Review of: "Effective distributed representations for academic expert search"

Silvio Peroni[1]

1 University of Bologna

Potential competing interests: The author(s) declared that no potential competing interests exist.

Metadata

Title: Effective distributed representations for academic expert search

Author: Anonymous (due to double-blind peer review process)

Submitted to: First Workshop on Scholarly Document Processing

Review

This work describes how different approaches based on embeddings to represent scholarly documents (e.g. journal articles and conference papers) can be used to implement systems for experts search in scholarly communication. In particular, the authors present a comparison between different approaches, which include text embeddings and citation links.

While, in principle, this work would be an acceptable contribution to the workshop, there are a few aspects that should be clarified and improved in the camera-ready of the paper, listed as follows.

Possible applications

Even if this work is entirely on the embeddings-based algorithms for expert finding, the expert finding activity per se has several applications in the scholarly domain all with different scopes (reviewers finding, collaborators finding, etc.). It would be good, for instance in the introduction, to list some of such applications to provide robust justifications about why the authors run that work and how embedding-based algorithms for expert finding can be used in critical applicative contexts. It would also be useful to hypothesise to which contexts the work run by the authors better applies to – e.g. is the strategy adopted in their study to work better on reviewer finding than on collaborators finding?

Reproducibility of the experiment

For reproducibility purposes, it would be crucial to introduce all the steps and data that have been used to come up with the results presented in this paper. However, it seems that some of these explanations are lacking in the paper. In particular:

1. The author used MAG, but it is not indicated which release in particular. They generally refer to the Open Academic Graph initiative (https://www.openacademic.ai/oag/), but not to the particular dataset used (including its version). It would be helpful, in this case, to properly cite these data within the reference list (see https://www.force11.org/datacitationprinciples and https://doi.org/10.7717/peerj-cs.1), to have a clear view of what has been used in the experiment.
2. "we expanded this set with the references of all the 29237 papers": how did the authors retrieve such references? By looking at the MAG data available? By extracting them from the arXiv sources downloaded? In the former case, are the references complete, i.e. did they include all the article cited or something is missing? In the latter case, which approach did the authors use to extract, harmonise and reconcile metadata of references, in particular when they reference the same cited paper?
3. "we performed a bounded stratified sampling for the authors to retrieve a subset of 5,000 authors who are representative of both highly-, medium- and less prolific author populations": how did the authors obtained this sample? How many representatives for each of the three categories have been retrieved? Since the authors are evaluating the CS domain, how much of these authors are CS researchers? Did the authors use some approach to filter only CS researchers?
4. Since all the data extracted by the authors come from freely available material, it would be crucial to make them available somewhere (e.g. Zenodo or Figshare). This would allow others to reuse them and, even, to replicate the experiment. Would it be possible to publish such dataset somewhere?
5. The authors often refer to material included in the Appendix (tables, algorithms, etc.), which is crucial to fully understand the rationale of some of the steps of the workflow used by the authors in the experiment, as well as to replicate them. However, the Appendix seems to be missing in the paper.

Other general comments

1. Using the papers as sources for creating the embeddings proposed by the authors means to consider all the authors of a paper equally expert on its topics. However, this is not necessarily the case, in particular in interdisciplinary domains (see https://doi.org/10.1016/j.joi.2017.03.003). Did the author handle this aspect in their work?
2. The claim in which the authors say that "authors that cite each other can be considered having similar interest" should be supported by a citation. E.g. see https://doi.org/10.1016/j.technovation.2008.03.009 and similar studies.
3. There is a most updated work about MAG that could be cited instead of the one specified in the reference list, i.e. https://doi.org/10.1162/qss_a_00021.
4. The author said they started from the CS papers in arXiv, obtaining more than 130K papers. However, why other and more comprehensive sources were not chosen for this kind of study? For instance, DBLP (https://dblp.uni-trier.de) would have provided more CS documents (now they have 5,265,753 documents in their DB). Having more input material, in principle, would have allowed selecting a larger number of relevant papers from MAG. Could the author justify their choice of using arXiv only?
5. There are most recent versions of MAG available online (e.g. see https://archive.org/search.php?query=microsoft%20academic%20graph) than the most recent available in the OAG website (MAG included in OAG 2 is dated 2018). Why did the authors decide to prefer such an older version than the new one available?
6. The authors say that they "chose the citation counts of the papers as a proxy for expertise" of their authors. However, this is

an oversimplification of a complicated issue. In the past, studies have shown how the choice of what to cite may derive from the perceived authoritativeness of (some of) the authors of cited papers. However, the authors' expertise, while meaningful, is not the only dimension that seems to be involved when selecting what to cite. For instance, the study in https://doi.org/10.1145/2063576.2063757 shows that also the venue of publication has a role in the citation activity.

Also, there are cases in which a high citation count can be due to different reasons, such as fraud or misconduct. For instance, see the 1998 article published in The Lancet "Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children", which has been cited (source: OpenCitations COCI, as of 27 September 2020, API call http://opencitations.net/index/api/v1/citations/10.1016/S0140-6736(97)11096-0?filter=creation:%3E2010) 620 times after he was retracted in 2010 vs the total citation count of 1,008 (same source). Several of the citations after (but even before!) the retraction are *negative* - i.e. they cite it explicitly as a retracted article and warn against its results. Thus such citations do not strictly highlight the expertise of the authors of the document, in this case.

All these aspects should be taken into account in the current work, and at least should be presented as current limits of the study.