Evaluation and Modification of Local Citation Recommendation Systems:
A Comparative Study of available Embedding and Similarity Search Techniques.

Kabir Jaiswal, Nigam Patel, Spriha Jha

Tandon School of Engineering, New York University, NY, USA
{kj2294, np2726, sj3520}@nyu.edu

Problem Statement

This project attempts to analyze and improve the pre-fetching step of the two-stage Local Citation Recommendation System [1] through a comparison study of GloVe, OpenAI, and Cohere embeddings, as well as cosine similarity, Pinecone vector search, and Nomic's Atlas similarity search approaches. The purpose is to carry out a thorough comparative examination of how various embeddings and similarity search methods affect the effectiveness of the citation suggestion pipeline. The portion of the prefetching stage that we want to compare is depicted in the accompanying image.

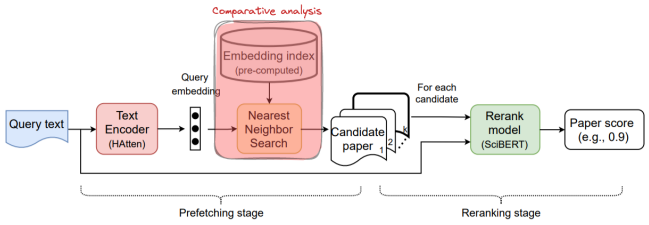

Fig. 1: Overview of our two-stage local citation recommendation pipeline.

Literature Survey

Primary focus areas:

1. Citation Recommendation Systems (CRS) : In-depth research has been done on content-based, collaborative filtering, and hybrid citation recommendation systems. A two-stage local CRS pipeline with prefetching and reranking steps for improved recommendations was suggested by Gu et al. in 2022 [1] [Colab][Github]
2. Text Embeddings: Text embeddings from OpenAI and Cohere are two examples of embedding models that have been created for the purpose of capturing semantic information and are frequently used in NLP tasks including document similarity and information retrieval.
3. Similarity Search Techniques: Different similarity search methodologies, including Pinecone's vector search and Nomic's Atlas, have been created for high-dimensional spaces and used in recommendation systems.

Dataset

The datasets listed below, as suggested by Gu et al. (2022)[1], will serve as a good starting point for examining the effects of different text embeddings and similarity search methods on the local Citation Recommendation System pipeline.

- ACL-200, FullTextPeerRead, RefSeer, arXiv [1] Google Drive
- This Github repo contains a "pseudo" custom dataset that is actually ACL-200 as used in Colab.

Model(s)

Prefetching Stage:

- Hierarchical-Attention Text Encoder
- OpenAI, Cohere for document vectors.
- Pinecone vector search and Nomic's Atlas for similarity search techniques.

Reranking Stage:

- SciBERT-based [2] reranker to refine the list of prefetched citations based on contextual relevance.

Expected Results

- Analyze the performance of the SciBERT-based Reranking pipeline and the Hierarchical-Attention Text Encoder using quantitative methods using the aforementioned embeddings and similarity search methods.
- Give a metric table, similar to Figure 4 in the given research [1], that illustrates the effects of various combinations on performance indicators.
- Choose the combination that will increase the accuracy and productivity of the pipeline and allow for more accurate citation recommendations.

References

1. Gu et al. (2022, April). Local citation recommendation with hierarchical-attention text encoder and SciBERT-based reranking. (DOI)
2. Beltagy, et al. (2019, Nov). SciBERT: A Pretrained Language Model for Scientific Text (DOI)