The Solution of WhoIsWho-IND-KDD-2024

Zhe Zhang
Nanjing University of Science and Technology
Nanjing Shi, China
qianlan.z@qq.com

Weili Guo
Nanjing University of Science and Technology
Nanjing Shi, China
wlguo@njust.edu.cn

Qingyuan Jiang
Nanjing University of Science and Technology
Nanjing Shi, China
qyjiang24@gmail.com

Abstract

The rapid growth of online publications has complicated the problem of disambiguating authors with the same name, resulting in errors in author rankings and award fraud. To tackle this, the OGA-Challenge Team published a large-scale dataset and hosted the KDD Cup 2024 Challenge to detect paper assignment errors based on author and paper metadata.

In this paper, we propose a method using two strategies. The first strategy involves extracting features from papers and authors, and using machine learning techniques, specifically tree models, for prediction. The second strategy constructs a graph neural network to capture the relationships between authors and papers. By integrating these two approaches, our method achieves promising results in detecting misassigned papers.

ACM Reference Format:
Zhe Zhang, Weili Guo, and Qingyuan Jiang. 2024. The Solution of WhoIsWho-IND-KDD-2024. In . ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

The rapid growth of online publications has significantly complicated the problem of disambiguating authors with identical or similar names. This ambiguity leads to inaccurate author rankings and misattribution of research.

To tackle this challenge, we developed a method that combines feature extraction with machine learning and graph convolutional networks (GCN) [17]. First, we extract features such as co-authorship patterns, publication venues, keywords, and publication years, and apply tree-based models like XGBoost [3] to predict paper assignments. Second, we construct a GCN to model the relationships between authors and papers, using node and edge features along with graph convolutional layers to aggregate information.

Our approach integrates these two strategies through an ensemble model, combining their strengths to achieve superior accuracy in detecting misassigned papers. Our main contributions are as follows:

- Data Cleaning and Feature Extraction: We thoroughly cleaned and preprocessed the data, handling missing values and normalizing attributes. We then used various embedding models to extract features from paper titles, abstracts, keywords, institutions, authors, and publication years. This comprehensive feature extraction enabled our machine learning models, specifically XGBoost, to achieve high performance.
- Graph Convolutional Networks for Enhanced Disambiguation: We applied a graph convolutional network (GCN) to model relationships between authors and papers. Although the GCN alone scored lower, integrating its results with XGBoost improved overall performance, demonstrating the value of combining graph-based methods with traditional machine learning techniques.

These contributions advance author name disambiguation by integrating effective data processing with advanced embedding and graph-based methods, enhancing the accuracy of disambiguation systems.

2 Related Work

Author Name Disambiguation (AND) is a fundamental problem in academic information retrieval and analysis. As the volume of scholarly publications grows, distinguishing between authors with similar or identical names becomes increasingly complex. Several approaches have been explored to tackle this challenge, each with its strengths and limitations.

Content-Based Approaches: Early methods for AND focused on analyzing textual information from publication metadata. Techniques such as TF-IDF and word embeddings, like Word2Vec [4], have been employed to capture semantic similarities between paper titles, abstracts, and keywords. Recent advances in large language models (LLMs)[5, 8, 14, 10, 8, 16, 11], including BGE M3-Embedding [2] and ChatGLM3-6B [6], have further enhanced content-based methods by providing richer semantic representations and improving disambiguation accuracy through better understanding of context and language nuances.

Graph-Based Approaches: With the rise of complex academic networks[15, 9, 12, 13, 14], graph-based methods have become increasingly popular. These approaches utilize the relationships between authors, their publications, and their co-authors to disambiguate names. For instance, co-authorship networks and citation relationships provide valuable context that can help distinguish between authors with similar names based on their collaboration patterns and citation metrics.

3 Method

In this study, we developed and implemented two distinct methods to address the WhoIsWho-IND task: Extreme Gradient Boosting (XGBoost) and Graph Convolutional Networks (GCN). The task involves detecting papers incorrectly assigned to authors based on detailed attributes of the papers, including title, abstract, authors,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference'17, July 2017, Washington, DC, USA
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

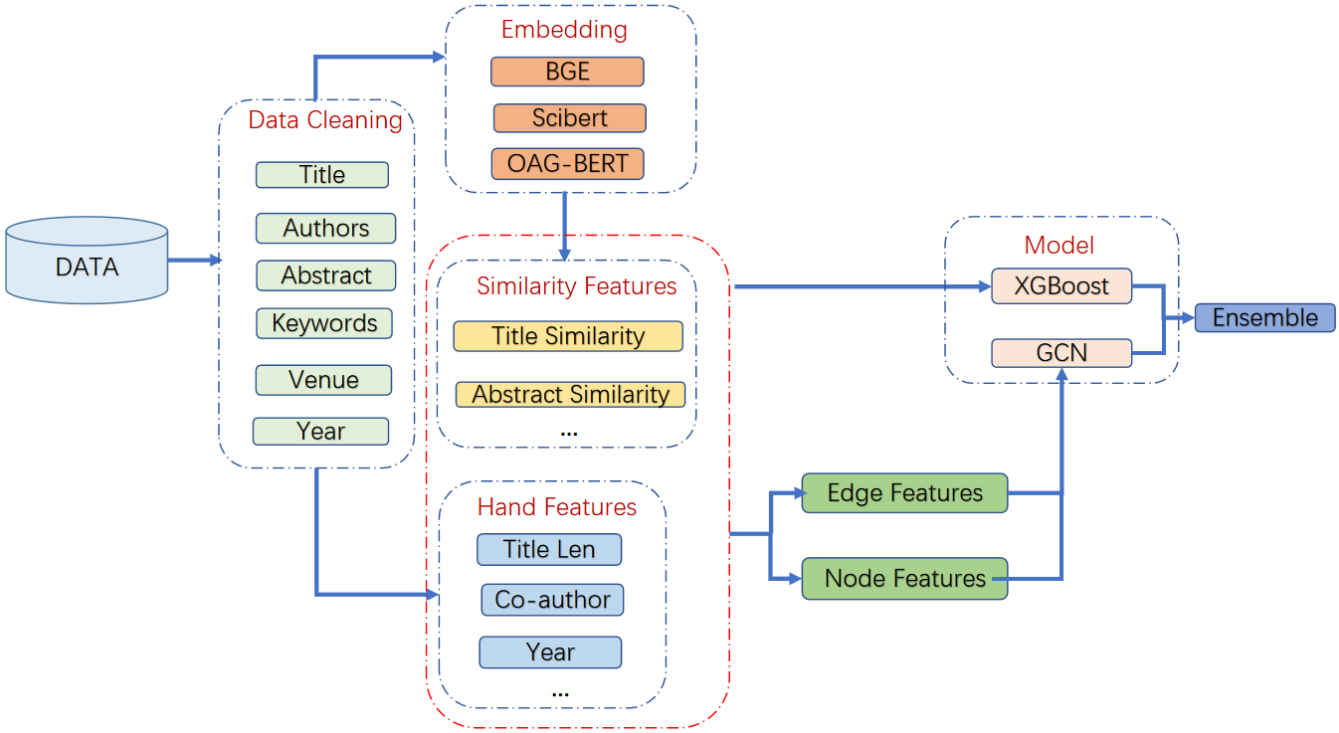

Figure 1: The illustration of our method

keywords, location, and publication year. Our approach combined traditional machine learning techniques with advanced deep learning models to leverage both the structured and relational aspects of the data.Our approach is shown in Figure 1.

3.1 Extreme Gradient Boosting

XGBoost, an efficient and scalable implementation of gradient boosting, was used for classification tasks. . The following are our feature engineering and key steps.

Data Cleaning. For the data cleaning process, we performed several steps to ensure the dataset was prepared for model training. We removed missing values and duplicates and normalized text fields such as titles, abstracts, venues, and author organizations. This involved converting text to lowercase, removing punctuation and stop words, and performing tokenization and lemmatization. Specific stop words relevant to our dataset were identified and excluded to enhance the quality of text features. The resulting cleaned and normalized data was stored in a JSON file for further processing.

Embedding. To represent the textual and metadata information of each paper, we generated embeddings using three different models: BGE, SciBERT [1], and OAGBERT [7]. For each paper, we computed embeddings for the title, abstract, keywords, and venue. The BGE and SciBERT models were used to encode these fields into vectors, while OAGBERT was utilized to create comprehensive embeddings that consider additional contextual information such as authors and affiliations. These embeddings were then stored for subsequent use in model training and evaluation.

Feature Engineering. The following list shows a subset of all features.

- total_w_co_title: the sum of the weights of the common title words of a paper with the same author and other papers.
- co_author: the average weight of the common author words in a paper with the same author and other papers.
- total_title_bge_sim: the sum of similarity is calculated by embedding the bge of a paper with the same author and other papers' titles.
- title_length: the number of words in the paper title.
- year: year of publication
- title_vector: bge embedding vector of paper title

3.2 Graph Convolutional Neural Network

Feature Extraction: The process begins with comprehensive feature extraction from each publication. Essential attributes, including the title, abstract, keywords, institution, and publication year, are extracted. These textual features are converted into numerical representations using several embedding models: OAG-BERT, Sci-BERT, and BGE-M3. OAG-BERT provides semantic embeddings tailored to academic contexts, Sci-BERT offers specialized embeddings for scientific texts, and BGE-M3 delivers high-quality semantic representations. The concatenation of these embeddings creates robust feature vectors for each paper, forming the basis for subsequent analysis.

Table 1: parameters setting of XGBoost

parameters	value
booster	gbtree
objective	binary:logistic
eval_metric	auc
max_depth	20
learning_rate	0.05
n_estimators	2900
colsample_bytree	0.9
colsample_bynode	0.9
random_state	2024
reg_alpha	0.1
reg_lambda	10
max_bin	255
tree_method	hist

Graph Construction: A graph is then constructed where each node represents a paper, and edges denote the similarity between papers. Similarity is determined by calculating the cosine distance between the feature vectors obtained from the embedding models. Edges are weighted according to these similarity scores, reflecting the potential for misassignment between papers. This graph effectively captures the complex relationships between papers, facilitating a more nuanced approach to disambiguation.

Graph Convolutional Network (GCN): A Graph Convolutional Network (GCN) is applied to the constructed graph to model the relationships between papers. The GCN aggregates information from each node's neighbors to update node features based on their contextual relationships. Through iterative graph convolutional layers, the GCN learns patterns and features from these aggregated representations. The model is trained on labeled data to classify paper assignments accurately. During the prediction phase, the GCN utilizes the learned patterns to identify potential misassignments, leveraging the graph's structure to enhance accuracy.

4 Experiment

We conducted experiments on XGBoost and GCN respectively.

4.1 XGBoost

Through a large number of experiments, we set training parameters to maximize the effect of model training as shown in Table 1.

We employed 10-fold cross-validation to train our model and assess its performance on the validation set. Additionally, we used an early stopping strategy, halting training when there was no improvement in validation performance over 100 consecutive epochs.The results of the trained model on the test set were shown in Table 2.

4.2 GCN

In this section, we describe the experiments conducted using Graph Convolutional Networks (GCNs) for author name disambiguation. We evaluated the performance of three different embedding models—BGE-Small, Sci-BERT, and OAG-BERT—by extracting features for each node in the graph.

Table 2: Experimental results of different methods on test set

Model	Score on test set
XGBoost	**0.781**
GCN+BGE-Small	0.765
GCN+Sci-BERT	0.756
GCN+OAG-BERT	0.776

Feature Extraction and Graph Construction:
For each embedding model, features were extracted from the papers using the following methods:

- **BGE-Small**: Utilized for generating embeddings that capture semantic information in a compact representation.
- **Sci-BERT**: Provided domain-specific embeddings tailored to scientific literature.
- **OAG-BERT**: Offered high-quality embeddings optimized for academic contexts.

The graph was constructed where each node represents a paper, and edges denote the similarity between papers. Similarity scores were calculated based on the cosine distance between feature vectors extracted from the embedding models. The similarity between a given paper and other papers authored by the same author was used as the edge weight, reflecting the strength of the relationship.

Training and Evaluation:
The GCN model was trained using the constructed graphs, incorporating edge weights to enhance flexibility and accuracy in handling weighted graphs. Edge weights provide additional contextual information, representing the strength or importance of connections between papers.

The performance of the GCN models was evaluated based on the accuracy of author name disambiguation. The results are summarized in Table 2.

Analysis: The results indicate that OAG-BERT features provided the most discriminative power, achieving the highest accuracy of 0.776. This suggests that OAG-BERT's embeddings are particularly effective in distinguishing between different authors based on their publication features. The performance of BGE-Small and Sci-BERT was also competitive, with accuracies of 0.765 and 0.756, respectively. These findings demonstrate the effectiveness of integrating GCNs with high-quality embedding models for author name disambiguation.

4.3 Ensemble

At the end, we integrated the results from different solutions. The ensemble model achieves the highest accuracy of 0.801, surpassing the individual performances of XGBoost and GCN models. This result highlights the effectiveness of integrating machine learning and graph-based approaches for author name disambiguation. By combining the strengths of feature extraction and relational modeling, the ensemble model provides a robust solution to the complex problem of author name disambiguation in academic publications.The results of the fusion of different proportions are shown in Table 3.

Table 3: The result of different fusion ratios

Model	Score on test set
Ensemble Model (9:1)	**0.801**
Ensemble Model (8:2)	0.796
Ensemble Model (5:5)	0.789

5 Conclusion

In this paper, we addressed the problem of author name disambiguation by combining machine learning and graph-based methods. We utilized BGE-Small, Sci-BERT, and OAG-BERT to extract features from papers and employed XGBoost for initial predictions. We then constructed a graph of papers based on their feature similarities and applied a Graph Convolutional Network (GCN) to refine the assignments. Our ensemble model, which integrates XGBoost and GCN, achieved the highest accuracy of 0.801, demonstrating the effectiveness of our approach in improving author name disambiguation.

References

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

[2] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

[3] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: a scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

[4] Kenneth Ward Church. 2017. Word2vec. *Natural Language Engineering*, 23, 1, 155–162.

[5] Zhongtian Fu, Kefei Song, Luping Zhou, and Yang Yang. 2024. Noise-aware image captioning with progressively exploring mismatched words. In *Proceedings of the AAAI Conference on Artificial Intelligence* number 11. Vol. 38, 12091–12099.

[6] Team GLM et al. 2024. Chatglm: a family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

[7] Xiao Liu, Da Yin, Jingnan Zheng, Xingjian Zhang, Peng Zhang, Hongxia Yang, Yuxiao Dong, and Jie Tang. 2022. Oag-bert: towards a unified backbone language model for academic knowledge services. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3418–3428.

[8] Fengqiang Wan, WU Xiangyu, Zhihao Guan, and Yang Yang. [n. d.] Covlr: coordinating cross-modal consistency and intra-modal relations for vision-language retrieval.

[9] Yang Yang, Zhao-Yang Fu, De-Chuan Zhan, Zhi-Bin Liu, and Yuan Jiang. 2019. Semi-supervised multi-modal multi-instance multi-label deep network with optimal transport. *IEEE Transactions on Knowledge and Data Engineering*, 33, 2, 696–709.

[10] Yang Yang, Zhen-Qiang Sun, Hengshu Zhu, Yanjie Fu, Yuanchun Zhou, Hui Xiong, and Jian Yang. 2021. Learning adaptive embedding considering incremental class. *IEEE Transactions on Knowledge and Data Engineering*, 35, 3, 2736–2749.

[11] Yang Yang, Hongchen Wei, Zhen-Qiang Sun, Guang-Yu Li, Yuanchun Zhou, Hui Xiong, and Jian Yang. 2021. S2osc: a holistic semi-supervised approach for open set classification. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16, 2, 1–27.

[12] Yang Yang, Yi-Feng Wu, De-Chuan Zhan, Zhi-Bin Liu, and Yuan Jiang. 2018. Complex object classification: a multi-modal multi-instance multi-label deep network with optimal transport. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2594–2603.

[13] Yang Yang, Yi-Feng Wu, De-Chuan Zhan, Zhi-Bin Liu, and Yuan Jiang. 2019. Deep robust unsupervised multi-modal network. In *Proceedings of the AAAI Conference on Artificial Intelligence* number 01. Vol. 33, 5652–5659.

[14] Yang Yang, Jia-Qi Yang, Ran Bao, De-Chuan Zhan, Hengshu Zhu, Xiao-Ru Gao, Hui Xiong, and Jian Yang. 2021. Corporate relative valuation using heterogeneous multi-modal graph neural network. *IEEE Transactions on Knowledge and Data Engineering*, 35, 1, 211–224.

[15] Yang Yang, De-Chuan Zhan, Yi-Feng Wu, Zhi-Bin Liu, Hui Xiong, and Yuan Jiang. 2019. Semi-supervised multi-modal clustering and classification with incomplete modalities. *IEEE Transactions on Knowledge and Data Engineering*, 33, 2, 682–695.

[16] Yang Yang, Da-Wei Zhou, De-Chuan Zhan, Hui Xiong, Yuan Jiang, and Jian Yang. 2021. Cost-effective incremental deep model: matching model capacity with the least sampling. *IEEE Transactions on Knowledge and Data Engineering*, 35, 4, 3575–3588.

[17] Si Zhang, Hanghang Tong, Jiejun Xu, and Ross Maciejewski. 2019. Graph convolutional networks: a comprehensive review. *Computational Social Networks*, 6, 1, 1–23.