Multi-source information fusion for tracing the sources of papers

Yuxuan Wu*
wux521@zju.edu.cn
School of Software Technology,
Zhejiang University
Hangzhou, Zhejiang, China

Genhang Shen*
shengenhang@zju.edu.cn
School of Software Technology,
Zhejiang University
Hangzhou, Zhejiang, China

Hongyu Fan
fanhongyu@shu.edu.cn
School of Economics,
Shanghai University
Shanghai, China

ABSTRACT

With the rapid advancement of academic data mining technologies, tracing the origins of papers has become more thorough and precise, aiding in the identification of pivotal papers that significantly impact entire research fields. The PST task in KDD Cup 2024 was introduced to address this challenge. This paper proposes an effective solution that employs a dual-network architecture to integrate information from both the papers and their references. To address the issue of insufficient information, we utilized data from DBLP Citations as well as contributions from the introduction and conclusion sections. Additionally, we set two optimization objectives: maximizing the similarity between related papers and minimizing the similarity between unrelated papers. The model was trained using reverse gradient propagation. Our team, pigpigwin, achieved 10th place in KDD CUP 2024, with a final test set score of 0.38159, demonstrating the effectiveness of our proposed solution. Codes are available at https://github.com/kddcup24-10th/PST-KDD2024.

CCS CONCEPTS

• **Information systems → Information retrieval**.

KEYWORDS

Natural Language Processing, Text Similarity, Paper Ranking

ACM Reference Format:
Yuxuan Wu, Genhang Shen, and Hongyu Fan. 2024. Multi-source information fusion for tracing the sources of papers. In *Proceedings of KDD 2024 OAG-Challenge Cup.* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

In today's rapidly advancing technological landscape, the number of academic papers published each year has been increasing dramatically. According to data from the Scopus database, as of 2021, the total number of academic journal articles published globally has reached 220 million, encompassing a wide range of fields from natural sciences to humanities. This vast volume of literature presents unprecedented challenges for researchers in acquiring new knowledge, understanding scientific progress, and formulating research strategies. Identifying key influential source papers from this sea of documents has become a critical issue that both academia and industry urgently need to address.

With the development of academic data mining techniques, Paper Source Tracing (PST) [6] has garnered widespread attention. The goal of PST is to identify the references in a given paper that significantly influenced its core ideas or main methods, referred to as "source papers" (ref-source). This identification not only helps researchers quickly grasp key literature in a specific field but also provides crucial insights for scientific policy-making and research direction planning. For example, a study [4] published in Information Retrieval in 2014 explores the enhancement of query expansion in patent retrieval through citation analysis. Additionally, this research indirectly examines the role of citation networks in identifying key literature and evaluating their influence.

In this context, the KDD Cup 2024 has introduced the OAG-Challenge to advance the frontier of academic knowledge graph mining techniques. As a significant component of this challenge, the Paper Source Tracing (PST) task requires participants to design automated models that extract useful information from XML files and evaluate the importance of references. The ultimate goal is to identify the source papers that have significantly influenced the given paper and are indispensable to its development.

We aim to explore methods for tracing the origins of academic papers based on natural language processing and machine learning techniques. In the KDD Cup competition, we achieved a tenth-place finish. In the following sections, we will briefly analyze the dataset structure, provide a detailed description of the methods employed, and outline the experimental setup. Finally, we will conclude with a summary of our results.

2 TASK DESCRIPTION

In the paper source trace task, with the given full text of the paper p, it is required to identify the references that have the most profound impact on the paper p, also called the "source paper". A paper may cite one or more sources or none at all. Source papers usually have one or more of the following characteristics:

- The core idea of the paper p was inspired by the reference.
- The main research methodology of the paper p originated from that reference.
- Reference is so crucial to thesis p that thesis p would not be complete without its work.

Specifically, given a paper's ID p and its full document D, we can extract from D a list of references for paper p named $rps = [rp_1, rp_2, \cdots, rp_k]$. The goal is to find the source papers of paper

[*] Both authors contributed equally to this work.

KDD 2024 OAG-Challenge Cup, June 01–21, 2024, Barcelona, Spain

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

p from rps. It is needed to develop an algorithm to compute the correlation between paper p and each element of rps and quantify this correlation as a value between 0 and 1, therefore identifying the source papers based on the value of their correlation.

3 PROPOSED METHOD

3.1 Data Preprocessing

We first employ data preprocessing methods to extract the abstracts from the XML files of the articles. However, since some articles may not be included in the XML files or may lack abstracts, we need to use other data sources to fill these gaps. To address this, we extract relevant information from the DBLP Citation database and retrieve the main contributions from the introduction and conclusion sections of the articles. By doing so, we aim to gather comprehensive information to compensate for the missing abstracts.

However, this approach may introduce a significant amount of noisy data, affecting the accuracy of subsequent analyses. To improve data quality and accuracy, we considered using large language models (LLMs) to clean and process the data. LLMs possess powerful natural language processing capabilities, enabling them to identify and filter out irrelevant information while extracting key details. Nevertheless, due to resource constraints, we currently employ rule-based filtering techniques to eliminate some of the noisy data and have not yet implemented the LLM-based approach.

Additionally, for abstracts that we are unable to obtain, we constructed a 768-dimensional learnable vector initialized to zero to represent these missing data. This approach allows us to cover all necessary information as much as possible while enabling the optimization of the representation of these missing data through the learning mechanism during model training.

3.2 Model Construction

As shown in Figure 1, we adopted a dual-network architecture. First, the paper data is divided into context and abstract sections. The abstract is used to compare the similarity between the paper and all its references, thus requiring the abstracts of both the paper and its cited references. We have preprocessed and stored these data in JSON files, which can be directly accessed when loading the data.

The context information is fed into Module 1, generating a 768-dimensional embedding vector. This is similar to the Scibert model used in the baseline; however, the difference is that we extract the final layer's embedding vector directly, rather than passing it through an additional classifier layer as done in the baseline.

The abstract information is fed into Module 2, with inputs consisting of the current paper's abstract and the abstracts of the papers it cites. We pass these two inputs through the same network, generating two 768-dimensional embedding vectors. We assume that if a paper is inspired by another, the core ideas in their abstracts should be related. We use the oagbert-v2-sim model to extract the content of these abstracts and identify their similar parts. Then, we concatenate the embedding vectors of the two papers and pass them through a feedforward neural network (FFN) layer composed of two MLPs, ultimately producing a single 768-dimensional embedding vector.

Finally, the embedding vectors generated by Module 1 and Module 2 are concatenated and passed through a single-layer MLP to produce prediction scores. This approach effectively integrates the context and abstract information of the papers, enhancing the model's accuracy in paper source tracing tasks.

3.3 Optimization Objective

The overall model is designed using the cross-entropy loss function, as shown in formula 1. In this formula, p_i represents the model's predicted probability of a paper being a source paper, while y_i denotes the actual label indicating whether the paper is a source paper or not. This choice of loss function allows us to effectively train the model to accurately classify papers based on their source status.

$$\mathcal{L}_c = \frac{1}{N}\sum_{i=1}^{N} y_i \log(p_i) + (1-y_i)\log(1-p_i) \tag{1}$$

Furthermore, to strengthen the relationship between a paper and its source paper while distancing from unrelated papers, we introduced cosine similarity calculation in Module 2. Specifically, we compute the cosine similarity between two vectors outputted by OAG-BERT to obtain $\hat{y}_i$, and use Mean Squared Error (MSE) loss function to optimize the model. During training, we denote the labels as y_i, assigning a value of 1 to source papers and 0 to unrelated papers. Through backpropagation, we continuously adjust model parameters to achieve the desired outcomes. The formula 2 is as follows, where h_1 and h_2 represent 768-dimensional features extracted by OAG-BERT, and σ denotes the sigmoid activation function. This method is akin to contrastive learning, maximizing similarity between related papers and minimizing similarity between unrelated ones, thus enabling the model to effectively distinguish between source and non-source papers.

$$\mathcal{L}_m = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \sigma\left(\cos(\theta_{h_1,h_2})\right)\right)^2 \tag{2}$$

The overall loss function is as follows,

$$\mathcal{L} = (1-\lambda)\cdot\mathcal{L}_c + \lambda\cdot\mathcal{L}_m \tag{3}$$

where λ is the hyperparameter which is used to control the weights of two losses.

4 EXPERIMENTS

4.1 Experimental Setup

4.1.1 *DataSets.*

In the training phase, we used the official dataset[1] of 788 labeled records provided by the tournament, which includes papers' references and source papers. For the validation phase, we utilized the officially provided validation set. To obtain additional metadata about the papers, we leveraged the DBLP [5] and OAG [7] datasets, which contain information on authors, publication years, abstracts, and more.

4.1.2 *Baselines.*

We compare our approach with three baselines on the validation set.

- Random Forest (RF) [2] is an ensemble learning technique that constructs multiple decision trees during training and combines their predictions to improve accuracy and prevent overfitting.

[1] https://www.biendata.xyz/competition/pst_kdd_2024/data/

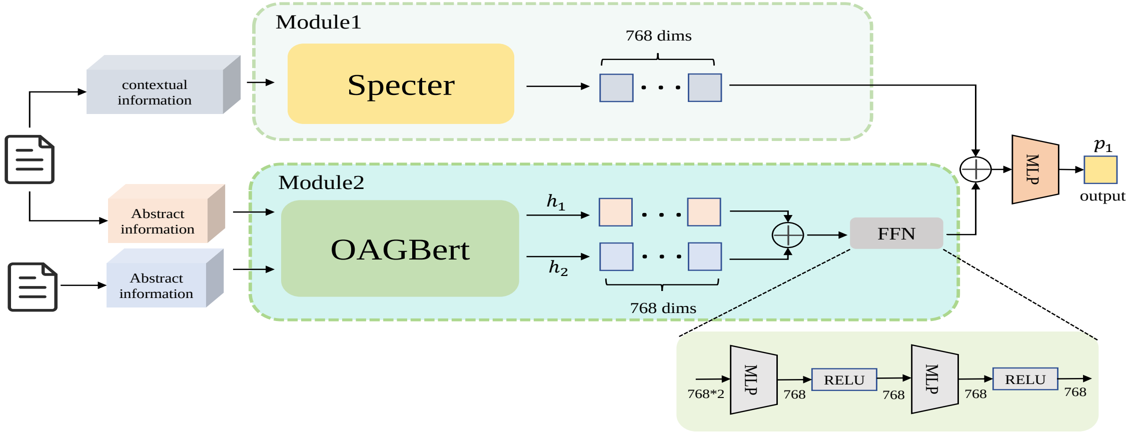

Figure 1: An overview of our proposed model architecture. We extract contextual information of references from the paper and abstract information from two corresponding articles. The contextual information, processed by Module 1, is transformed into 768-dimensional vectors. Similarly, Module 2 extracts 768-dimensional vectors from the abstracts. To align these vectors effectively, we incorporate an FFN (Feedforward Neural Network) layer within Module 2 for further refinement.

- ProNE [8] is a fast and scalable network embedding model, by using sparse matrix factorization and spectral modulation, significantly reducing computation time for large-scale networks.
- SciBERT [1] is a pre-trained language model tailored for scientific text, optimized with domain-specific vocabulary and structure to improve performance on scientific tasks.

4.1.3 Evaluation Metrics. We evaluate the model's performance using Mean Average Precision (MAP), which averages the Average Precision (AP) scores across all queries to quantify the quality of ranking algorithms. AP measures precision at each relevant document retrieved in the ranked list, offering a comprehensive assessment of the retrieval system's effectiveness and accuracy.

4.1.4 Parameters Settings. We implement our method with Pytorch, which is a widely used deep learning framework, using the AdamW optimizer [3] with the learning rate $3e-5$. The value of λ is searched in $\{0.05, 0.1, 0.15, 0.2\}$. And we chose the hidden layer dimensions of MLP in [200], [500], [768], [1000]. All experiments are done using a single RTX 4090 GPU.

4.2 Overall Performance

We compared our model with the baseline model on the validation set and the results are shown in Table 1. Our approach achieved considerably performance gain over the baseline solutions. This finding underscores the importance of multi-source information fusion in the task of paper source traceability within the domain of academic research.

4.3 Ablation Study

We conducted ablation study to systematically eliminate the component loss $\mathcal{L}_m$. Specifically, we set the lambda to different sizes, thus adjusting the weight of the two losses. And the result is shown in Tabel 2. It can be observed that the incorporation of the loss term $\mathcal{L}$ contributes to the enhancement of the model's performance, which corroborates the effectiveness of information aggregation across multiple modules.

Table 1: Overall performace on validation set.

Methods	MAP
RF	0.21420
ProNE	0.21668
SciBERT	0.29489
Ours	**0.42815**

Table 2: Ablation study on validation set. $\lambda = 0$ indicates that the model has only $\mathcal{L}_c$

Methods	MAP
$\lambda = 0$	0.40063
$\lambda = 0.05$	0.4045
$\lambda = 0.1$	**0.42815**

5 CONCLUSION

This paper presents our approach for the PST task in KDD CUP 2024. We designed two pathways: one uses Specter to extract contextual information from papers, while the other extracts abstracts from XML files and DBLP citation and uses OAGBert for similarity calculation. These two pathways are then fused through a concatenation operation and processed by an MLP layer to output prediction scores. Additionally, we designed two optimization objectives to guide the model training process. Ablation experiments validated the effectiveness of our approach, showing that our method improved the baseline model (using Scibert) from a MAP of 0.29489 on the validation set to 0.42815. As a result, our solution achieved 10th place in the PST task of the KDD CUP 2024 challenge.

REFERENCES

[1] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676* (2019).

[2] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32.

[3] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[4] Parvaz Mahdabi and Fabio Crestani. 2014. The effect of citation analysis on query expansion for patent retrieval. *Information retrieval* 17 (2014), 412–429.

[5] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 990–998.

[6] Fanjin Zhang, Kun Cao, Yukuo Cen, Jifan Yu, Da Yin, and Jie Tang. 2024. PST-Bench: Tracing and Benchmarking the Source of Publications. *arXiv preprint arXiv:2402.16009* (2024).

[7] Fanjin Zhang, Xiao Liu, Jie Tang, Yuxiao Dong, Peiran Yao, Jie Zhang, Xiaotao Gu, Yan Wang, Evgeny Kharlamov, Bin Shao, et al. 2022. Oag: Linking entities across large-scale heterogeneous knowledge graphs. *IEEE Transactions on Knowledge and Data Engineering* 35, 9 (2022), 9225–9239.

[8] Jie Zhang, Yuxiao Dong, Yan Wang, Jie Tang, and Ming Ding. 2019. Prone: Fast and scalable network representation learning.. In *IJCAI*, Vol. 19. 4278–4284.