Finding Related Works by AI

DM Chiu, Feb 18, 2023

The other day, a former student sent me the pre-print of a paper to appear in a prestigious venue with a comment: “This reminded me of your work on X”. Not surprisingly, the paper does not cite our work. In the ensuing discussion, the student speculated on the reason for the possible negligence. Though it is a slight letdown, I told him this happens, can be committed by you and me, and is a result of “the challenge of scale”.

The fact is, as we all know as an academic, that there is increasingly large amount of related work to keep track of in research; this is especially the case if you are working on a hot topic area (such as AI) or in an inter-disciplinary area (in which case you may not have been keeping track of the area on a continuing basis). Is there a way AI can help us with this problem? As people discuss daily what ChatGPT can do and cannot do (though I have not tried out ChatGPT personally), it occurs to me that ChatGPT has demonstrated that building such a tool is quite feasible. We may refer to it as RW-GPT (RW stand for Related Works). Its purpose is to dig out all the related works given the draft of a research paper. For this purpose, RW-GPT can be fed and pre-trained only with documents related to academic research. Such specialization should make training of RW-GPT less compute intensive, and help it do the job more accurately. If such a RW-GPT is available, it can be used at least in two ways: (a) by authors drafting a research paper, and (b) by reviewers of papers submitted for publication. Similar AI tools have already been used widely to screen out plagiarism, for both publications and student course work. The task of finding related works, however, can be considerably more demanding, as it is not a matter of simply matching sequences of words, but comparing the ideas discussed by different sequences of words.

In the last few weeks, I attended two talks: one by Professor James Evans[1], and the other by Professor Aaron Clauset[2]. Both talks are on the topic of “science of science”, in other words trying to understand how scientific research is done by the community of researchers. Evans’s talk summarized their research on how novel and innovative research are produced, for which he gave several interesting insights. In the Q&A session, I asked him if they find the rate of publication has increased a lot from the past, and if so why; is it because people are more productive? He agreed that the rate of publication has increased a lot, and claimed that the reason is not people are (on average) more productive, but rather because many more people are doing research. This is consistent with my own understanding[3]. What is interesting is that Evans also claimed that the rate of novel discoveries has not increased likewise. This implies an increase of cost or decrease in efficiency. In Clauset’s talk, he

[1] Available on YouTube: https://www.youtube.com/watch?v=JjykhfkrElM&t=59s

[2] Available on YouTube: https://www.youtube.com/watch?v=cX2sXEMkKhw

[3] Quite a few years ago I met Michiel Kolman (a VP of Elsevier) and he told me that according to their statistics, on average each author published one paper a year and this rate has been roughly the same for many years; and I later did some study using other datasets and verified that.

explained that they found there is a strong causal relationship between prestige (the affiliation of an author) and an author's academic standing/success. This correlation, for example, over-shadows where the author was educated. There can be interesting speculations on the implication. Of course, affiliation with a more prestigious institution likely implies more resources and more valuable connections to other researchers. But it may also imply prestige brings more attention to one's research. But, as we briefly discussed in the Q&A session of that talk, ideally attention should be accorded more due to the relevance and novelty of the research rather than simply prestige associated with the author, for the benefit of more efficiently advancing knowledge discovery. The RW-GPT tool mentioned above may be helpful in this regard.