# Important Text Features For Paper Source Tracing

Solution for the 4th Place in the PST Task of the 2024 KDD Cup OAG Challenge

Guolin Xu†
School of Computer Science and Technology
Chongqing University
Chongqing 404100 China
xgl0626@163.com

Kehuan Feng
College of Hanhong
Southwest University
Chongqing 404100 China
mayuanwoer@gmail.com

Ao yu
School of Computer Science and Technology
Chongqing University of Posts and Telecommunications
Chongqing 404100 China
1360522976@qq.com

## ABSTRACT

With the rapid advancement of science and technology, the numbe r of academic papers has surged significantly, making it increasing ly challenging for researchers to trace the historical context of tech nological development. The Paper Source Tracing (PST) task aims to identify the references that most inspire a specific paper, know n as the source papers. We propose a method based on BERT for p aper source trace. By enhancing the model's performance through parameter optimization, feature construction, and model fusion, w e achieve favorable experimental outcomes on the test datasets. He re is the code https://github.com/xgl0626/paper-source-trace-chine segpt-rank4.git

## KEYWORDS

Paper source trace, BERT, Identifying source paper

## 1 Introduction

The swift progress in science and technology has led to a dramatic surge in academic paper publications. Annually, millions of papers are released across the globe, and this trend continues to accelerate. According to data from the Scopus database, by 2021, the total number of academic journal papers published worldwide had reached 220 million, encompassing all academic papers published since the 17th century across various fields, including natural sciences, social sciences, and humanities. For researchers, it is becoming increasingly difficult to trace the origins and direction of technological development amidst such a vast volume of literature. Therefore, tracing the direction of the source has garnered significant attention from researchers. Several methods have been proposed to address this problem. "Identifying Meaningful Citations"[1], a paper presented at the AAAI workshop in 2015, can identify significant references in academic papers and achieves 90% accuracy with a recall rate of 65%. In the paper "MRT: Tracing the Evolution of Scientific Publications"[2], published in TKDE in 2021, the proposed method not only sorts and filters the crucial references of the target paper but also provides a traceability tree. In 2024, Zhang et al.[3] proposed a benchmark for academic graph mining, providing an overview of recent advancements in tracing paper origins, and introduced both traditional methods and current datasets.

### 1.1 Paper Source Tracing Task

The purpose of the paper source tracing task is to identify the most crucial reference, known as the source paper, from the full text of a given paper, denoted as p. The source paper is typically the literature that is most enlightening for the given paper. Each paper may have one or more source papers or none at all. For each reference within the paper, an importance score ranging from 0 to 1 should be assigned. The following aspects determine whether a reference constitutes a source paper:

1. Whether the main idea of paper p is inspired by this reference.
2. Whether the main method of paper p is derived from the reference.
3. Is the reference indispensable for paper p? That is, without the work of this reference, paper p could not have been accomplished.

In this competition, participants are tasked with rating the importance of each reference for a given paper in XML format. The training set and validation set collectively contain 394 data items, represented as a list of dictionaries.

## 2 Method

In this section, we introduce the method of source paper tracing, mainly including data construction, model and impovement.

### 2.1 Data Construction

For the data construction part, we extract two main fields: reference fragments and titles. Since each reference is cited in the paper, all citation fragments for each cited document are located within the paper. These fragments are truncated using a fixed window of 200 characters in length. Considering the number of source papers is relatively small, the proportion of positive to negative samples is managed by randomly collecting samples, ensuring that the ratio of positive to negative samples is 1:10 without repetition.

It is important to note that the test dataset includes data from papers on arXiv at the time of model inference. Since published papers are generally not sourced from arXiv, when constructing the training data, the titles of cited papers that have been published are included, while the titles of cited papers from arXiv are left empty and categorized as negative samples. So the data example for the input model is {title of the cited paper, fragment of the cited paper}.

## 2.2 Model And Improvement

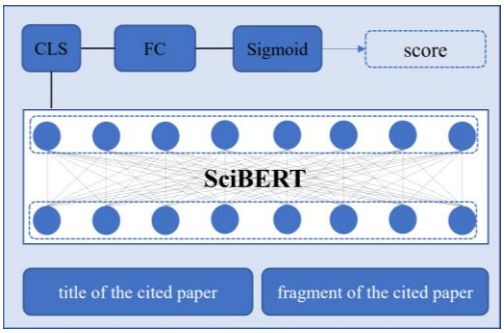


**Figure 1: Model Structure**

For the pre-training model, BERT[4] is chosen as the classification model, with SciBERT[5] providing the pre-trained weights. The approach involves inputting the citation fragment text and the title of the cited literature into the classification model to enable feature interaction and determine whether it is the source paper.

For Parameter optimization, by adjusting the learning rate, batch size, training epoch, warmup proportion and gradient clip, the model can better learn the important features in the data to identify the source paper.

For model fusion, the train dataset is randomly divided into ten parts and ten models are trained. The three models with the best test dataset scores are selected for model fusion to enhance the performance of the model.

# 3 Experiments

## 3.1 Setup

The experiments are conducted on the NVIDIA RTX4090 GPU. The model parameter settings were as follows: Batch Size is 16, Epoch is10, learning rate is 1e-5, warmup proportion is 0.1, and max gradient norm is 10.

## 3.2 Main results

The single model score of test dataset A is 0.454, and it is 0.465 after model fusion. The single model score of test dataset B is 0.434, and it is improved to 0.449 after model fusion, which is significantly improved compared with the feature engineering method based on machine learning.

## 3.3 Ablution studies

On test dataset A, the baseline score of this task is 0.283, and the model effect is improved to 0.414 after parameter optimization, by 4% after data reconstruction, and by 1.1% after model fusion.

# 4 Conclusion

Overall, we implement a classification method for paper source trace based BERT. The model we used serves as a simple baseline method, and there are numerous areas for optimization in feature construction.

# ACKNOWLEDGMENTS

Thanks to the 2024 KDD Cup OAG Challenge organizers for providing the competition opportunity and dataset.

# REFERENCES

[1] Valenzuela M , Ha V , Etzioni O .Identifying Meaningful Citations[J]. 2015.
[2] Yin D , Tam W L , Ding M ,et al.MRT: Tracing the Evolution of Scientific Publi cations[J].IEEE Transactions on Knowledge and Data Engineering, 2021, PP(9 9):1-1.DOI:10.1109/TKDE.2021.3088139.
[3] Zhang, F., Shi, S., Zhu, Y., Chen, B., Cen, Y., Yu, J., … & Tang, J. (2024). OAG -Bench: A Human-Curated Benchmark for Academic Graph Mining. arXiv prep rint arXiv:2402.15810.Conference Name:ACM Woodstock conference
[4] Devlin J , Chang M W , Lee K ,et al.BERT: Pre-training of Deep Bidirectional Tr ansformers for Language Understanding[J]. 2018.DOI:10.48550/arXiv.1810.04 805.
[5] Beltagy I , Lo K , Cohan A .SciBERT: A Pretrained Language Model for Scientif ic Text[J]. 2019.DOI:10.18653/v1/D19-1371.