The language of innovation

A. Tacchella[1], A. Napoletano[2], L. Pietronero[2]

[1]ISC-CNR Rome

[2]La Sapienza University of Rome

Predicting innovation is a very peculiar problem in data science. Following its definition, an innovation is always a never-seen-before event and this makes the usual approach of learning patterns from the past a useless exercise. Here we propose a strategy to address the problem in the context of innovative patents, by defining an high dimensional space in which dynamics predict specific innovation events as well as global trends of technology popularity and abandon.

We define an innovation as a never-seen-before association of couples of technologies inside a patent (e.g. cameras and motorized steering wheels to build a self-parking car). Each patent is regarded as a collection of two or more technological codes. We define a distance among technological codes and we spot events of codes never occured together but with a small distance. To achieve such result we train a one-layer neural network to guess what technological code has been removed from a real patent (CBOW approach). The internal structure (the rows of a matrix) of such neural network provides an high dimensional embedding of each possible technological code, and distances (or angles) in this high dimensional space prove to be an excellent predictor of innovation. By looking at the distances among the embeddings of never co-occured codes we find 2 main results: 1) close codes have an extremely higher probability of being patented together in the future than randomly selected couples; 2) innovation events happening among close codes are persistent (i.e. patenting activity continues for years after the first innovation), while the rare event of innovation among far-apart codes tend to be a one-shot event.

By looking at the general distance vs. "number of co-occurences" relationship (even for couples that had been patented together in the past) we observe a very rich dynamics with clearly separated laminar flows going towards popularization or abandon of specific technological bonds. We therefore think that this approach provides a solid way of building a consistent euclidean "technological space", where distances and relative position are meaningful with respect to a technological semantic. This "technological space" opens to a vast set of possibile analytic strategies that can be implemented in real space, from clustering methods to dimensionality reduction approaches and any method that benefits from a continuous space embedding.

[l] T. Mikolov, K. Chen et al., Arxiv1301.3781, (2013).

[2] B.H. Hall et al., RAND J. Ec. **-**, - (2005).

[3] C. Martinez, Scientometrics **86**, 39 (2011).