Comparing website measures on R&D with patenting indicators

Submitted to the 2019 Global Tech Mining Conference as a Power Talk

Empirical research on emerging technologies and innovative firms traditionally relies on any number of established data sources, including primary qualitative collections and surveys, as well as secondary databases of patents, publishing, and business metrics. However, firm survey response rates continue to decline; existing business databases are pricey and may limit access to certain researchers and institutions; and not all firms patent or publish. New data sources, such as websites and social media, are readily available at modest to no cost, but little is known about how to operationalize valid and reliable measures. The purpose of this power talk is twofold: (1) to operationalize R&D keyword-based measures derived from firm websites, and (2) to correlate those measures to common indicators of patenting activity.

The sample consists of 1,146 patenting firms whose websites were successfully crawled. Each firm is in one or more of the following industries: synthetic biology, nanotechnology, or renewable energy. R&D keywords are operationalized as matching any of the following terms: "lab" OR "laboratory" OR "research" OR "development" OR "R&D" OR "researcher" OR "scientist". Patenting intensity, on the other hand, is measured as the number of patents granted by the US Patent and Trademark Office since the firm's founding, or since 1976. Even when both variables are logged, they are not bivariate normally distributed, suggesting that a non-parametric correlation test is appropriate. The results show that there is a weak correlation between R&D mentions on websites and patenting intensity: Kendall's tau is 0.16 and is significant at α = .01.

This preliminary result suggests that websites could provide a new set of innovation indicators with respect to firm R&D activities. However, first there is a need to create additional measures for comparison purposes to better understand what R&D mentions on websites may signify vis-à-vis known indicators from the patenting literature. As part of the power talk, we will review a limited set of other measures drawing on keyword-derived search terms in Gok et al. (2013) and Li et al. (2018) and derive simple correlations, and possible multivariate regression outputs, to tease out how patenting firms make use of their websites to convey R&D and innovation-related themes. This work builds on past research in the sense that it starts with a frame list of patenting firms, rather than firms identified through business databases. The findings therefore should be generalizeable to the population of patenting firms.

References cited

Gök, A., Waterworth, A., and Shapira, P., "Use of web mining in studying innovation," *Scientometrics*, 2015, 1, 653-671. http://dx.doi.org/10.1007/s11192-014-1434-0

Li, Y., Arora, S., Youtie, J., & Shapira, P. (2018). Using web mining to explore Triple Helix influences on growth in small and mid-size firms. *Technovation*, 76, 3-14.