Evaluating pre-trained transformer models for similar article recommendation in PubMed

Li Zhang

May 22, 2021

Abstract

The work presented here promises to make the exploration of scholarly material faster and more accurate. Though the simialr method identificaiton method shows in this paper is for PubMed, we believe it can be migrate into other bibliographic databases/digital libraryies and border field.

1 Introduction

Simialr article recomendation is an important feature in many academic seaching database/digital libraries. It enable users to go though relative researher quackly, play an impor6tant role in improving user searching experimence, while more importantly, is can help to accurate dismi the valuable biomedical findings. Improved literature search engines can save researchers time and effort by making it easier to locate the most important and relevant literature. [**?**]

How long has PubMed supplied "Simialr article" function? Like other system, ResearchGate, Scopus, NCBI's PubMed system, has integrated this feature since *, However, the their method to find the similar article is still very *. To date, with the fast developemnt of nature language proccessing, the cutting-edge techniques have prove an opputunity to improve simialr article recommendation performance.

In this paper, we show the simialr article performance are be largely promoted by using large-heavy pertrained language model.

PubMed related article links identify closely related articles and enhance our ability to navigate the biomedical literature [**?**].

PubMed has integrated the "Simialr article" function for a long time,

show the case here.

significence: Fig * shows the similar article recommendation functionality, this function is very helpful for biomedical scholar, as a recent works by NLM/NCBI team suggest user needs can be largely improved while users explore related articles. can power PubMed user experience,...

Why similar article recommendation is very important for further improveing search exerperience, why it is a necessary functionality? Related works: how PubMed improving use searching experience. To improving user searching experience, NCBI has provided many in place measurements from serverl aspects. such as autocomplition, ..., . refrequent search terms recommendation. However, this measures can be not necessary enough, user may expore other What did europen Pubmed did? and what did other platform did?

In many academic service platform, such as web of science,..., they

How did others find similar article?

Our contribution are three parts. we provide an method to automaticilly build similar article dataset for development models we evaluated the most well-known pre-trained models on four dataset, and emperical evaluation shows fine-tuned * model shows state-of-the-art result. Using this method, we obtain the paper distribution vector for whole PubMed papers.

This paper targets for three groups of readers. 1. NCBI's PubMed, we created a enhanced simialr article methods for PubMed, and show better resuts than the same current result in PubMed. Moreover, we successful compress the orign 768-long vector into a short concise vecotr for entir PubMed paper. This exercise help to make this method praticle. Base on the consise vector, we also pre-computed the "nearest" paper for all PubMed papers, and relase the paper vectors for facilicate future studies. 2. Research group for science of science. This paper them to better visual the clustering of indexed paper in PubMed, and help them to manage/mantuance. For each paper, one can the tool in our released code to fast identify the simialr articles by simialarity in descding order.

In the following section. We review exisiting or potential methods for this task, and review evaluation datasets in Relatied work section. Then, in method section, we show how to buliding dstribution representtation for articles, and how to build our evaluation dataset.

2 Related works

2.1 Method review

2.1.1 PubMed Similar Article Function

How NCBI did? To identify similar article in PubMed, the NCBI group has invested many efforts. According to the recent description of the implementation of this functionality [1], [2]. measuremnets like word-level * is used to computed the similarity. However, this suluation may be not an optimal one, becasue the words extracted from citaiton abstract already helpful for accurately simialrity becasue some general words can be convery the research content appropriately. epsecially for synonyms/, it will result in the fact that similar paper will be remove due to words with close meanings appear in similar articles. They mainly rely on lexical overlap, they cannot learn the semantic similarity, this is especifically occurs for some complexity compund and long biomedical phrase.

How other researcheres did for PubMed? In recent years, there are several studies on similar paper identification task for PubMed.

Assume this task is text similarity problem, is there any methods can do it better? From a broder view, this task is a spefific case of text similarity problem, i.e., text similarity for biomedical articles. Therefore, techiniques

[1] https://pubmed.ncbi.nlm.nih.gov/help/computation-of-similar-articles
[2] https://pubmed.ncbi.nlm.nih.gov/help/similar-articles

on text similarity can be borrowed for improving the experimence of simialr article exporelation in PubMed. In addition to the existing studies for this task, Recent advances on natural language processing has achieve great progress, many effective models such as deep text matching, Bert cite are proving to be effective on similar text task. These models have removed the barrie of words mearning variants and close words meanring between articles by considering the simialrity from a semantic prospective. What we can do for improving it? In this work, we investigate the cutting-edge text similarity approaches in NLP for biomedical arcicle and evaluating them for this helpful functionality.

2.2 Dataset review

Show detailed is there any evaluation dataset? how others bulding the dataset? similar paper finding as a information retrivel problem.

3 Method

We employ SBert model [3], a increasing number of state-of-the-art pretrained models for more than 100 languages, fine-tuned for various use-cases. Specifically, BioBert, SciBert and ... are introducted for comparsion. Since these models are pre-trained on a vast scientific-text corpus for general tasks. it may not achieve perfect performance on the specific task. Therefore, in order to obtain better results, we fine-tune these models and evlauate the original models and the fine tuned models seperatly.

3.1 dataset

We ranomdly selected 1M articles from PubMed citations, and developed programs to bulk retrieve the similar articles for the 1M samples from PubMed official site using its open APIs via biopython [**?**] package.

In this section, we show how we did to bulid the dataset. For evaluation dataset. We ...

Note that we did not consider the publication timeframe, as we can a later-published article can exist in the related article list.

plot a database enhanced here by NER and Europen PubMed.

Rules to selected ground truth simiar paper for buliding evaluation dataset. In order to buliding evaluation dataset. We used two rules to identify the

[3] https://www.sbert.net/

similar paper from entire PubMed citations. These rules ars inspired by the characteristics of simiar papers.

Actually, very similar paper usuaally share simiar research purpose, and they use simialr methods or tools. These identifiable feature can be extracted by the representative terms-MeSH, which a standard resource for the whole sprctrum of biomedical scientific papers. Currently, the MeSH terms in each citations are assgined carefully by the NLM's curator team. MeSH are used for widely applications not restricted to paper search. MeSH are indexed though the coolabration and machine-aided indexing and human curators. Another promint advantage of MeSH is semantic indexing over lexical, the same MeSH term can labeled to a paper though it not exists in the paper's title and abstract.

In addition, the citing-cited relationship between articles is another fearture need exploration for building dataset. Its can be understood that the cited articles in refernece usually have semantic connection for the citing paper. Thus, the relationship can be much helpful for bulding and accurate paper similariy dataset.

However, not mantter MeSH list or refernece list is very sparsity for PubMed paper. On average, there are * percentage paper having MeSH, and there are only * number of MeSH for these articles. For citations, the low density of citation relationship of PubMed is a well known problem. Similar to MeSH list, only * percentage paper having reference, and only average * number of citations for these articles.

To handle this issue, we made considerable efforts to enhance the two metadata. For MeSH enhancement, we use the NER result of PubMed Knowledge Graph [**?**]. The goal of the NER task is to recognite named entity from PubMed citation content (i.e, abstract). Our empericall evaluation found there are nearly *% entities are included by MeSH. To supplement the additional MeSH, we fist filter MeSH terms from NER entity list and then assicaiting them to the whole PubMed citations by the via PubMed paper ID (PMID). with repsect to the refernece sparsity problem, we compensate it by retrieve the reference information from Europe PMC [4], which is a service of the Europe PMC Funders' Group, in partnership with the European Bioinformatics Institute ; and in cooperation with the National Center for Biotechnology Information at the U.S. National Library of Medicine (NCBI/NLM) . It includes content provided to the PMC International archive by participating publishers. By the open API [5] of Europe

[4]https://europepmc.org/

[5]https://europepmc.org/RestfulWebService!/Europe32PMC32Articles32RESTful32API/references

PMC, We are able to obtain a relatively complete reference list. Through this exercise, the overall density of refernence for PubMed is * in comparsion to *% of original refernce density.

To incorperate it. we

In this paper, Enhanced Mesh and Enhanced Ciation. Simialr paper usually has.

However,

MeSH/Citation

3.2 fine-tuning

4 Results

4.0.1 Dataset

4.0.2 Competing methods

As discussed in relative words section, there are many methods in NLP for text similarity. For evlauation, we introduce Bert, ... For comparsion, we introduction the recommendation result shown in PUbMed interface (see "Simialr article" in page navigation of a particular paper)

4.0.3 Evaluation metrics

Ranking metrics, Information retrivel metrics and paper title overlapping

5 Discussion

5.1 how this works can be intregtared into PubMed system?

fast, very short words embeddings, inductive-infering.

5.2 user study

how it can be

5.3 limitation

5.4 future protential imporovement

This study only recomnedation paper from semantic perspective, however, in many commercial recommender system, the recommeded items may not

ony semantic relatedness, other relationship the relationships that a mature system should consider is not equal to semantic relatedness. other crucial aspect such as ... also need consideration. However, we can not obtain such real word dataset, i.e, integrating PubMed searching log to develop more powerful recommneder system. Thus, we can image the paper recommnedation system in PubMed can be more powerful by leveraging state-of-the-art techinique in recomendation system and information retrival. for example, recommendation with intrepretation,

exploring more user intellience that can be available in NCBI

6 Conclusions

In this works, we shows an effective and effecient way to locate similar PubMed for power user search experpence in PubMed system. This study provide initial, portary investigation on simiar article for PubMed system. ... Our future works intent more relationship in paper recomendation.