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

Alessia Iancarelli[1]

1 Northeastern University

Potential competing interests: No potential competing interests to declare.

Overall comment: The author presented an interesting work. Indeed, the paper has potential, presents novel approaches and interesting results, however it can use some work (including improving some sections and proof-reading). Some concerns should be addressed prior to publication.

REVIEW:

The study introduces an alternative method to measure knowledge relatedness across patents that goes beyond relying on patent citations. The method uses unsupervised machine learning algorithm Doc2Vec to derive vector space representations of patents using patent abstract text, and then uses cosine similarity to measure their proximity in ideas space.

The use of machine learning methods to categorize and analyze similarity for over one million patents provides a novel approach to uncovering behavioral biases in inventors' citation patterns. The study finds evidence of strategic behavior in citation behavior, such as inventors citing their own prior inventions less after they change firms and applicants strategically omitting citations to patents in different cities.

The findings suggest that the validity of patent citations as a measure of knowledge spillovers is challenged by strategic biases, and the study proposes using patent text similarity as an alternative measure. The study's contribution is important because it sheds light on the limitations of relying solely on patent citations as an indicator of knowledge relatedness across patents and proposes an alternative method that could provide a more accurate measure.

Main suggestions

1) I would suggest to re-write the introduction starting with the research significance of the paper. The author should provide a clear explanation of the research problem, why it is important, and how it contributes to the field. This can help the reader understand the context of the research and its potential impact. at the present state, the author starts the introduction summarizing what he did.
I also recommend adding more literature review.

2) I would mention in the discussion or methodology other approaches for measuring patent similarity (and quickly justifying your choice of using Doc2vec) to establish the validity and robustness of the findings.

Points that I found unclear

2) in the intro, the sentence "The problems with using patent citations to proxy for knowledge flows have been well documented" needs actual citations.

3) I would have appreciate more clarity in how the author determined the change of firm for the inventors.

Other:

1) the following part needs i) citations, ii) full explanation of acronym.
2.2. Patent Abstracts to Vector Space Representations:
[...]I use procedures standard in the NLP literature [...]
2) for clarity, I would suggest to separate the results from the interpretation of the results/citations (belonging to the discussion section).

3) in the section "4.1 Rate of self-citation before and after firm change", in the final period there is a space missing: "I compare the rate *ofinventor* self-citation [...]". This is just one example, I suggest proof-reading.

4) I would add in the discussion the limitation that the paper only uses patent data from the US, so the results may not be generalizable to other countries.

Best of luck,

Alessia