Review of: "Strategic Citations in Patents: Analysis Using Machine Learning"

Shiyan Ou[1]

1 Nanjing University

Potential competing interests: No potential competing interests to declare.

The main idea of this study is to propose measuring the knowledge relevance between two patents by calculating their cosine similarity based on the Doc2Vec representation, instead of the previous citation relationship. Based on this idea, this study provides an interesting analysis of the patterns of patent applicants citing other patents. This idea has some novelty. However, there is an implied assumption that the more similar a patent is to the target patent, the more likely it is to cite the target patent. First of all, it is deserved to demonstrate or explain the correctness of the assumption.

There are still some issues that need to be addressed.

1. The author did not demonstrate the feasibility of Doc2Vec to calculate the similarity of patent abstracts. There are other algorithms to calculate text similarity, which should be compared and selected.
2. The similarity of patent abstracts was divided into several bins. However, the selection process of bins is briefly mentioned, which seems arbitrary. It's unclear whether the concerned conclusion would still hold if the threshold for highest similarity bin was set at 0.5 instead of 0.6. Therefore, the author needs to provide a more detailed explanation about the bin selection.
3. The author introduced patent numbers when generating vector representations in order to enrich the information contained in the generated vectors. However, based on my experience, string such as “US7502754” may actually bring noise and potentially affect the resulting vectors.
4. The author introduced patent numbers when generating vector representation in order to enrich the information contained in the text vectors. However, based on my experience, string such as “US7502754” may actually bring noise and potentially affect the resulting vectors.
5. The article should be proofread further to correct some grammar and spelling errors, such as “Ci-tations”. Additionally, it is uncommon to use the first-person singular in scientific literature.
6. Some statements require necessary citations to support them, such as “Following prior literature, the patent's location is determined as the MSA where the highest proportion of inventors are located.”
7. The arrangement of chapters is confusing. This article lacks a separate "literature review" section. The “introduction” section focuses on the current work itself and provides an insufficient literature review. And the “Discussion” section is actually about the situation of patent applicants changing their firms.
8. It is difficult to locate the data on which several conclusions are based. For instance, in the statement, “However those made to patents within the same MSA has increased, although not consistently over the period: the share of local

backward citations rose from 9.3% in 1985 to 12% in 2015.", the specific data source for the change is unclear and difficult to find. The author should provide more details about the data sources used in this study.