Journal of Computers Vol. 34 No. 4, August 2023, pp. 41-52
doi: 10.53106/199115992023083404004

Author Name Disambiguation Based on Heterogeneous Graph

Chuang Ma*, Helong Xia

School of Software Engineering, Chongqing University of Posts and Telecommunications,
Chongqing 400065, China

903294996@qq.com,s211201026@stu.cqupt.edu.cn

Received 12 July 2022; Revised 24 November 2022; Accepted 4 January 2023

Abstract. Since multiple people share the same name in the real world, this will cause performance degradation to academic search systems and lead to misattribution of publications. The author name disambiguation algorithm has not yet to be well solved. In this paper, we propose a disambiguation method that combines heterogeneous graph-based and improved label propagation, first we construct a publication heterogeneous graph network, then graph neural networks is applied to aggregate the nodes representation and relation types, finally combined with the improved label propagation algorithm to realize clustering. The task of author name disambiguation is completed to improve the retrieval performance. Experimental results on two public datasets show that our method was improved by 2.8% and 4.9% over the suboptimal method, respectively. Our method can effectively reduce the number of publications returning the wrong author and improve the performance of the academic retrieval system.

Keywords: data mining, author name disambiguation, disambiguation, heterogeneous graph, deep learning

1 Introduction

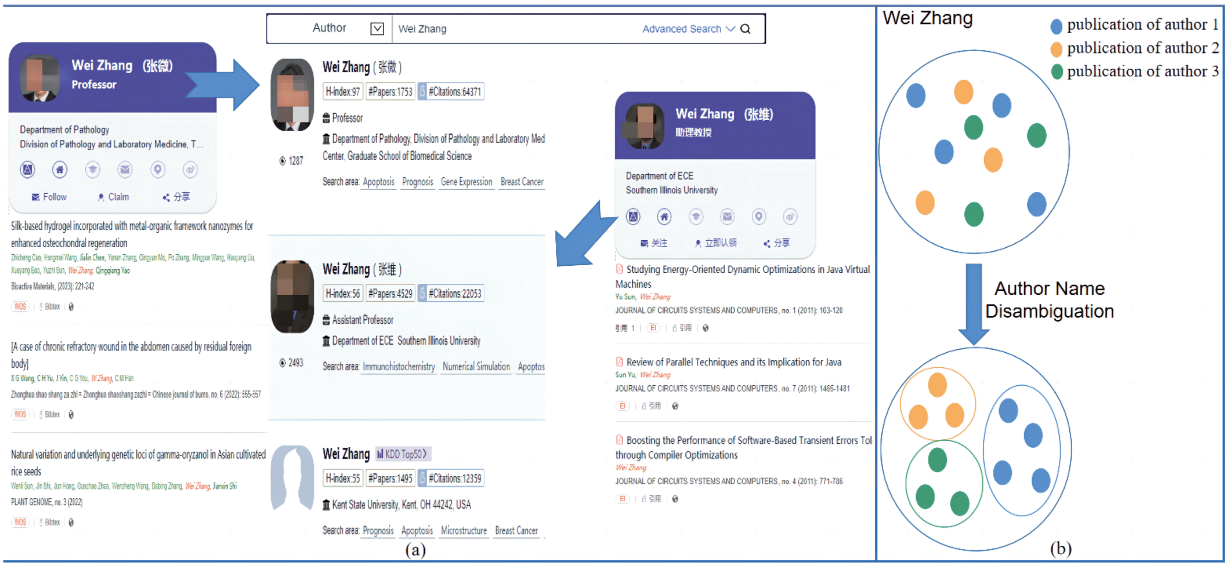
The rapid growth of the world population has led to many people sharing the same names. Some research [1] shows that 1.1 billion people share 90,000 names. In other words, about 12,000 people have the same name, which will lead to the inconvenience of the academic search system. For example, the database of the academic search system stores many publications with the same author's name, and the staff must manually distinguish these publications. Fig. 1(a) shows the results of the disambiguation named Wei Zhang. When one wants to search the publications of Wei Zhang scholars, the publications that have already been grouped will be returned, and such a result will make the usage more efficient. The process of distinguishing these publications with the same author's name is called AND (the abbreviation of Author Name Disambiguation), and it can be defined as a clustering task with an unknown size of clusters. As in Fig. 1(b), the small circles represent the set of all publications with the author's name Wei Zhang. The disambiguation process is to divide the publications set into multiple clusters, and each cluster represents all papers of a real author with the name Wei Zhang. Many methods of dealing with the AND have been proposed, while many challenges exist.

The first challenge of AND is to obtain accurate publication representations. Some studies, such as [1, 2], only utilize the topology of the publication network and do not consider the publication's title, abstract, and other attribute information, which is not conducive to obtaining accurate node representations. The general approach is to use the attribute information of the publication as the initial node representation and then combine it with the topological features of the publication network to get a better node representation [3-6]. However, the features of relations in a publication heterogeneous graph should be considered. The semantic information of the relations can enrich the node features and get better node representations. Some methods depend on feature engineering, which would spend much precious manpower to mine features.

The second challenge is determining the number of distinct authors with the same name, that is, to determine the size of clusters. Since common clustering methods that require a specified size of clusters cannot be applied to this task, many methods similar to hierarchical clustering and AP clustering are used [3, 7]. However, these clustering algorithms also require specifying the relevant parameters that control the size of clusters. The optimal parameters vary from different data, which will take much time to choose the parameters. The number of distinct authors could be estimated by neural networks [8] from labeled data. However, labeled data are scarce and difficult to obtain in large quantities. The community discovery algorithm can perform community partitioning of the

* Corresponding Author

nodes of the graph, which does not require any additional parameters to be specified. Therefore, it can be applied to the publication graphs to obtain the clustering partitioning of the publications.



Author ☑ Wei Zhang Advanced Search ∨ Q

Wei Zhang (张微)
Professor

Department of Pathology
Division of Pathology and Laboratory Medicine, T...

Follow · Claim · 分享

Silk-based hydrogel incorporated with metal-organic framework nanozymes for enhanced osteochondral regeneration
Zhicheng Cao, Hongmei Wang, Jialin Chen, Yanan Zhang, Qingyun Mo, Po Zhang, Mingyue Wang, Haoyang Liu, Xueyang Bao, Yuzhi Sun, Wei Zhang, Qingqiang Yao
Bioactive Materials, (2023): 221-242
WOS | Bibtex |

[A case of chronic refractory wound in the abdomen caused by residual foreign body]
X G Wang, C H Yu, J Yin, C G You, W Zhang, C M Han
Zhonghua shao shang za zhi = Zhonghua shaoshang zazhi = Chinese journal of burns, no. 6 (2022): 555-557
WOS | Bibtex |

Natural variation and underlying genetic loci of gamma-oryzanol in Asian cultivated rice seeds
Wenli Sun, Jin Shi, Jun Hong, Guochao Zhao, Wensheng Wang, Dabing Zhang, Wei Zhang, Jianxin Shi
PLANT GENOME, no. 3 (2022)
WOS | Bibtex |

Wei Zhang (张微)
H-index:97 | #Papers:1753 | #Citations:64371
Professor
Department of Pathology, Division of Pathology and Laboratory Med Center, Graduate School of Biomedical Science
1287
Search area: Apoptosis Prognosis Gene Expression Breast Cancer

Wei Zhang (张维)
H-index:56 | #Papers:4529 | #Citations:22053
Assistant Professor
Department of ECE Southern Illinois University
2493
Search area: Immunohistochemistry Numerical Simulation Apoptos

Wei Zhang KDD Top50 〉
H-index:55 | #Papers:1495 | #Citations:12359
Kent State University, Kent, OH 44242, USA
Search area: Prognosis Apoptosis Microstructure Breast Cancer

Wei Zhang (张维)
助理教授

Department of ECE
Southern Illinois University

关注 · 立即认领 · 分享

Studying Energy-Oriented Dynamic Optimizations in Java Virtual Machines
Yu Sun, Wei Zhang
JOURNAL OF CIRCUITS SYSTEMS AND COMPUTERS, no. 1 (2011): 103-120
引用 1 | EI | 引用 |

Review of Parallel Techniques and its Implication for Java
Sun Yu, Wei Zhang
JOURNAL OF CIRCUITS SYSTEMS AND COMPUTERS, no. 7 (2011): 1465-1481
EI | 引用 |

Boosting the Performance of Software-Based Transient Errors Tol through Compiler Optimizations
Wei Zhang
JOURNAL OF CIRCUITS SYSTEMS AND COMPUTERS, no. 4 (2011): 771-786
EI | 引用 |

(a) The result of author name disambiguation (b) The clustering of author name disambiguation

Fig. 1. The result

(Author name disambiguation: There are multiple authors with the name Wei Zhang in the system, and papers written by different authors are automatically grouped.)

The third challenge is incremental disambiguation, where new publications in the academic system need to be categorized in corresponding author documents. In [3], an incremental disambiguation method is proposed to obtain new node information by sampling random paths and categorizing the new publication nodes. The algorithm that can achieve incremental disambiguation has better practicality.

In this paper, we proposed an author name disambiguation method based on a combination of heterogeneous graphs and improved label propagation. First, we constructed a publication heterogeneous graph network. Then, graph neural networks are applied to aggregate node representations and the features of relation types, and finally, an improved label propagation algorithm is used to achieve clustering. The main contributions of this publication include the following aspects: (1) We used multidimensional features to construct a publication heterogeneous graph. To preserve the structural information and simplify the computation of processing the constructed heterogeneous graph, we fused the relation features between publications to the feature aggregation process. (2) We propose an improved label propagation algorithm for the disambiguation task without specifying the size of clusters. It could achieve better clustering by choosing the direction and priority of label propagation.

The remaining sections are organized as follows: the related works about author name disambiguation are presented in section 2. Section 3 explains our proposed method for (1) initializing the publication node features. (2) constructing graphs from the information of publications. (3) the final representation of nodes obtained by aggregating node information. (4) the objective function to be optimized. (5) description of the improved label propagation algorithm. Section 4 mainly describes the experiments we have done. Section 5 summarizes the methods made in this paper and future work.

2 Related Work

Author name disambiguation has been an important topic in information retrieval, and researchers have proposed many methods to solve it, which can be divided into supervised learning and unsupervised learning.

Supervised methods require expensive manually labeled labels. Moreover, due to the uneven distribution of the number of author's publications, these methods may fall into the imbalance problem [8-10]. Han et al. [9] assumed that the number of authors is known and use statistics models for supervised training. The diversity of the data is also important, so [10] introduced other databases to enhance the features of the data and achieve better performance. Chen et al. [11] used attribute information like institution, coauthors, years as the criteria for the uniqueness of the author, and then adopt the clustering merging strategy to improve the model recall rate. Silva et al. [12] proposed a new feature extraction method for the AND task, which can extract the local features associated with the author and the publications. With the development of Generative Adversarial Networks (GAN) in the computer vision field, it was applied to the AND task by some searchers [13, 14]. [13] was the first to obtain node representations in a heterogeneous graph with GAN and applied it to AND. Later, a novel generative adversarial network for disambiguation to mine higher-order relationships between publications was proposed [14].

Unsupervised methods aim to divide the publications into multiple clusters, and each cluster represents the publications of a distinct author. Fan et al. [1] compared the similarity of the publications by designing paths to obtain the similarity matrix of publications. Some researchers wanted to protect the author's privacy, so anonymized graphs were constructed without the text information of publications to process the task [2]. In contrast, in this paper, the initial representation of publication nodes was denoted as the text of publication. Since the HDBSCAN clustering algorithm is inclined to cluster nodes into a few clusters, the HDBSCAN algorithm was used to combine with the AP algorithm [15]. Peng et al. [16] improved [15], increase the information types, and propose a semi-supervised clustering algorithm. The Graph Auto Encoder, as a classical unsupervised feature extraction algorithm, was also used to extract publication features by researchers [7], and they used the HAC algorithm in clustering. However, it could not mine a deep relationship between publications because the homogeneous graph can not preserve a complex relationship. At the same time, our method constructed a heterogeneous graph to preserve it. Xiong et al. [4] obtained the result of clustering by converting heterogeneous graph to homogeneous graph and using semantic and relationship joint embedding module. The unsupervised algorithms mentioned above have parameters associated with the size of clusters. These parameters may be different on different datasets, which will consume valuable manpower to find the optimal parameters. At the same time, some researchers are concerned about determining the size of clusters. For example, Qiao et al. [3] proposed a modularity-based clustering method to minimize the modularity to achieve clustering without specifying the size of clusters. In contrast, in our work, an improved label propagation algorithm was used in clustering, which also could automatically determine the size of clusters. Zhang et al. [7] propose an end-to-end method for predicting the size of clusters based on neural networks, which can alleviate the poor performance caused by clustering size errors. Jie Tang et al. [5] utilize the X-means algorithm to determine the size of clusters, but this method tends to merge clusters on large data sets, resulting in a decrease in accuracy. In summary, most past work has focused on obtaining high-quality publication representations, while a few methods focusing on clustering are computationally complex [3, 5, 7]. Our method can obtain better publication representations and cluster results with low complexity.

3 Proposed Method

The framework we proposed is shown in Fig. 2, divided into three parts: a, b, and c. Part a focus on obtaining node representations, specifically initializing publication node representation, constructing publication heterogeneous graphs, and aggregation of node features and relation features. In Part c, meta paths are used to calculate the objective function according to the representation of publication nodes. Part b mainly uses the topological properties of the graph to realize unsupervised clustering.

3.1 Initialize Node Representation

Doc2vec [17] model is an unsupervised algorithm, which can learn fixed length feature representation from variable length text (such as sentences, paragraphs or documents). So, we utilize Doc2Vec to train publications' titles and abstracts to obtain the initial node's representation. It will be used as input to the feature aggregation later.

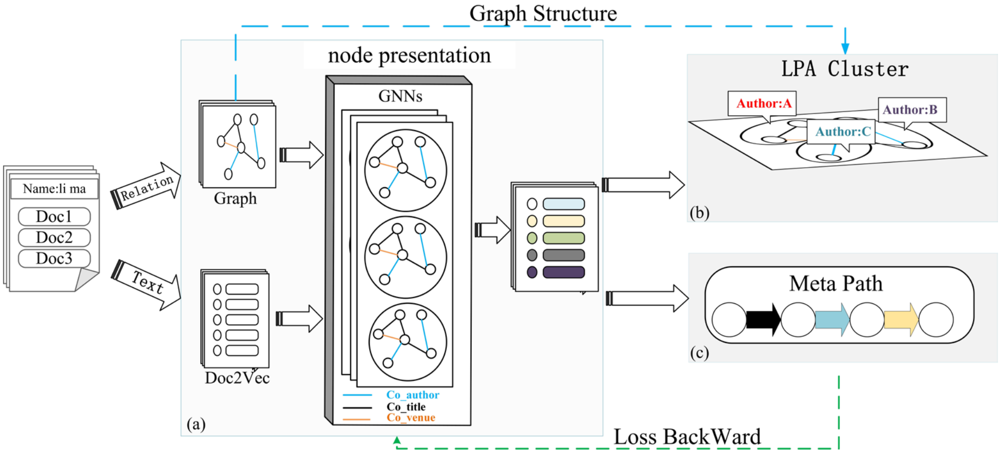


Fig. 2. The illustration of our proposed framework

3.2 Construct Publication Heterogeneous Graph

Heterogeneous graphs contain different types of nodes and edges, which can explain and preserve the relationship between nodes and capture more features. Therefore, we will construct a publication heterogeneous graph to explore more relationships and features between publications.

Given an ambiguous author name a , we construct a publication heterogeneous graph $G^a = (V, E, T)$. V is the node set, each node $v_i \in V$ denotes the different publication with author name a . E is the edge set, which preserve the connection relationship between publication nodes. Finally, T is the set of relation types, and it includes coauthor, covenue, and cotitle relation type. The relation types between node v_i and node v_j can be expressed as T_{ij} and the details are as follows:

(1) Coauthor: The relation exists if there is an overlap between the authors of two publications (excluding a), the weights of the relation are the number of overlaps.
(2) Covenue: The relation exists if two publications are published in the same journal.
(3) Cotitle: The relation exists if the title sets of two publications are overlapped after removing stop words, the weights of relation are the number of overlaps.
(4) Self-Loop: self-loop for each node

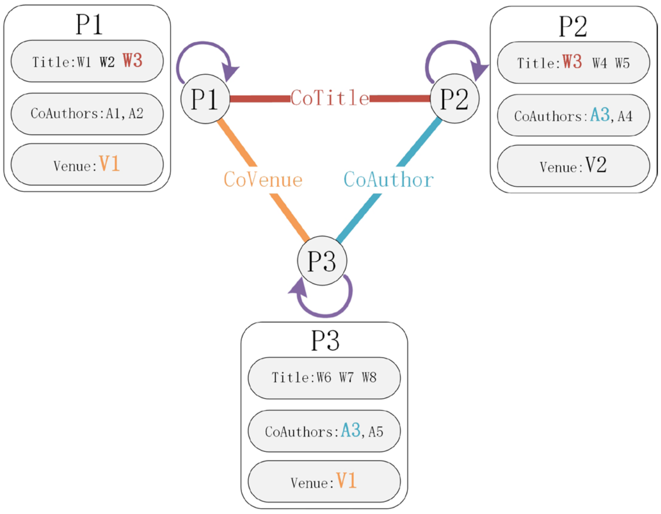


Fig. 3. Construct graph

As shown in Fig. 3, the title of publication P1 is w1w2w3, and the title of publication P2 is w3w4w5. There are common fields in the titles of publication P1 and publication P2, and the number is 1. Therefore, a cotitle relation should be established between P1 and P2, and the weight is 1. The publications of P1 and P3 were published at the V1 conference, so the covenue relation should be established between P1 and P3 and the weights is 1. The coauthors of publication P2 are A3 and A4, and the coauthors of publication P3 are A3 and A5. There is overlap between the coauthors, so coauthor relation between P2 and P3 is also established, with a weight of 1. Finally, a self-loop will be in each publication. Through the above steps, we have established a publication heterogeneous graph with the number of node types being 1 and the number of relation types being 3. It will be used in the following features aggregation.

3.3 Aggregation of Node Features

Many previously proposed graph-based author disambiguation methods [1, 2, 15] do not take into account the relation types, which does not take full advantage of the topological structure information of the graph. So, we fuse the features of relations to the feature aggregation in our experiments to obtain a better representation. The details are as follows. First, we map the relation T_{ij} between node v_i and node v_j as representations z_{ij}^T . However, the relations between two publications may be composite, i.e., there is both a coauthor relation and a covenue relation between two publications, we average all the relation representations as the final relation representations. Specially, we set $|T_{ij}|$ to be the number of relations between v_i and node v_j and the final representation of the relation e_{ij} can be described as:

$$e_{ij} = \frac{\sum_{t \in T_{ij}} z_{ij}^t}{|T_{ij}|} . \tag{1}$$

Then the node representations can be obtained by aggregating the neighbor's node representation and the relation representation, as following formula:

$$h_{\text{i}}^l = \delta(\sum\nolimits_j^n a_{ij}(W_s^{(l)} s_{ij}^{(l-1)} + b^{(l)})) . \tag{2}$$

$$s_{ij}^{(l-1)} = Concat(h_j^{(l-1)}, W_e^{(l-1)} e_{ij}) . \tag{3}$$

Where h_i^l is the representation of the node v_i in the l-th layer, and h_i^0 is the initial feature of the publication node obtained from Section 3.1. $b^{(l)}$ is the bias of l-th layer and a is the normalized adjacency matrix of the weighted graph. δ is an activation function of Relu and e_{ij} represents the representation of relation between node v_i and v_j. The node representation thus obtained not only aggregates the features from the neighbor nodes themselves, but also aggregates the features of the relations. It could enrich the obtained representation. Fig. 4 is a diagram of the process.

3.4 Objective Function

The meta-path can capture the correlation of different nodes through heterogeneous relationships [18], so minimizing the distance between adjacent nodes on the meta-path is considered as the optimization objective. As shown in Fig. 5, path (P1 → P2) can be formed through coauthor relation between node P1 and P2, which means node P1 and node P2 were written by a same coauthor, and it may increase the probability that these two publications belong to the same author.

This paper defines a relationship sequence *coauthor → covenue → coauthor → cotitle*, which means meta-path (P1 → P2 → P3→ P4 → P5) is defined. We minimize the distance of nodes in a window.

$$loss = \log\sigma(h^v \cdot h^u) + \sum_{i=1}^{m} \log\sigma(-h^{z^i} \cdot h^v) . \tag{4}$$

Where u is the target node, v is the neighbor node of the node u. m is the number of negative samples and z^i is the i-th negative sampling node. The final representation of each node can be obtained by using gradient descent to optimize the above loss function. and the node representation contain the network topology information and the text information.

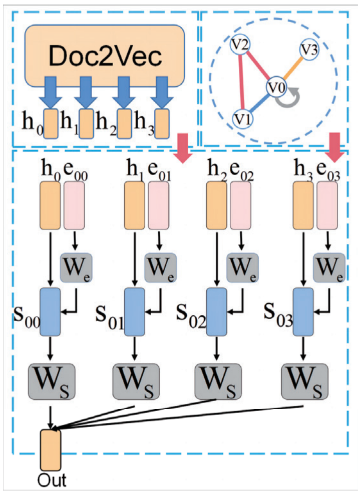


Fig. 4. The diagram of proposed features aggregation

(The nod V0 representation are obtained by aggregating neighbor node's relation features and node features.)

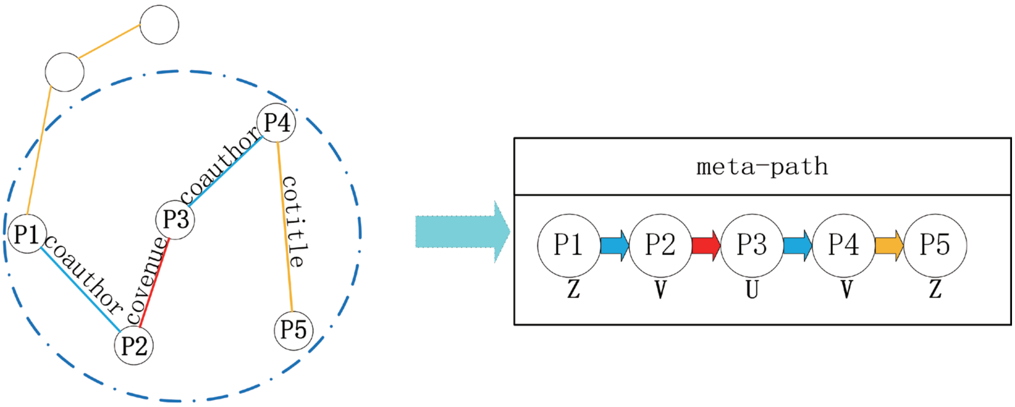


Fig. 5. Meta-Path

3.5 Clustering Partition

After the objective function is optimized to obtain the final representations of the node, the nodes need to be clustered, and how to determine the size of clustering is very important. HAC clustering algorithm is often used in past words, because it does not require specifying the size of clusters. However, it requires manually determining the parameters associated with the size of clusters. Moreover, the optimal threshold parameters may vary from one dataset to another, which can be time-consuming to manually tune the parameters on the data. Therefore, we use the label propagation algorithm to automatically determine the size of clusters by using the topological structure of the publication graph and node representations. We also improved the standard label propagation algorithm so that it can be applied in the heterogeneous graph of publications.

The idea of the standard labeling algorithm is to use labeled nodes to predict the label information of unlabeled nodes. The algorithm steps are simple, and as follows.

(1) Each node is given a unique label during initialization.

(2) Update the node's labels. Specifically, select the labels that appear most in neighboring nodes as the node's label. If the labels appear the same times, one is randomly selected.

(3) Continue to perform step (2) until there is no node label to update or iteration finished.

However, the standard label propagation algorithm has the disadvantages of unstable results, strong randomness, and low accuracy. These drawbacks arise from the uncertainty of the propagation direction and the lack of consideration of the correlation of nodes. So, to make it have a better effect in this experiment, we improve the algorithm as follows.

For the uncertainty of the propagation direction, we make its label propagate in a fixed direction. Intuitively, the importance of different relation types in label propagation is different. For example, if a publication node has a relation of coauthor and a relation of cotitle type, the label should be propagated from the relation of the coauthor type. Therefore, priorities about relations should be established to make the label propagate from a more important relation.

For considering correlation of nodes, a node may have multiple relations of the same type, and the relations connect its different neighbors. When this situation exists, the similarity between neighboring nodes and this node is calculated, and the node with higher similarity is selected for label propagation.

As shown in Fig. 6, the label of node h0 is propagated from node h1, h2, h3, h4, and h5. Since the predefined coauthor edge-type priority is higher than the covenue and cotitle edge-type priority, only nodes h1 and h2 can propagate labels to node h0. Moreover, the similarity between node h1 and node h0 is greater than the similarity between node h2 and node h0, the label of node h0 is updated with the label of node h1. Note that the node representation for calculating the similarity is obtained from the feature aggregation in the previous section.

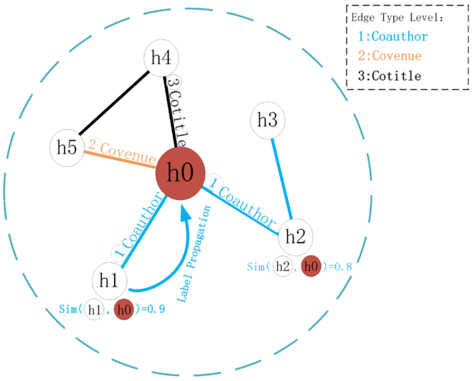


Fig. 6. The process of label propagation algorithm

4 Experiment

We perform several experiments on datasets to evaluate our proposed method. The results of experiments show that our method has advantages in performance compared with other methods.

4.1 Datasets

We conduct experiments on two benchmark datasets: Aminer and ECNU (the abbreviation of East China Normal University disambiguation datasets). The Aminer disambiguation dataset contains 110 disambiguation names, 1,723 distinct authors, and 8,505 publications. Each publication contains data such as title, year, venue, authors, etc. The ECNU contains 110 disambiguation names, 31822 distinct authors, and 152256 publications. Each publication includes abstract data in addition to the above data in Aminer. The sample of the Aminer is show as

Table 1 below.

Table 1. Sample data from Aminer

Filed	Content
Title	Editorial: introduction to the special issue on innovative applications of computer vision
Authors	Name: Manish Bhide Name: Alok Gupta
Org	Liver Cancer Institute and Zhongshan Hospital, Fudan University, Shanghai, China Siemens Corporate Research Inc., David Michael, Cognex Corp.
Venue	Machine Vision and Applications
Year	2022

4.2 Baselines

To evaluate the effectiveness of our method, we compared it with other methods.

HAC: The representation of nodes is obtained by training the title, abstract, venue of publications with the Doc2Vec model. Then the nodes are clustered by HAC clustering algorithm.

Metapath2Vec [18]: Methpath2vec is widely used in graph representation learning on heterogeneous graph. The meta-paths are obtained according to the predefined relationship sequence to produce node representation.

Zhang et al. [2]: The author-author graph, publication-publication graph, and author-publication graph are constructed, and minimize the distance between adjacent nodes on these three graphs to obtain the node representation.

PHNet [3]: They construct heterogeneous graphs according to publications, then the PHGCN is applied to the graphs.

GAE: Autoencoder extensions on the graph to obtain node embeddings by reconstructing the graph structure and input features.

We set the dimension of Doc2Vec model as 64, and the dimension of representation of relation type is 32. The relationship sequence of meta-path is defined as *coauthor* → *covenue* → *coauthor* → *cotitle*. Finally, the size of the window is 5, the number of negative samples is 5.

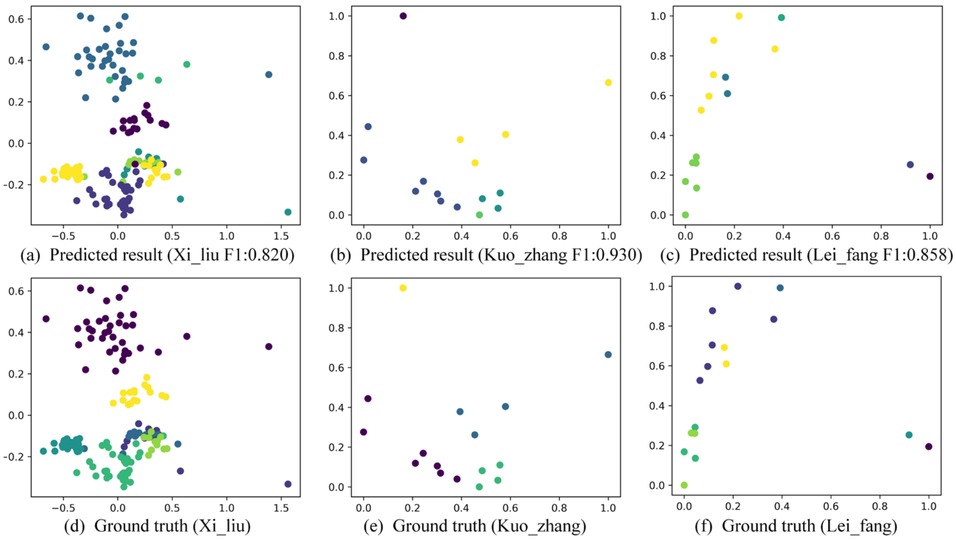
4.3 Experimental Results

Table 2 and Table 3 show the performance of our and other methods on the dataset. For example, for the publications with the author's name Kuo Zhang on the Aminer dataset, 70% of the publications belong to 1 distinct author, and it falls into an unbalanced distribution. However, our method achieves a 93.0% F1 score on it, which shows that our method can handle unbalanced data well. The F1 score on all names in each dataset (shown in the last row) indicates that our method outperforms all the baselines (+2.2-35.4% in Aminer dataset, +4.9-35.2% in ECNU dataset).

Our method can outperform the other methods because we can obtain high quality node representations of heterogeneous networks. Among the algorithms we compared, the HAC algorithm only utilizes attribute information, such as the publication's title and abstract. It lacks topological structure features and causes terrible results, about 30% lower than our algorithm. Metapath2vec [18] only utilizes the topological structure of the network and does not fully exploit the textual information. So, it leads to low performance, with only 56.4% F1 score in the Aminer dataset and 45.7% F1 score in the ECNU dataset. The performance of Metapah2vec [18] algorithm is higher than that of the HAC algorithm, which means that topological structure features of the graph network is more effective than textual features for the author's name disambiguation task. Zhang et al. [2] used multiple relationships to build multiple graphs and then used a network embedding method to obtain a representation of the publication. However, this method reaches only 76.1% F1 score in the Aminer dataset and 45.7% F1 score in the ECNU dataset. The reason for this is that privacy is protected and the textual information of the publication is missing. Graph Autoencoder reconstructs the textual features of the input publication and the network structure by decoder. However, this method cannot capture multiple relationships capability and can only mine shallow features with 71.1% F1 score in Aminer dataset and 50.2% F1 score in ECNU dataset. In contrast, our method preserves the heterogeneous information of the graph by fusing the relation features into the feature aggregation process. It can mine deeper associations between nodes and obtain a higher quality node representation.

In the clustering process of the publications, our method combines the topological features of the heterogeneous network and the node embedding representation to jointly assign labels to the publication nodes. At the same time, algorithms such as Zhang et al. [2], Graph Autoencoder, and others do not use topological features in the clustering process and result in ineffective results.

Fig. 7 is a visualization of the node representation of the publication, with the predicted results in the top row and the corresponding Ground truth in the bottom row. It can be found that our model can embed the publication nodes of the same author in a similar space and the publication nodes of different authors in the space separated from each other. From Fig. 7(a) and Fig. 7(d), it is found that even if the nodes of different authors are close in representation space, the method in this publication can divide these nodes. The reason is that this method not only considers the representation of the node itself but also considers the topology of the graph structure in the clustering process.



(a) Predicted result (Xi_liu F1:0.820) (b) Predicted result (Kuo_zhang F1:0.930) (c) Predicted result (Lei_fang F1:0.858)

(d) Ground truth (Xi_liu) (e) Ground truth (Kuo_zhang) (f) Ground truth (Lei_fang)

Fig. 7. Visualization of embedding spaces on publications of three different names

4.4 Component Analysis

To evaluate that our proposed method considering features of relation type can extract better node features than the standard GCN model, we cluster the node features extracted by the standard GCN model and the node features extracted by our proposed method.

From the experimental results in Fig. 8, we know that the method with relation type features considered in both datasets can obtain better clustering partitions than the standard GCN model. This indicates that fusing relation type to the aggregation process can effectively preserve the heterogeneous graph structure information. So our method can capture the potential relationship between nodes and nodes and obtaining a better node representation.

We also did comparative experiments to verify that our proposed improved label propagation can get better results than the standard label propagation algorithm. The experimental results of Fig. 9 show that our proposed method outperforms the standard method on both datasets. This is because we consider the priority of relation types in the label propagation process. In the publications, some information is more beneficial for author name disambiguation, such as coauthor and covenue information. So, the labels should be propagated with these relation types first in the label propagation process. Our proposed method relies on relations in heterogeneous graphs, and it can be better in heterogeneous graphs with more nodes and relations. This may be why our method works well on larger ENCU datasets with more relations and nodes.

Table 2. F1 score values for each algorithm on the Aminer dataset

Names	Our method	HAC	GAE	Metapath2vec	PHNet	Zhang et al [2]
Ajay Gupta	**0.734**	0.362	0.652	0.677	0.687	0.656
Lei Fang	0.858	0.364	0.822	0.667	0.824	**0.880**
Bo Liu	**0.818**	0.329	0.801	0.206	0.800	0.811
Kuo Zhang	**0.930**	0.477	0.799	0.757	**0.930**	0.797
Michael Smith	**0.844**	0.384	0.712	0.431	0.700	0.809
David Cooper	**0.789**	0.392	0.574	0.621	0.679	0.786
Bob Johnson	0.782	0.483	0.722	0.700	**0.863**	0.773
Yu Zhang	0.549	0.462	0.553	0.450	0.594	**0.606**
Avg.	**0.788**	0.407	0.711	0.564	0.760	0.761

Table 3. F1 score values for each algorithm on the ECNU dataset

Names	Our method	HAC	GAE	Metapath2vec	PHNet	Zhang et al [2]
li_ma	**0.510**	0.224	0.462	0.346	0.455	0.457
hui_gao	0.485	0.184	0.441	**0.553**	0.513	0.530
jie_sun	0.528	0.173	0.501	0.405	**0.529**	0.339
jing_yu	**0.708**	0.196	0.529	0.422	0.644	0.442
xi_liu	**0.820**	0.298	0.642	0.613	0.646	0.537
jing_huang	0.544	0.181	0.544	0.184	0.468	**0.622**
yan_gao	**0.541**	0.163	0.467	0.520	0.388	0.342
jie_gong	**0.621**	0.322	0.543	0.571	0.618	0.589
Avg.	**0.590**	0.215	0.502	0.457	0.536	0.457

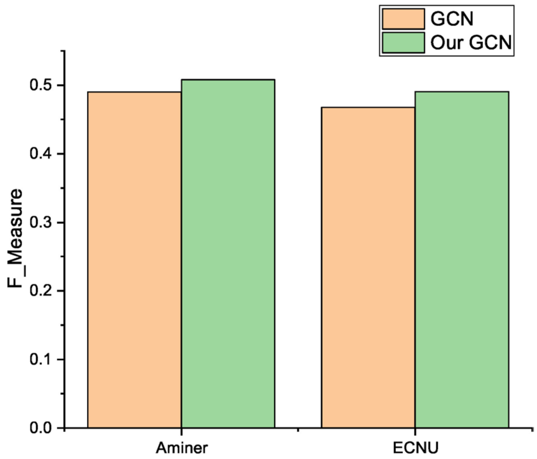


Fig. 8. Node representation test

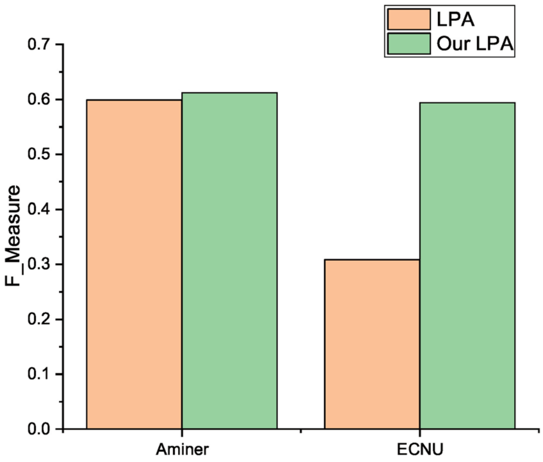


Fig. 9. LPA test

4.5 Clustering Size Estimation

We use commonly used clustering algorithms to compare the predicted number of distinct authors. The experimental results are shown in Table 4, and the results show that the MSLE of our method reaches 0.098, and the error is smaller than that of other clustering methods. For author name David Cooper and Bob Johnson, we were completely accurate in predicting the number of distinct authors. For the author's name Yu Zhang, which has 72 real authors, other methods have too much error in predicting the number of authors, while our method predicts the number closest to 72. This shows that relying on the network structure and node features can better predict the number of real authors. The reason for this lies in the structure of the publication heterogeneous graph. Since many authors of the same name have only a few or even only one publication, there are many isolated nodes and communities consisting of a small number of publication nodes in the graph. Our proposed topology-based label propagation algorithm can handle these isolated publication nodes well, so our method can achieve good results. At the same time, our method does not need to determine any parameters, which can save researchers time and improve efficiency.

Table 4. Clustering number estimation

Name	Actual	HAC	AP	XMeans	Our
Ajay Gupta	9	3	3	5	12
Lei Fang	7	6	4	4	4
Bo Liu	47	2	5	27	27
Kuo Zhang	4	5	3	5	3
Michael Smith	19	23	15	12	15
David Cooper	7	4	2	2	7
Bob Johnson	7	3	3	2	7
Yu Zhang	72	12	9	27	52
MSLE	-	1.536	1.360	0.479	**0.098**

5 Conclusion

In this paper, we proposed a heterogeneous graph-based author name disambiguation method. In obtaining publication representations, we used multidimensional publication information with attribute features and structural features as input to GNN. Furthermore, we fused relation features to the feature aggregation process to preserve the heterogeneous graph structure and simplify the processing of the constructed heterogeneous graph. For the clustering process, we proposed a label propagation algorithm with a selecftable propagation direction for the disambiguation task, avoiding the problem of determining the size of clusters. Experiments on two disambiguation datasets proved that our method performed better than other commonly used methods. We made some improvements, but some related issues still need to be addressed. For example, the literature search system collects lots of new publications every day, and how our approach is compatible with this new publication is an important issue. So, in future work, we will focus on this issue to achieve incremental disambiguation, which can increase the method's practicality. Moreover, we will continue investigating the joint optimization of representation learning and clustering processes, which can achieve better performance.

References

[1] X.-M. Fan, J.-Y. Wang, X. Pu, L.-Z. Zhou, B. Lv, On Graph-Based Name Disambiguation, Journal of Data and Information Quality 2(2)(2011) 10.

[2] B.-C. Zhang, M. Hasan, Name Disambiguation in Anonymized Graphs using Network Embedding, in: Proc. The 26th ACM International Conference on Information and Knowledge Management, 2017.

[3] Z.-Y. Qiao, Y. Du, Y.-J. Fu, P.-F. Wang, Y. Zhou, Unsupervised Author Disambiguation using Heterogeneous Graph Convolutional Network Embedding, in: Proc. 2019 IEEE International Conference on Big Data (Big Data), 2019.

[4] B. Xiong, P. Bao, Y.-L. Wu, Learning semantic and relationship joint embedding for author name disambiguation, Neural Computing and Applications 33(6)(2021) 1987-1998.

[5] J. Tang, A.C.M. Fong, B. Wang, J. Zhang, A Unified Probabilistic Framework for Name Disambiguation in Digital Library, IEEE Transactions on Knowledge and Data Engineering 24(6)(2012) 975-987.

[6] Y. Chen, H.-L. Yuan, T.-T. Liu, N. Ding, Name Disambiguation Based on Graph Convolutional Network, Scientific Programming 2021(2021) 1-11.

[7] Y.-Y. Ma, Y.-L. Wu, C.-Q. Lu, A Graph-Based Author Name Disambiguation Method and Analysis via Information Theory, Entropy 22(4)(2020) 416.

[8] Y.-T. Zhang, F.-J. Zhang, P.-R. Yao, J. Tang, Name Disambiguation in AMiner: Clustering, Maintenance, and Human in the Loop, in: Proc. 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018.

[9] H. Han, L. Giles, H.-Y. Zha, C. Li, K. Tsioutsiouliklis, Two supervised learning approaches for name disambiguation in author citations, Computer, in: Proc. 2004 Joint ACM/IEEE Conference on Digital Libraries, 2004.

[10] L. Zhang, Y. Huang, J.-Q. Yang, W. Lu, Aggregating large-scale databases for PubMed author name disambiguation, Journal of the American Medical Informatics Association 28(9)(2021) 1919-1927.

[11] Y.-B. Chen, Z.-Y. Jiang, J.-L. Gao, H.-L. Du, L. Gao, Z. Li, A supervised and distributed framework for cold-start author disambiguation in large-scale publications, Neural Computing & Applications 35(18)(2023) 13093-13108.

[12] J. Silva, F.M.A. Silva, Feature extraction for the author name disambiguation problem in a bibliographic database, in: Proc. Symposium on Applied Computing, 2017.

[13] L.-W. Peng, S.-Q. Shen, D.-S. Li, J. Xu, Y.-Q. Fu, H.-Y. Su, Author Disambiguation through Adversarial Network Representation Learning, in: Proc. 2019 International Joint Conference on Neural Networks, 2019.

[14] H.-W. Wang, R.-J. Wang, C. Wen, S.-H. Li, Y. Jia, W. Zhang, X. Wang, Author Name Disambiguation on Heterogeneous Information Network with Adversarial Representation Learning, Proceedings of the AAAI Conference on Artificial Intelligence 34(1)(2020) 238-245.

[15] J. Xu, S.-Q. Shen, D.-S. Li, Y.-Q. Fu, A Network-embedding Based Method for Author Disambiguation, in: Proc. The 27th ACM International Conference on Information and Knowledge Management, 2018.

[16] L.-W. Peng, S.-Q. Shen, J. Xu, Y.-Q. Fu, D. Li, A.L. Jia, Diting: An Author Disambiguation Method Based on Network Representation Learning, IEEE Access 7(2019) 135539-135555.

[17] Q. Le, T. Mikolov, Distributed Representations of Sentences and Documents, Proceedings of Machine Learning Research, Proceedings of the 31st International Conference on Machine Learning 32(2)(2014) 1188-1196.

[18] Y.-X. Dong, N.-V. Chawla, A. Swami, metapath2vec: Scalable Representation Learning for Heterogeneous Networks, in: Proc. 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data, 2017.