

Drafting Alignment of Business Method Patent Documents: A Text-mining Approach

Speaker

Davit Khachatryan, PhD

Assistant Professor of Statistics and Analytics

Babson College; Division of Mathematics and Science

Email: dkhachatryan@babson.edu; Phone: (781) 239 – 6475

Abstract

Patent review and examination, in other words the patent “prosecution” is a complicated process. Patent examiners operate under stringent deadlines often faced with voluminous backlogs of pending patent documents. In fact, according to one study, on average, examiners spent about eighteen hours for the review of an application. Further, there exists documented evidence that United States Patent and Trademark Office (USPTO) tends to “over-grant” patents at times when it is under resource constraints. As noted by Lichtman and Lemley (2007), “it is hardly a surprise that the PTO makes mistakes during the initial process of patent review, granting patents that, on the merits, should never have been issued.” At times such patents lose their validity to exclude others as a result of legal rulings, after being granted.

But entrepreneurs have to be careful in generating and wisely spending the cash required for survival. Because a positive cash flow is one of the crucial factors helping startups stay in business, entrepreneurs can’t afford to waste cash in unnecessary legal fees, in case their patents get legally challenged and invalidated after being granted. It has been shown that failure of the patent specification to prove that the patentee invented what is being claimed in the patent document is one reason that can lead to legal proceedings and ultimately invalidation. In other words, patent specification should be closely “aligned” with what is being stated as the invention(s) as part of patent claims. Having ways to measure and subsequently paying close attention to accurate alignment of claims and specification is in entrepreneurs’ best interest before sending the patent application to the patenting office.

In this work, we define a measure of *drafting alignment* that quantifies the similarity between claims and specification of patent documents. Given that both the claims and the specification are text which usually are big by any standard measure of data, we use methods from text mining to accomplish our objective. To measure the alignment, we first pre-process the data by removing from each of the two texts (i.e. from the set of claims and the specification) the punctuation, numbers and “stop words,” followed by further steps to clean the data, such as by “stemming” each word in each constituent text

document to reduce it to its stem. Subsequently, we convert each resulting pre-processed document to a vector, in line with the vector-space model (Salton, et al. 1975). Further, the document-term matrix is constructed corresponding to the two pre-processed text documents. The definition and calculation of drafting alignment is then carried out using the obtained document-term matrix and appropriate term frequency-inverse document frequency (tf-idf) weighting scheme (Jones 1972; Baumer, et al. 2017).

After defining the metric, we build a statistical (regression) model to investigate the relationship between various patent-specific factors and the defined drafting alignment. The objective of such an investigation is to determine factors of which entrepreneurs or inventors should be cognizant, as potential “drivers” of drafting alignment. Some of the factors considered in this investigation are: presence of a representative during patenting, number of independent claims, average length of independent claims, the ratio of the length of the claims section to that of the specification section, etc. Investigation of these and similar factors and the resulting implications carry the promise of helping the entrepreneur draft applications that represent a better alignment between claims and specification and as such less likely to be ultimately invalidated, if granted.

Commercial database repositories provide data on patent documents. To apply the newly defined measure of drafting alignment and empirically investigate its properties, we use data on non-provisional patent documents applied to USPTO which have USPC class 705 as their main class. Note, that these data were extracted from the Thomson Reuters’ *Thomson Innovation* (currently Clarivate Analytics) patent database.

References

- Baumer, B. S., Kaplan, D. T., & Horton, N. J. (2017). *Modern Data Science with R*. CRC Press.
- Spark-Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(5), 111-121.
- Lichtman, D., & Lemley, M. A. (2007). Rethinking Patent Law's Presumption of Validity. *Stanford Law Review*, 45-72.
- Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613-620.

Biography

Dr. Davit Khachatryan is an applied statistician with research interests in analyzing intellectual property data to study the formation and diffusion of knowledge in emerging industries. Davit's current and former research has produced publications in academic, peer-reviewed journals such as *Scientometrics*, *IEEE Transactions on Engineering Management*, *Journal of the Royal Statistical Society (Series C)*, *The American Statistician*, and *Journal of Statistics Education*. Prior to joining Babson College, Davit was a Senior Associate at the National Economics and Statistics practice of PricewaterhouseCoopers (PwC). In the latter role he consulted in the area of predictive modeling and advanced data analytics, helping clients from financial, healthcare, and government sectors with building automatic predictive models and enhancing business intelligence solutions. Davit has earned his B.S. in Applied Mathematics and Informatics from Yerevan State University, M.S. in Statistics and Ph.D. in Management Science from the University of Massachusetts, Amherst.