Map of Technology: Topic modeling full-text patent data

Arho Suominen* and Hannes Toivanen

*arho.suominen@vtt.fi
VTT Technical Research Centre of Finland,
P.O.Box 1000, 02044 Espoo, Finland

A central challenge for the mapping patents is the creation of valid and accurate coordinates. Our study discusses the choice of the origin of coordinates in order to make a map of technology, and, in particular, demonstrates the advantages of unsupervised learning-assigned coordinates over those created by human reasoning.

A number of studies have focused on creating patent maps (Yoon et al. 2002; Lee et al. 2009; Kim et al. 2008). Many of the patent map studies rely on traditional bibliometrics, such as Karki (1997) using citation analysis to form a patent mapping to public policy analysis and Daim et al. (2006) shows what can be regarded as the de facto bibliometric technology forecasting example. Recently the focus has turned to the use big data and data mining, specifically text mining, in patent mapping. In 2007, Tseng, Lin and Lin (2007) illustrated text mining techniques for patent analysis. Tseng, Wang, Lin and Lin (2007) focused on creating machine produced summarizations and mapping of patents. A data mining approach has also been used in technology roadmapping (Yoon and Phaal 2013). As computational methods in the machine-learning field are becoming more available and stable, studies have moved towards using stable processes and focusing on applying the methods to for example management of technology. For example, unsupervised learning methods, such as Latent Dirichlet Allocation, are relatively stable methods relatively easily accessed by science scholars outside the computer domain. In this study, we show the capabilities of unsupervised learning with a single-node computer in learning the thematic areas of all full-text patent documents published by the USPTO in 2014 (N=374,704). We further discuss the key challenges in running the study, interpreting the outcome and further developments.

Background

Unsupervised learning produces an outcome based on an input while not receiving any feedback from the environment. Unsupervised learning differs from supervised or reinforced learning by its reliance on a formal framework that enables the algorithm to find patterns. The majority of unsupervised methods rely on a probabilistic model of the input data. An unsupervised learning method estimates the model that represents the probability distribution for an input either based on previous inputs or independently. Topic models are unsupervised learning methods and Latent Dirichlet Allocation (LDA) is one topic model that draws out latent patterns from text. In 2007, Blei and Lafferty (2007) showed the usability of topic models in modeling the structure of semantic text. In presenting the methodology Blei and Lafferty (2007) noted that topic models " ...can extract surprisingly interpretable and useful structure without any explicit "understanding" of the language by computer". The basic idea behind the model is that each document in a corpus is a random mixture over latent topics, and each latent topic is characterized by a distribution over words. In the LDA model, each document is a mixture of a number of topics based on the words attributable to each of the topics. LDA allows us to uncover these latent probability distributions based on the sematic text used in the document, thus classifying the documents based on the latent patterns within them. For a detailed explanation on the algorithm refer to for example Blei and Lafferty (2009) and for an evaluation analyzing scientific publications refer to Yau et al. (2013).
We analyze USPTO published patent data from the year 2014 (N=374,704). The data consists of all patents published in 2014 and the analysis uses the full-text description as source data for unsupervised learning. Prior to analysis the abstract texts were pre-processed with a Python script. The Python script removes stopwords and punctuations. Terms that occur only once in the whole data were also removed at

this stage. After all of the before-mentioned terms were removed, the text was tokenized and each token was transformed to a corresponding number, to further reduce the complexity of the data. As, LDA requires a fixed number of topics, we employed the KL divergence based evaluation of the natural number of topics (Arun et al. 2010). The qualitative evaluation of KL divergence values and multiple runs of the algorithm, we produced 200 topics. The topics were visualized using wordclouds.

Results

Our results show, how we are able to draw out meaningful latent patterns from a large text corpus with a single-node computer. Our setup classified the 374,704 full-text documents in a practical time, creating a model that can be used to infer the classification of new documents. Our results question the usefulness of human-given labels, such as IPC classes, as classifiers as unsupervised learning produces a practical division of technology that is not reliant on a historical human generated classification scheme.
The key challenge of LDA based analysis is estimating the number of topics built and pre-processing needed. The method proposed by for example Arun et al. (2010) takes significant computational time and produces limited value for the analysis. For pre-processing, Yau et al. (2013) suggested limiting the pre-processing of data prior to analysis. Our results however show that there is an added value of taking on a more aggressive approach. More research is however needed. Clearly, methodological development in machine-learning methods is in a point where algorithms are available "of the shelf". Our abilities of visualizing matrices of size 374,704 times 200 is however more challenging and there is a clear need to turn focus on creating actionable results for users.

References

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.

Blei, D. M. and J. D. Lafferty, (2007). "A correlated topic model of science," The Annals of Applied Statistics, vol. 1, no. 1, pp. 17–35.

Blei, D. M. and J. D. Lafferty, (2009). Text Mining: Classification, Clustering, and Applications, ch. Topic Models, pp. 71–94, 10th ed. Taylor and Francis.

Daim, T., G. Rueda, H. Martin, and P. Gerdsri, (2006). "Forecasting emerging technologies: Use of bibliometrics and patent analysis," Technological Forecasting & Social Change, vol. 73, no. 8, pp. 981–1012.

Karki, M. M. S. (1997). "Patent citation analysis: A policy analysis tool." World Patent Information 19.4, pp. 269-272.

Kim, Y. G., J. H. Suh, and S. C. Park, (2008). "Visualization of patent analysis for emerging technology," Expert Systems with Applications, vol. 34, no. 3, pp. 1804–1812.

Lee, S., B. Yoon, and Y. Park, (2009). "An approach to discovering new technology opportunities: Keyword-based patent map approach," Technovation, vol. 29, no. 6, pp. 481–497.

Tseng, Y.-H., C.-J. Lin, and Y.-I. Lin, (2007). "Text mining techniques for patent analysis," Information Processing & Management, vol. 43, no. 5, pp. 1216–1247.

Tseng, Y.-H. ,Y.-M. Wang, Y.-I. Lin, C.-J. Lin, and D.-W. Juang, (2007). "Patent surrogate extraction and evaluation in the context of patent mapping," Journal of Information Science.

Yau, C.-K., A. Porter, N. Newman, and A. Suominen, (2013). “Clustering scientific documents with topic modeling,” Scientometrics, pp. 1–20.

Yoon B. and R. Phaal, (2013). “Structuring technological information for technology roadmapping: data mining approach,” Technology Analysis & Strategic Management, vol. 25, no. 9, pp. 1119–1137.

Yoon, B.-U., C.-B. Yoon, and Y.-T. Park, (2002). “On the development and application of a self–organizing feature map–based patent map,” R&D Management, vol. 32, no. 4, pp. 291–300.