Leveraging Diverse Text Embeddings and Mediums for Incorrect Assignment Detection

Lingrui Zhang
SouthEast University
Nan Jing, China
220224398@seu.edu.cn

Yuxin XIE
Shanghai Normal University
Shanghai, China
1000527330@smail.shnu.edu.cn

Abstract

Academic knowledge graph mining seeks to enhance our understanding of scientific evolution and trends, unlocking substantial potential for guiding policy, facilitating talent discovery, and advancing knowledge acquisition. However, the field's progress is hindered by the absence of standardized benchmarks. To address these issues through tasks focusing on name disambiguation complexities and developing models to detect misattributed papers, leveraging detailed paper attributes We used diverse text embedding methods to extract semantic features of paper attributes, and established an isomorphic graph structure based on the connections between papers to capture potential associations between different papers. By integrating the tree-base model and the graphsage model achieved 5th place in WhoIsWho-IND-KDD-2024 competition.

CCS Concepts

• **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

Keywords

Incorrect Assignment Detection,text embedding, GBDT,GNN

ACM Reference Format:
Lingrui Zhang and Yuxin XIE. 2018. Leveraging Diverse Text Embeddings and Mediums for Incorrect Assignment Detection. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Deep mining of academic data can enhance understanding of the development, nature, and trends of science, encourage researchers to share their achievements and experiences, and promote the growth of the entire academic community. Paper author homonym disambiguation is a significant challenge in academic search systems. Improving the accuracy of existing disambiguation systems will help clarify researchers' academic contributions, enhance academic retrieval results, and foster academic collaboration and scientific policy-making.

1.1 Dataset

A unique and high-quality academic graph mining dataset-WhoIsWho-IND[6], is provided by Tsinghua University's Knowledge Engineering Group (KEG) and Zhipu AI. The training set includes authorsID, the IDs of papers correctly attributed to authors, and the IDs of papers incorrectly attributed to them. Additionally, each paper provides information such as the title, author names, author affiliations, venue, publication year, keywords, and abstract.

1.2 Task

Based on the given author and paper information in the training set, participants need to train a model to detect papers that are incorrectly attributed to authors in the test set. The authors included in the test set should be independent of those in the training set. The model's performance will be measured using the AUC, which is widely adopted in anomaly detection. For each author:

$$\text{Weight} = \frac{\#\text{TotalErrors}}{\#\text{ErrorsOfTheAuthor}}$$

For all authors(M represents the number of authors) :

$$\text{WeightedAUC} = \sum_{i=1}^{M} \text{AUC}_i \times \text{weight}_i$$

2 RELATED WORK

Mainstream methods for homonym disambiguation include content-based and graph-based approaches. Content-based methods use techniques such as Word2vec[2] and TF-IDF[4] to extract text vectors from publication metadata, determining whether pairs of publications relate to the same author. Recently, advancements in large language model (LLM) technology, such as BGE M3-Embedding[1] and ChatGLM3-6B[3], have significantly enhanced natural language processing. These models capture higher-quality semantic information, offering new possibilities for improving performance in IND tasks. Graph-based methods, on the other hand, focus on capturing the structural relationships within the academic knowledge graph[5], such as collaboration and citation relationships between authors, to aid in homonym disambiguation.

3 SOLUTION OVERVIEW

In this section, we outline the key components of our solution. We operate under the assumption that for each author, the majority of papers assigned to them constitute their related papers. Hence, we treat the competition as an anomaly detection problem where

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

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

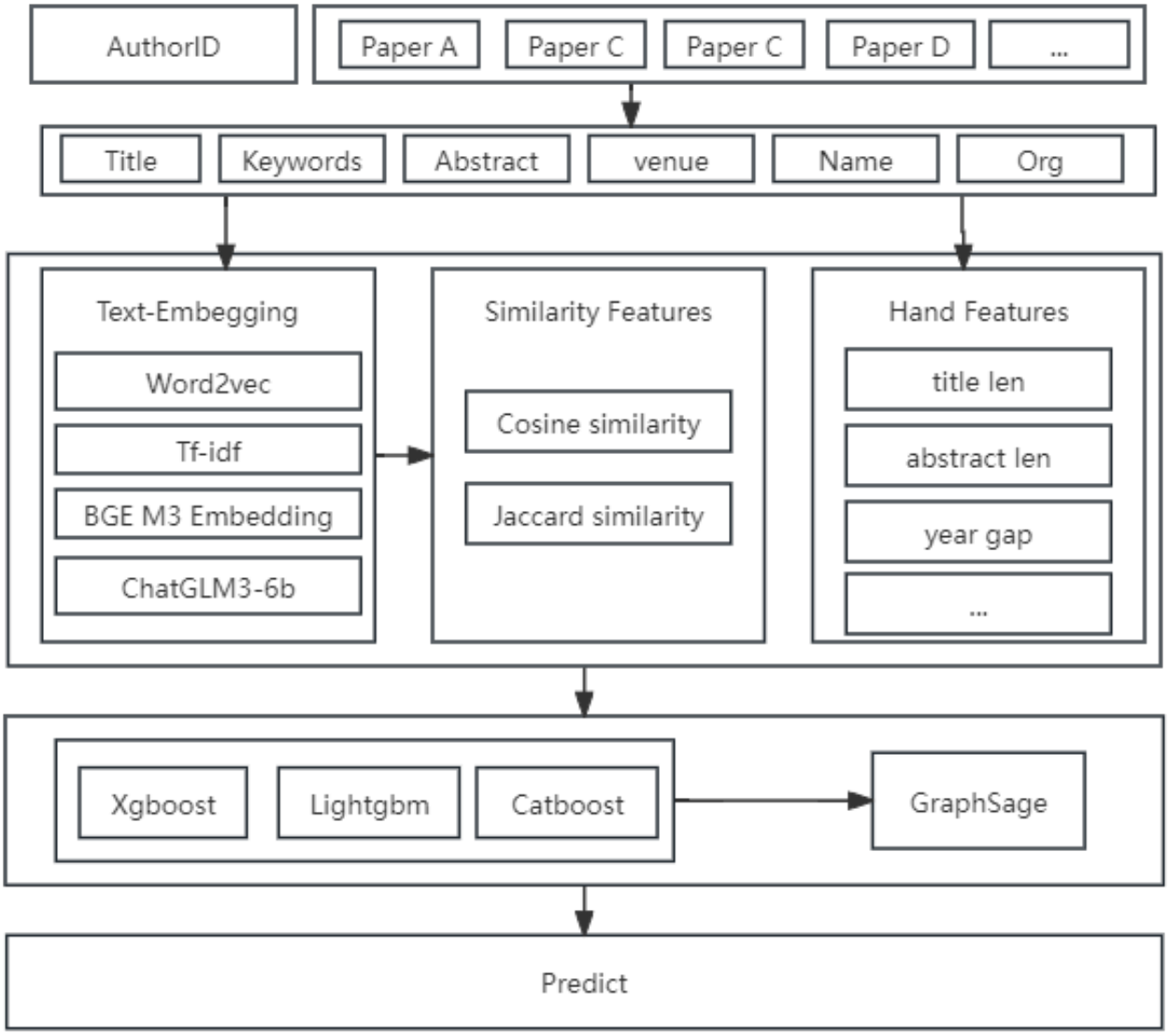

Figure 1: Solution architecture

we determine whether each paper assigned to an author belongs to a major paper cluster. The overall architecture of our model is illustrated in Figure. 1, which is divided into three main parts. Firstly, we utilize a variety of Text-Embedding methods to extract semantic representations of papers, including Word2vec, Tfidf2vec, BGE-M3-Embedding, and ChatGLM3-6b. Secondly, we conduct extensive feature engineering efforts, focusing on similarity features between paper pairs and manually crafted features that characterize paper attributes such as length and author order. Finally, we employ decision tree-based models such as LightGBM, XGBoost, CatBoost, as well as models based on graph neural networks like GraphSAGE in the third part.

4 DETAILED METHOD

4.1 Data Preprocessing

In this section, we focused on text cleaning within paper titles and abstracts by removing low-frequency special characters and converting all text to lowercase. Additionally, to mitigate potential performance impacts from variations in author name formats across different papers, we conducted fuzzy matching and combined author names within papers associated with the same autherID.

Modules	Local weighted AUC	Leaderboard weighted AUC (Phase)	Leaderboard weighted AUC (Phase)
Baseline	0.652876	0.61130591	-
Baseline+Hand Features	0.743627	0.65929681	-
Baseline+Similarity Features	0.846354	0.7316034	-
Baseline+All features	0.854505	0.75628129	0.78125766
Emsemble tree-based models	0.861291	0.76112896	0.79945476
Emsemble tree-based models + Graphsage	0.897844	0.77084034	0.80720296

Table 1: Performance comparison of modules

4.2 Features

Enhancing the semantic information of thesis-related elements forms the foundation of our approach, with high-quality thesis text vectors serving as the cornerstone. Our solution incorporates various word vector methodologies: (1) Tf-idf measures the importance of a term in a document by considering its frequency within the document and across a collection of documents. (2) Word2vec is used to generate distributed representations of words in a continuous vector space based on their context in a corpus.,(3) bge-m3-embedding, and (4)chatglm-3.

A direct approach to assess the inclusion of a given paper within the main paper cluster involves computing the cosine distance between the paper in question and others authored by the same author. Specifically, we calculate the cosine distance for each pair of papers based on each type of text vector, as well as the Jaccard similarity at the word level. The extracted multiple text embedding are combined with other manual features describing the paper attributes as model input to characterize the basic properties of the paper.

4.3 Model

Tree-based models such as LightGBM, XGBoost, and CatBoost demonstrate strong performance by transforming unstructured text data into structured tabular data through feature engineering. These models excel in handling feature-rich and complex datasets. Given the significant score variability observed in our results, even with grouped cross-validation by author ID, we adopt an ensemble approach using all three models simultaneously to enhance the stability and robustness of our overall scheme.

GraphSAGE is a classical domain-agnostic graph neural network algorithm designed specifically for processing large-scale graph data. It efficiently learns node representations by aggregating neighbors on graph sub-samples, eliminating the need for full graph traversal. This characteristic enables GraphSAGE to excel in tasks involving graph structures. Treating each paper as a node, we establish edges between nodes based on relationships such as shared co-authors, institutional affiliations, and intersecting paper keywords. This approach constructs a homogeneous graph structure where the previously extracted features serve as node features. By aggregating information from neighboring nodes, GraphSAGE further explores potential relationships between papers, thereby enhancing error detection accuracy.

4.4 Training

We utilize one GTX 3090 for both feature extraction and model training. For methods such as Word2vec and Tfidf2vec, we perform training and inference on the entire paper dataset. Conversely, for models like BGE M3-embedding and ChatGLM3-6b, we rely solely on open-source pre-trained weights for inference without fine-tuning, primarily due to computational resource limitations. The extraction of text-embedding consumes approximately 300 minutes, while other feature computations, particularly paper similarity assessments, take about 60 minutes. Training tree-based models and graph neural network-based models also require approximately 120 minutes.

5 EXPERIMENT

5.1 Validation Strategy

The competition comprises a training dataset and two test datasets in the online phase, strictly partitioned based on author ID. To replicate this setup, we partitioned the groups offline by author ID and conducted cross-validation across these groups. However, we encountered inconsistent results in ablation experiments across the training set and the two test datasets. These discrepancies may stem from inherent distributional variations in data across author IDs or differences in data acquisition sources across different phases.

5.2 Results and Comparison

In this section, we compare the performance of various modules across three datasets. Our results demonstrate that diverse methods of text-embedding extraction can mutually complement each other, thereby enhancing semantic comprehension. Furthermore, establishing distinct node relationships enriches interconnections among papers. Through meticulous feature extraction, tree-based models outperform graph neural networks. However, graph neural networks offer richer connections within thesis graph relations, making them valuable as part of an ensemble approach.

6 CONCLUSION AND FUTURE WORK

We present our solution in the KDDCUP-2024 OAGchallenge-WhoisWho incorrect assignment detection track. Leveraging diverse text-embedding methods effectively characterizes the semantic information of papers. Papers are interconnected based on author information, institutional affiliations, and paper keywords. The integration of tree models and graph neural networks markedly enhances in accuracy. Nonetheless, there are opportunities for further enhancement, such as fine-tuning the language model (LLM) or implementing end-to-end paper attribution detection model.

7 ACKNOWLEDGMENTS

We express our gratitude to Tsinghua University, Knowledge Engineering Group (KEG), and Zhipu AI for organizing the KDD Cup 2024 OAG-challenge. Participating in this challenging task was a rewarding experience for us. We also extend our thanks to all the participants whose competitive spirit kept us motivated and engaged throughout the competition.

References

[1] 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*.

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

[3] 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*.

[4] Thorsten Joachims et al. 1997. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In *ICML*. Vol. 97. Citeseer, 143–151.

[5] Baichuan Zhang and Mohammad Al Hasan. 2017. Name disambiguation in anonymized graphs using network embedding. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 1239–1248.

[6] Fanjin Zhang et al. 2024. Oag-bench: a human-curated benchmark for academic graph mining. *arXiv preprint arXiv:2402.15810*.