The Patent Similarity Dataset – Enabling More Nuanced Data-Driven Innovation Research

Ryan Whalen

Overview:
This project provides a patent dataset that both facilitates new innovation metrics, while offering improvements on previously established measures. By providing over 6 million 300-dimension patent doc2vec vectors, the dataset enables researchers to more-accurately address issues of patent content. Meanwhile, the provision of almost 700 million similarity scores provides useful context for a wide variety of data-driven innovation research. The dataset is provided in both JSON, and where-possible CSV formats, and maps easily onto other publicly-available USPTO data.

Background:
Much empirical innovation scholarship relies on data about patents to infer relationships between them (e.g. citations) or to estimate their substantive content (e.g. classifications like the CPC). There is a large body of work that relies on these metadata signals, using them to measure things like research impact, patent value, the existence of patent thickets, and patenting trends. However, these metadata-based measures provide rough approximations at best. For instance, relying on classification data to infer content lumps together many inventions with varying-levels of similarity. Likewise, relying on patent citations as binary signals of relationship or impact glosses over much heterogeneity between different types of citations. This project engages with the long-tradition of data-driven patent research by sharing a new data set that allows for more nuanced and accurate innovation metrics, while also enabling types of analyses not previously possible using existing datasets.

Creation of the dataset:
The dataset creation entailed the following steps:
• Using the full text of patents granted since 1976 until the end of 2018 (n = 6,183,713) to compute a doc2vec model of patent semantic space.
• Extracting the 300-dimension vectors for each patent published between 1976 and the end of 2018.
• Calculating the pairwise vector similarity for all cited/citing patent pairs (n = 74,619,582); and
• Identifying the 100 most-similar patents for each patent, and calculating their similarities.
The result is a dataset containing 6,183,713 300-dimension vectors and 692,990,882 pairwise patent similarity scores. Creating the dataset required not only programming and data collection/cleaning resources, but also approximately 3-weeks of computer wall time. Sharing it with other intellectual property researchers allows others to capitalize on this research investment.

Uses for the dataset:

There are many potential applications for the patent similarity dataset. These include improved measures of impact or value, increased granularity in measures of substantive patent content, and better insight into the patent examination system. For example, consider the widespread use of patent citations to measure an invention's impact. These are usually binary measures, with impact assessed by simply counting the number of citations that a patent receives. However, doing so does not address any of the different ways that inventions can have impact. For instance, some inventions have impact on their own technological areas, whereas others have more general impact across a wide-variety of

technical areas. The patent similarity dataset allows us to capture these and other qualitative differences that are overlooked by binary citation measures.

As another example application, I have a recent paper in Research Policy that uses an earlier version of this dataset to measure "boundary spanning" inventions, and subsequently demonstrates that these inventions which draw on disparate areas of technical knowledge have increased in recent years and that examining them strains PTO resources. I have another paper (joint with Laura Pedraza-Fariña and forthcoming in the University of Chicago Law Review) that advocates for more research into empirical measures of nonobviousness and argues that semantic similarity measures may be a useful tool in developing these measures. By sharing the patent similarity dataset, I hope to enable future follow-on research in this vein.

In addition to these examples of ways to apply the patent similarity dataset to innovation research, there are many more that I haven't space to touch on here, and surely even more that I haven't even yet considered. Although the sharing of new datasets and papers about them is somewhat different than the usual IPSC presentation, I hope to not only publicize the dataset at the conference, but to also seek feedback on potential applications and improvements that will further strengthen this dataset and help it contribute to empirical innovation studies. This will be followed by revisions to the dataset description and publication of the dataset. Feedback is an essential part of this process.