of the language by computer" (Blei, Ng, Jordan, & Lafferty, 2003). LDA is able to classify scientific publications with significant precision (Yau, Porter, Newman, & Suominen, 2014). On a qualitative level, LDA has produced practical mapping of science (Suominen & Toivanen, 2015a) and patent landscapes (Suominen & Toivanen, 2015b). Due to difference between the language used in patents and publications, merging the two corpuses is challenging, but the approach offers a possibility to complement the existing sparse non-patent citations.

In this study, we analyse the relationship between publication and patents by looking at the intersection of human assigned and machine learned linkages between science and patents. We use a macro level approach focusing on the whole science publication and patents from one country. This specific vantage point to science and patent linkages looks if we can overlay knowledge patterns at a regional/country level and if this can be achieved through human given labelling or machine created classifications.

Data & Method

Machine learning approach

We use a publication dataset for years 1995-2011 obtained from the Web of Science (ISI-WOS) where at least one of the publication authors has Finnish affiliation. The publication data was obtained via two methods. First, the data until 2010 was delivered as a tagged XML data with the full article level information as recorded in ISI-WOS. Data delivery was done by Thomson Reuters in August 2012. This data was updated via web access to the Web of Science by which means data for 2011 was added to the first data set. Patent data is USPTO grants data, where at least one inventor resides in Finland. The data is obtained via a locally hosted PATSTAT database. The final merged dataset consists of 185 931 records. This merged dataset was pre-processed before running LDA using a Python script.

The Python script used removes punctuations, stopwords, tokens consisting of only numbers or including numbers and any special characters. In addition, the script uses Levenshtein distance to search for typographical errors and a procedure to identify bigrams. Data pre-processing also omits records that, after pre-processing, do not have sufficient data for analysis. After pre-processing the final dataset consist of 185 831 records, 16 393 patent grant documents and 169 438 publications documents. The LDA analysis was done using the implementation of variational EM for LDA by Blei et al. (2003). While LDA has been shown to produce a good estimation on the latent pattern in a given corpus, it requires the user to input the number of classes, usually unknown to the researcher. To estimate the real-world performance of different topic numbers we used a trial-and-error approach, finally setting the number of topics to 150.

The probability distributions for words and documents were analysed using Python, R and Excel. A Python code was used to merge metadata, namely publication year, document unique identifiers and document type, with document probabilities. The merged dataset was aggregated to descriptive data tabulating data into topic level. Python package NetworkX was used to transform the probability matrix to a two-mode network dataset, and then imported to Gephi software for further analysis. Using the OpenOrd layout algorithm (Martin, Brown, & Klavans, 2011) and the Modularity algorithm (Blondel, Guillaume, & Lambiotte, 2008) embedded in Gephi, we analysed proximities among documents and topics. The OpenOrd layout algorithm was used to create a visual layout. The Modularity algorithm was used to analyse if there are clusters among the documents and topics. The word probability distribution for each topic was visualized using wordclouds through the R package wordcloud.

Non-patent citations

The non-patent references for each patent was searched from the European Patent Office's ("EPO") Open Patent Services ("OPS"). The data was accessed using the patent number of the 16 393 patents in the machine learning sample and retrieving all non-patent references. The code was created using Python programming. The retrieved data contain numerous types of references, some of which are full citations to literature other less informative abbreviations. Matching science publications and patents used a fuzzy matching strategy focusing on the citations author name and title, if available. The approach first identified all title matches between non-patent references and publications. After this, patents where matched to publications titles through a fuzzy

match using the Python difflib library. Finally, author and inventor names where used to identify any remaining matches in the datasets.

Results

The sum of probabilities for patents (n=150) average 107,05 (s = 124,72) and for publications (n=150) average 1084,30 (s = 753,86) per topic. The topic distribution of both patents and publications are near equally unconcentrated. Calculating the Herfindahl-Hirschman Index (HHI), patents have a HHI index of 11,01 over topics and publications 11,69. To analyse if we find topics with significant disparity in patent or publication difference, we calculated the absolute difference of sum of probability shares Abs(diff). The differences between sum of patent and publication probabilities (N=150) average 0,46 % (s=0,632), but there is significant variation from the average with a minimum of 0,003 % and maximum of 4,43 %. As expected, patents and publications are thematically differently focused. Largest in patents, Topics 66 and 68, refer to the telecommunication and electronics. This is large due to companies such as Nokia, which has rank highest in new patent application in Finland on several years. Publication topics are more challenging to interpret. Suominen & Toivanen (2015a) showed in their study that by using LDA Finnish science is classifies into an large interdisciplinary cluster of social sciences and management and several smaller focused medical research areas. This structure is also visible here. Topic 81, shared by both, is clearly a derivate of the natural language used to write both patents and publications and relates more to (research) methods

Table 1. Topics with highest sum of probabilities in publications and patents.

Patents			**Publications**		
Topic	**%**	**Top terms**	**Topic**	**%**	**Top terms**
Topic 66	5,4	system, communication, station	Topic 57	3,5	plurality, management, transmit
Topic 68	4,3	device, method, electronic	Topic 47	2,2	ci, mortality, risk
Topic 81	3,6	data, method, measurement	Topic 81	2,1	data, method, measurement

To validate if we can identify relevant science technology pathways, we highlight two university derived patent. Table 2 shows, in two different domains, medical research and energy technology, how a university based invention and publications by the same scholars has been classified by the algorithm. The selected patents do not have a non-patent citation to the selected publication. At a hard classification level, where we take the largest probability value for a document and assign the document to only that class, we see that the patent and corresponding publication are similarly classified. In energy technology, the highest probability topic is Topic 28, where the top terms are metal, electrode, ion and fuel. In medical technology, the highest probability topic is Topic 58, where top terms are detection, antibodie and serum.

Table 2. Topic distributions of university patents and publications by the inventors on the same topic.

Patent	**Topics %**	**Mod class**	**Publication**	**Topic %**	**Mod class**
Halme, A., Korhola, M., Appelqvist, A., Suomela, J., & Zhang, X. C. (2008).*U.S. Patent No. 7,384,701*. Washington, DC: U.S. Patent and Trademark Office.	Topic 99 (11%) Topic 28 (31%) Topic 15 (19%) Topic 62 (15%)	6	Zhang, X. C., Ranta, A., & Halme, A. (2006). Direct methanol biocatalytic fuel cell—Considerations of restraints on electron transfer. *Biosensors and Bioelectronics, 21*(11)	Topic 99 (7%) Topic 28 (16%) Topic 83 (7%) Topic 1 (7%)	0
Eriksson, S., & Pettersson, K. (2009). *U.S. Patent No. 7,638,292*. Washington, DC: U.S. Patent and Trademark Office.	Topic 58 (29%) Topic 67 (18%) Topic 14 (16%)	1	Eriksson, S., Halenius, H., Pulkki, K., Hellman, J., & Pettersson, K. (2005). Negative interference in cardiac troponin I immunoassays by circulating troponin autoantibodies. Clinical chemistry, 51(5)	Topic 58 (25%) Topic 67 (11%) Topic 147 (7%)	1

However, looking at the soft classification of the documents, we can identify difference in how the documents probability distribution in seen. Both of the examples have one additional shared lower probability class in the top classified topics, but also several that are different. Finally, if we look for how the patents where classifies using the Modularity algorithm, that takes into account the whole probability distribution in the dimension reduction, we see that the medical technology patent publication pair is classifies to the same class, but this is not the case for the energy technology pair.

Our API search retrieved 36217 NPR, these being from a total of 3136 patents. This means that for roughly 19 percent of patents we have the option of finding a science publication. On average a patent that has non-patent references has 12 different references (s=23, N=3136), this implies that patents that have non-patent references have strong connections to citations other than patents. From these patents, only a marginal portions had links within the Finnish science, making further analysis challenging.

DISCUSSION

The central objective of this paper has been to demonstrate the differences between human assigned non-patent literature citations and machine learned classifications when using a merged corpus of patents and publications and analysing linkages at a national level. Machine learning is able to pick-up links, even if not made explicit by the inventor, simultaneously creating significant noise that needs to be handled. Non-patent references are inadequate at an national level to create a knowledge overlap, specifically due to the low number of patents that contain references making the analysis challenging. Future research should focus on developing a mixed methods approach in creating a holistic view of patent science linkages.

References

Blei, D. M., Ng, A. Y., Jordan, M. I., & Lafferty, J. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 2003.

Blondel, V., Guillaume, J., & Lambiotte, R. (2008). Fast unfolding of communities in large networks. *Journal of Statistical*.

Callaert, J., Looy, B. Van, Verbeek, A., & Debackere, K. (2006). Traces of prior art: An analysis of non-patent references found in patent documents. *Scientometrics*.

Glänzel, W., & Meyer, M. (2003). Patents cited in the scientific literature: An exploratory study of'reverse'citation relations. *Scientometrics*.

Martin, S., Brown, W., & Klavans, R. (2011). OpenOrd: an open-source toolbox for large graph layout. *IS&T/SPIE*.

Meyer, M. (2000). Does science push technology? Patents citing scientific literature. *Research Policy*.

Meyer, M. (2006). Are patenting scientists the better scholars?: An exploratory comparison of inventor-authors with their non-inventing peers in nano-science and technology. *Research Policy*.

Narin, F., Hamilton, K., & Olivastro, D. (1997). The increasing linkage between US technology and public science. *Research Policy*.

Suominen, A., & Toivanen, H. (2015a). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology*, n/a–n/a.

Suominen, A., & Toivanen, H. (2015b). Unsupervised learning based patent landscapes using full-text patent data. In *2015 Portland International Conference on Management of Engineering and Technology (PICMET)* (pp. 2195–2203). IEEE.

Watts, R., & Porter, A. (1997). Innovation forecasting. *Technological Forecasting and Social Change*.

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