Zhou Q, Chen W, Zhao PP *et al.* Towards Effective Author Name Disambiguation by Hybrid Attention. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 33(1): 1–23 January 2018. DOI 10.1007/s11390-015-0000-0

Towards Effective Author Name Disambiguation by Hybrid Attention

Qian Zhou(周　乾), Wei Chen*(陈　伟), *Member*, *CCF*, *ACM*, Peng-Peng Zhao(赵朋朋), *Member*, *CCF*, *ACM*, An Liu(刘　安), *Member*, *CCF*, *ACM*, Jia-Jie Xu(许佳捷), *Member*, *CCF*, *ACM*, Jian-Feng Qu(瞿剑峰),*Member*, *CCF*, *ACM*, and Lei Zhao*(赵　雷), *Member*, *CCF*, *ACM*

School of Computer Science and Technology, Soochow University, Suzhou 215006, China

E-mail: {qzhou0, robertchen, ppzhao, anliu, xujj, jfqu, zhaol}@suda.edu.cn

Received July 15, 2022; accepted October 14, 2022.

Abstract Author name disambiguation (AND) is a central task in academic search, which has received more attention recently accompanied by the increase of authors and academic publications. To tackle the AND problem, existing studies have proposed various approaches based on different types of information, such as raw document features (e.g., co-authors, titles, and keywords), the fusion feature (e.g., a hybrid publication embedding based on multiple raw document features), the local structural information (e.g., a publication's neighborhood information on a graph), and the global structural information (e.g., interactive information between a node and others on a graph). However, there has been no work taking all the above-mentioned information into account and taking full advantage of the contributions of each raw document feature for the AND problem so far. To fill the gap, we propose a novel framework named EAND (Towards Effective Author Name Disambiguation by Hybrid Attention). Specifically, we design a novel feature extraction model, which consists of three hybrid attention mechanism layers, to extract key information from the global structural information and the local structural information that are generated from six similarity graphs constructed based on different similarity coefficients, raw document features, and the fusion feature. Each hybrid attention mechanism contains three key modules: a local structural perception, a global structural perception, and a feature extractor. Additionally, the mean absolute error function in the joint loss function is used to introduce the structural information loss of vector space. Experimental results on the two real-world datasets demonstrate that EAND achieves superior performance, outperforming all state-of-the-art methods by at least +2.74% in terms of the micro-F1 score and +3.31% in terms of the macro-F1 score.

Keywords author name disambiguation, multiple feature information, hybrid attention mechanism, pruning strategy, structural information loss of vector space

1 Introduction

Over the past decade, we have witnessed the unprecedented growth of academic digital records[1, 2]. For instance, the latest estimation presents that there are more than 271 million publications, 133 million scholars, and 754 million citations on AMiner[3]. These numbers are significantly surpassed by those on Google Scholar, Digital Bibliography & Library Project (DBLP), and Microsoft Academic[4, 5]. While the rapid development of academic digital databases has indeed brought great convenience to researchers, it has also introduced novel problems. Specifically, when searching for publications in these databases, an author's personal name, intended to identify a certain individual, is often used as a keyword. However, different authors may have the same or similar names, and one author may use different spellings or name variants in the real world. Such phenomena lead to the author name disambiguation (AND) problem, also referred to as object distinction[6] and name

Regular Paper

A preliminary version of the paper was published in the Proceedings of ICWS 2021.

This work was supported by the Major Program of the Natural Science Foundation of Jiangsu Higher Education Institutions of China under Grant Nos. 19KJA610002 and 19KJB520050, and the National Natural Science Foundation of China under Grant No. 61902270.

*Corresponding Author

identification[7]. This problem can cause inconvenience in data mining communities and academic information retrieval[8, 9]. For instance, an online search in DBLP for "Michael Jordan" may yield profiles of professors from UC Berkeley, Germany Helmut Schmidt University, Glasgow Caledonian University, and other institutions. This has become a common problem across various online digital databases. Making the matter worse, the growth of publications and researchers shows an unprecedented increase in recent years. As a result of the aforementioned cases, the AND problem has become an urgent and pressing task[10].

The goal of AND is to split a set of publications under the same author name into several disjoint groups, where each group represents publications published by a unique person[11–13]. In this domain, dedicated researchers have made significant efforts, and various communities have proposed numerous methods[11]. The existing methods roughly utilize the following four types of information, such as raw document features (e.g., co-authors, titles, and keywords), the fusion feature (e.g., a hybrid publication embedding based on multiple raw document features), the local structural information (e.g., a publication's neighborhood information on a graph), and the global structural information (e.g., the interactive information between a node and others on a graph). For instance, some methods employ the fusion feature to generate the context embedding and extract the global structural information[11, 12], or directly disambiguate author names based on this feature[13–15]. Other studies[16–18] utilize raw document features to construct similarity graphs and extract the local structural information based on these graphs to model high-order connections capturing publications' neighborhood information. In general, existing studies have made significant contributions to author name disambiguation, but they still suffer from the following challenges.

• On the one hand, they merely utilize a part of features and have not effectively addressed the issue: missing of raw document features (MRDF)[19], which refers to the phenomenon that some publications contain only certain raw document features, such as titles and co-authors, while other features (e.g., venues and keywords) are missing. By way of illustration, the methods[11, 12, 14, 15], which are developed based on the fusion feature and the global structural information, cannot capture a publication's neighborhood information, due to the absence of local structure information. The studies[16–18] relying on raw document features and the local structural information have encountered the MRDF issue. The MRDF issue poses a significant challenge in measuring pairwise publication similarity using raw document features and extracting the local structural information from them[19]. Particularly, this issue significantly impacts [16] and [18], which exclusively depend on the local structural information extracted from graphs built upon raw document features. While [17] considers the fusion feature, the local structural information, and raw document features, it overlooks the global structural information used to further precisely measure the similarity between two publications from a global perspective.

• On the other hand, none of the existing methods takes full advantage of the contributions of each raw document feature. We illustrate this in detail by dividing the existing studies into three categories. In the first category of studies, researchers propose a set of carefully designed heuristics and similarity functions, which utilize each raw document feature to generate a weighted sum of similarities. By doing so, these methods can effectively exploit such differences in the contributions of each raw document feature[13, 15, 20]. However, these studies do not consider the issue: scale-difference of raw document features (SRDF). SRDF refers to the phenomenon that some publications have many words in one raw document feature, while others have only a few words. For example, given the raw document feature "Co-author", [21] has two co-authors, while [22] has seven co-authors. This phenomenon brings a great challenge for quantifying the similarity between publications with raw document features. The studies in the second category only utilize strong discriminative features selected by humans (e.g., co-authors) for disambiguation[16, 18]. Notably, to address the AND

problem, the work[18] divides publications into small blocks based on the discriminative author attribute (i.e., affiliations). It further exploits semantic information by using attention mechanisms and meta-paths on a heterogeneous graph constructed by publications, authors, topics, and venues. However, the work[18] only focuses on certain raw document features, overlooking the MRDF and SRDF issues. Consequently, the utilization of the multi-view attention mechanism designed by [18] to capture meta-path significance does not effectively leverage the contributions of each raw document feature to address the AND problem. In addition, studies in the third category[11, 17, 19] adopt the most raw document features, but similar to the previously mentioned methods, they also ignore the contributions of each raw document feature to the AND problem.

To overcome the challenges of the existing studies, we propose a unified framework named EAND (Towards Effective Author Name Disambiguation by Hybrid Attention), where all types of feature information are taken into account and a hybrid attention mechanism is developed. Specifically, the proposed framework EAND consists of the following four components. 1) We construct five similarity graphs based on five raw document features (i.e., co-authors, affiliations, venues, titles, and keywords) to generate the global structure information. 2) We extract the local structure information from the fusion feature, enabling a more comprehensive representation of a publication. This is because the fusion feature contains the interaction information between raw document features in one publication, which can robustly quantify the similarity of pairwise publications[12]. 3) We design a novel feature extraction model (EX) that captures the influence between multiple types of feature information, aggregates these features, takes full advantage of the contributions of each raw document feature, and extracts the key information required for addressing the AND problem. Note that EX is composed of three hybrid attention mechanism layers with a residual block. Each hybrid attention mechanism comprises three key modules, i.e., a local structural perception, a global structural perception and a feature extractor. 4) A decision model (DI) is developed to determine whether two publications belong to a unique author. In this model, the triplets composed of pairwise nodes' embeddings and the corresponding edges' embeddings are fed into a multilayer perceptron (MLP). Furthermore, our framework employs a joint loss function including the log-likelihood loss function (Cross-Entropy-Loss) and the mean absolute error function (L1-Loss) for training. Note that here L1-Loss serves to introduce the structural information loss of the vector space, which encourages the vectors of positive pairs to be closer than negative pairs in the vector space.

Our main contributions are outlined as follows.

- We propose a unified framework EAND, where multiple feature information is taken into account and the contributions of each raw document feature are used sufficiently, with the goal of addressing the AND problem more effectively.
- We design a novel generating strategy, which extracts the local structure information from the fusion feature and extracts the global structural information from raw document features, to solve the MRDF issue to a certain extent. To tackle the SRDF issue, we improve the traditional similarity coefficients. Our framework incorporates a novel feature extraction model comprising three hybrid attention mechanism layers that capture the influence between multiple types of feature information, aggregate these features, and fully utilize the contributions of each raw document feature. Moreover, L1-Loss in the joint loss function is applied to introduce the structural information loss of the vector space.
- The experimental results, based on two real world datasets, demonstrate the superiority of the proposed framework EAND over state-of-the-art methods.

In this paper, we implement more delicate processing. Compared with the conference version [19] of this work, we make the following improvements.

- To tackle the AND problem more effectively, we develop a novel framework EAND to learn and take full advantage of the contributions of each raw document feature. Specifically, a hybrid attention mechanism in EAND is designed to utilize the contributions of each

raw document feature and extract key information essential for resolving the AND problem.

• In addition to considering the MRDF phenomenon, we also consider the SRDF phenomenon. To effectively quantify the similarity of pairwise publications, we improve the traditional similarity coefficients.

• In the loss function, we introduce the structural information loss of the vector space using L1-Loss, which encourages the vectors of positive pairs to be closer than negative pairs in the vector space. Our codes are publicly available on github[1].

This paper is organized as follows. Section 2 presents the formulation of the AND problem and the definition of the similarity graph. Section 3 discusses the solution for addressing the AND problem. Performance evaluation results and comparative analysis are given in Section 4. Section 5 describes the related work. In Section 6, conclusions and future work are discussed.

2 Preliminaries

2.1 Problem Definition

For a given author name α, let $P^\alpha = \{p_1^\alpha, \ldots, p_N^\alpha\}$ represent a set of N publications associated with this author name α, where $p_i^\alpha \in P^\alpha$ is a publication that contains a series of raw document features, denoted as $p_i^\alpha = \{x_1, x_2, \ldots, x_K\}$. Here, K is the number of raw document features and x_i is the information of the raw document feature, such as co-authors, affiliations, titles, keywords, or venues. Moreover, we use $\Psi(p_i^\alpha, p_j^\alpha)$ to describe whether p_i^α and p_j^α have the same identity or not[11]. More precisely, if p_i^α and p_j^α have the same identity, we have $\Psi(p_i^\alpha, p_j^\alpha) = 1$; otherwise, we have $\Psi(p_i^\alpha, p_j^\alpha) = 0$. Notably, we omit the superscript α in the following description if there is no ambiguity (e.g., $p_i^\alpha \to p_i$). Prior to formulating the problem of author name disambiguation (AND), we introduce some relevant concepts as follows.

Missing of Raw Document Features (*MRDF*). This phenomenon is that some publications only have a part of raw document features such as titles and co-authors, while other features (e.g., affiliations, venues and keywords) are missing[19].

Scale-Difference of Raw Document Features (*SRDF*). This is the phenomenon that some publications have many words in one raw document feature, while others have only a few words. For example, given the raw document feature: "Co-author", [21] has two co-authors, while [22] has seven co-authors.

Contributions of Each Raw Document Feature. Each raw document feature exhibits different discriminative capabilities[19, 20] and contributes uniquely to the AND problem. We formulate it as follows:

$$KI = \sum_{m=1}^{K} a_m \times KI_{x_m},$$

where KI_{x_m} is the information of m-th raw document feature such as co-authors, affiliations, titles, keywords, or venues, a_m is the weight which represents contributions of the m-th raw document feature, and KI is a weighted sum of contributions of each raw document feature and represents the key information for dealing with the AND problem.

Disambiguating Function. The disambiguating function Θ is shown as follows:

$$\Theta = \{\mathrm{EX}, \mathrm{DI}\},$$

where EX is a feature extraction model including a hybrid attention mechanism, which considers the phenomenon: MRDF and SRDF, and the contributions of each raw document feature, to extract KI for the AND problem; and DI is a decision model that is proposed based on the extracted KI for disambiguation.

Problem Formulation. Given a publication set $P = \{p_1, \ldots, p_N\}$ associated with the author name α, author name disambiguation aims at finding the function Θ to partition P into a set of disjoint clusters based on multiple feature information (i.e., raw document features, the fusion feature, the local structural information, and the global structural information) and the contributions of each raw document feature, i.e.,

$$\Theta(P) \to C = \{c_1, c_2, \ldots, c_k\},$$

[1] https://github.com/wx-qzhou/EAND, June 2022.

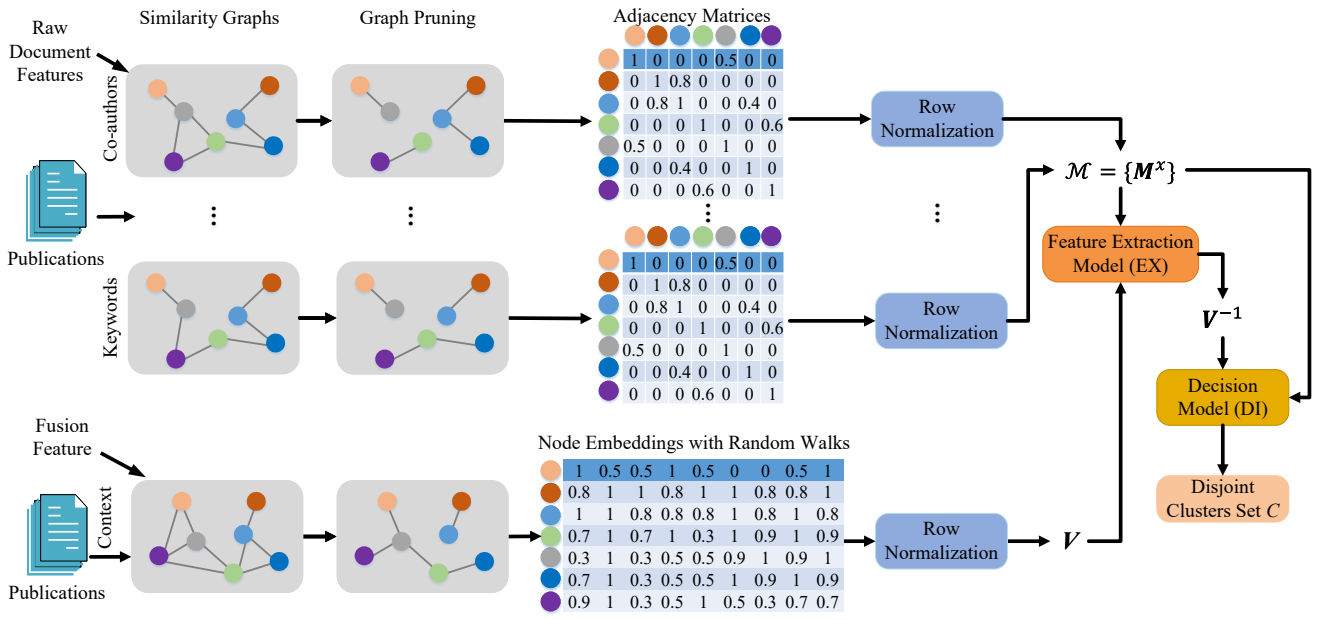


Fig. 1. Overview of the proposed framework EAND.

where C is the set of disjoint clusters; c_k is the k-th cluster that only contains publications of the same identity, i.e., $\forall(p_i, p_j) \in c_k \times c_k, \Psi(p_i, p_j) = 1$, and different clusters contain publications of different identities, i.e., $\forall(p_i, p_j) \in c_k \times c_{k'}, k \neq k', \Psi(p_i, p_j) = 0$.

2.2 Similarity Graph

Given a publication set $P = \{p_1, p_2, \ldots, p_N\}$ associated with the author name α, we construct a similarity graph $G_x = (D, E, x, S_x, w)$, where D and E are sets of nodes and edges in the graph, respectively. In detail, $D_i \in D$ is a node and represents the publication p_i. $E_{ij} \in E$ is an edge and represents that two publications p_i and p_j have a certain degree of similarity. The feature x is the information of one raw document feature (e.g., co-authors or affiliations) or the fusion feature. S_x is the similarity function to quantify the similarity of pairwise publications based on the feature x. The weight w_{ij} is the similarity of pairwise publications p_i and p_j calculated by S_x, and if pairwise publications p_i and p_j do not have the feature x or the intersecting set based on the feature x, the weight w_{ij} is zero. The similarity graph is an undirected weighted graph, i.e., $E_{ij} = w_{ij} = E_{ji} = w_{ji}$. Each node has a self-loop, which is an edge that connects a vertex to itself, i.e., $E_{ii} = w_{ii}$. We present an example of a similarity graph in Fig.2. Note that, we construct six similarity graphs, i.e., $G_{\text{co-author}}$, $G_{\text{affiliation}}$, G_{title}, G_{keyword}, G_{venue}, and G_{fusion} based on five common raw document features and the fusion feature.

3 Proposed Framework

An overview of our proposed framework, EAND, is illustrated in Fig.1. To be specific, we first describe how the similarity of pairwise publications is estimated using multiple types of similarity coefficients based on raw document features and the fusion feature. Then, five raw document feature graphs, which are used to extract the global structural information, are constructed based on co-authors, affiliations, venues, titles, and keywords. Next, we construct the fusion feature graph, which is utilized to extract the local structural information. After that, we introduce how the feature extraction model (EX) aggregates multiple feature information and extracts the key information for disambiguation. Finally, the decision model (DI) is employed to convert the AND problem into a binary classification between each publication pair.

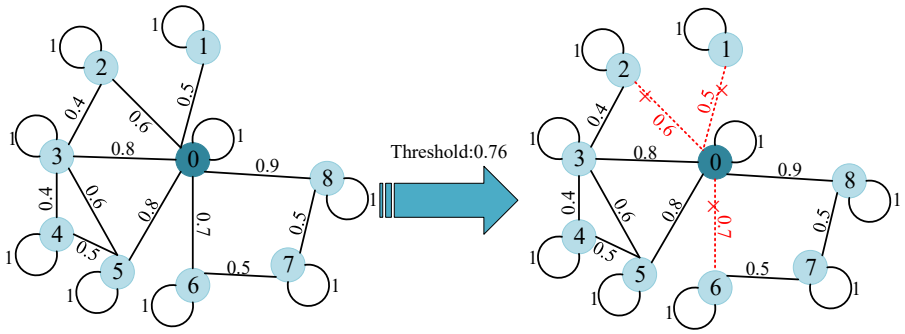


Fig. 2. An illustrative example of the similarity graph and the graph pruning.

3.1 Similarity Coefficients and Similarity of Pairwise Publications

Following previous work[13, 15, 20], we use multiple types of similarity coefficients to quantify the similarity between pairwise publications, with improvements made specifically to address the SRDF issue. To provide a comprehensive explanation, we commence by introducing relevant notations. Following that, we present the definitions of similarity coefficients and subsequently formulate the concept of pairwise publication similarity. Formally, given a publication set $P = \{p_1, \ldots, p_N\}$ associated with the author name α, $Z^x = \{z_1^x, \ldots, z_{\boldsymbol{\kappa}}^x\}$ is the set of words based on the feature x in these publications, where each word z_i^x in Z^x is not repetitive and x is one of the five raw document features (i.e., co-authors, affiliations, venues, titles, or keywords) or the fusion feature. Given two publications p_i and p_j, $Z_i^x \subset Z^x$ and $Z_j^x \subset Z^x$ are the word sets of p_i and p_j based on x, respectively. In addition, $\mathrm{Sim}(p_i, p_j)^x$ is used to denote the similarity of pairwise publications p_i and p_j based on Z_i^x and Z_j^x, and it is set to zero on condition that $Z_i^x \cap Z_j^x = \emptyset$.

3.1.1 *IDF or TF-IDF Based Similarity Coefficient*

Compared with MFAND[19], we introduce a penalty factor into the similarity coefficient based on IDF or TF-IDF to mitigate the influence of SRDF. Formally, let $Y^x = \{y_1^x, y_2^x, \ldots, y_{\boldsymbol{\kappa}}^x\}$ be the set of the weight of the words in Z^x, where y_i^x calculated based on IDF or TF-IDF is the weight of the word z_i^x. $\tilde{Y}^x = \{\tilde{y}_1^x, \tilde{y}_2^x, \ldots, \tilde{y}_{\boldsymbol{\kappa}}^x\}$ is the normalized representation of Y^x. The sum of the normalized weight of the words in the set $Z_i^x \cap Z_j^x$ is utilized to denote the similarity between two publications p_i and p_j, and we formalize it as follows:

$$\mathrm{Sim}(p_i, p_j)_1^x = \sum_{z_m^x \in (Z_i^x \cap Z_j^x)} \tilde{y}_m^x \times \epsilon, \epsilon = \frac{\sqrt{|Z_i^x \cap Z_j^x|}}{|Z_i^x \cap Z_j^x|}, \tag{1}$$

where $\tilde{y}_m^x$ is the normalized weight of the word z_m^x, $|\cdot|$ is the length of a set, and ϵ is a penalty factor.

3.1.2 *Jaccard-Based Similarity Coefficient*

The similarity coefficient of pairwise publications quantified based on the improved Jaccard is given as:

$$\mathrm{Sim}(p_i, p_j)_2^x = \frac{|Z_i^x \cap Z_j^x|}{2 \times \min(|Z_i^x|, |Z_j^x|) - |Z_i^x \cap Z_j^x|}, \tag{2}$$

where $\min(\cdot)$ is the function to get the smallest value.

3.1.3 *Dice-Based Similarity Coefficient*

The improved Dice-based similarity coefficient of pairwise publications is defined as:

$$\mathrm{Sim}(p_i, p_j)_3^x = \frac{|Z_i^x \cap Z_j^x|}{\min(|Z_i^x|, |Z_j^x|)}. \tag{3}$$

3.1.4 *Word2vec-Based Similarity Coefficient*

For some raw document features (e.g., titles and abstracts) that contain semantic information, we first utilize word2vec to embed them. Specifically, let $Y^x = \{\boldsymbol{y}_1^x, \boldsymbol{y}_2^x, \ldots, \boldsymbol{y}_N^x\}$ be the embedding set of $P = \{p_1, p_2, \ldots, p_N\}$ based on the feature x, where $\boldsymbol{y}_i^x$ is the embedding of i-th publication p_i. We estimate the

similarity of pairwise publications using both cosine similarity and Pearson correlation coefficient, i.e.,

$$\mathrm{Sim}(p_i, p_j)_4^x = s(\boldsymbol{y}_i^x, \boldsymbol{y}_j^x), \tag{4}$$

where $s(\cdot)$ is the operation of cosine similarity or Pearson correlation coefficient.

3.1.5 *Similarity of pairwise publications*

To obtain the final similarity $\mathrm{Sim}(p_i, p_j)^x$ of pairwise publications p_i and p_j, we aggregate all similarity values corresponding to the feature x, i.e.,

$$\mathrm{Sim}(p_i, p_j)^x = \sum \mathrm{Sim}(p_i, p_j)_m^x, \tag{5}$$

where $m \in \{1, 2, 3, 4\}$ and the selection of it is determined by which feature is to be explored. For example, m is set to 1, 2, and 3 if we quantify the similarity of a publication pair based on the feature "Co-author". The correspondence between features and similarity coefficients is presented in Table 1.

Table 1. Features and Corresponding Similarity Coefficients

Feature	Similarity Coefficient
Co-author	Based on IDF, TF-IDF, Jaccard and Dice
Affiliation	Based on IDF, TF-IDF, Jaccard and Dice
Venue	Based on IDF and Jaccard
Keyword	Based on IDF and Jaccard
Title	Based on IDF and Word2vec
Fusion feature	Based on IDF

3.2 Construction of Raw Document Feature Graph

Following the calculation of the similarity of pairwise publications, we construct five raw document feature graphs corresponding to co-authors, affiliations, venues, titles, and keywords, respectively. The construction process details are presented below.

3.2.1 *Raw Document Feature Graph*

Let $P = \{p_1, p_2, \ldots, p_N\}$ be a set of publications written by authors with name α, we employ an $N \times N$ adjacent matrix $\hat{\boldsymbol{M}}^x$ to denote the raw document feature graph $G_x = (D, E)$, where x is one of the five raw document features. $\hat{M}_{ij}^x$ is the similarity of pairwise publications p_i and p_j, which is calculated based on (1)–(5), and denotes the weight of edge E_{ij} between D_i and D_j. Note that the low similarity of pairwise publications indicates weak edges between them. All weak edges are considered to be noises. Hence, we adopt a pruning strategy as outlined in [23] to prune them within each raw document feature graph. The specifics of this strategy are detailed below.

3.2.2 *Raw Document Feature Graph Pruning*

Given the $N \times N$ adjacent matrix $\hat{\boldsymbol{M}}^x$ of the raw document feature graph G_x, the pruned adjacent matrix $\check{\boldsymbol{M}}^x$ is defined as follows:

$$\tilde{M}_{ij}^x = \begin{cases} 0, & \text{if } \hat{M}_{ij}^x < \frac{\sum_{m=0}^{N-1} \hat{M}_{im}^x}{N}, \\ \hat{M}_{ij}^x, & \text{else}, \end{cases}$$

$$\check{M}_{ij}^x = \frac{\tilde{M}_{ij}^x + \tilde{M}_{ji}^x}{2}.$$

This pruning strategy filters weak edges by setting a threshold, which is calculated as the mean value of each row and column in the matrix $\hat{\boldsymbol{M}}^x$[23]. Specifically, any element in $\hat{\boldsymbol{M}}^x$ that is below the threshold is set to zero. Fig.2 illustrates an example of the pruned process. The pruned adjacent matrix $\check{\boldsymbol{M}}^x$ is a comprehensive embedding for all edges and denotes the global structural information. Additionally, row normalization is applied to each $N \times N$ pruned raw document feature adjacency matrix to obtain the final adjacency matrix $\boldsymbol{M}^x$, i.e.,

$$M_{ij}^x = \frac{\check{M}_{ij}^x}{\sum_{m=0}^{N-1} \check{M}_{im}^x}.$$

3.3 Construction of Fusion Feature Graph

Apart from constructing the raw document feature graphs, we also build a fusion feature graph based on the fusion feature and the similarity of pairwise publications, as introduced in Subsection 3.1. The fusion feature, which is utilized to extract the local structural information, is generated by concatenating raw document features, including co-authors, affiliations, venues, titles, and keywords. Additionally, to ensure more precise local structural information, we utilize a novel pruning strategy to retain only the information that is highly

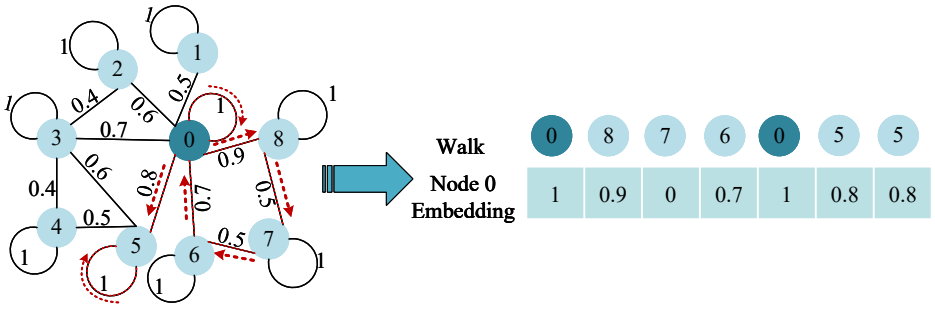


Fig. 3. Illustration of node embeddings.

relevant to the current node, filtering out weak edges. The details of them are presented as follows.

3.3.1 *Fusion Feature Graph Pruning*

Formally, we first construct an $N \times N$ adjacent matrix $\hat{\boldsymbol{M}}^f$, where each element in $\hat{\boldsymbol{M}}^f$ is the similarity of pairwise publications based on the fusion feature, which is calculated based on (1) and (5). Then, the pruned adjacent matrix $\boldsymbol{M}^f$ is defined as:

$$\theta = \frac{\beta \times \frac{\sum_{m=0}^{N-1} \hat{M}_{im}^f}{N} + \max(\hat{\boldsymbol{M}}_i^f)}{\beta + 1},$$

$$\tilde{M}_{ij}^f = \begin{cases} 0, & \text{if } \hat{M}_{ij}^f < \theta, \\ \hat{M}_{ij}^f, & \text{else}, \end{cases}$$

$$M_{ij}^f = \frac{\tilde{M}_{ij}^f + \tilde{M}_{ji}^f}{2},$$

where θ represents the pruning threshold for each row, and β is a parameter to balance the mean and maximum values, leading to improved pruning results.

3.3.2 *Fusion Feature Graph*

Based on the pruned $N \times N$ adjacent matrix $\boldsymbol{M}^f$, a fusion feature graph $G_f = (D, E)$ is built based on the following triples. If M_{ij}^f is nonzero, a triple (D_i, D_j, E_{ij}) is built, where D_i and D_j are the nodes of this graph and represent the publications p_i and p_j, respectively. E_{ij} denotes the weight of the edge between D_i and D_j, and the value of it is equal to M_{ij}^f.

3.3.3 *Embeddings of Nodes*

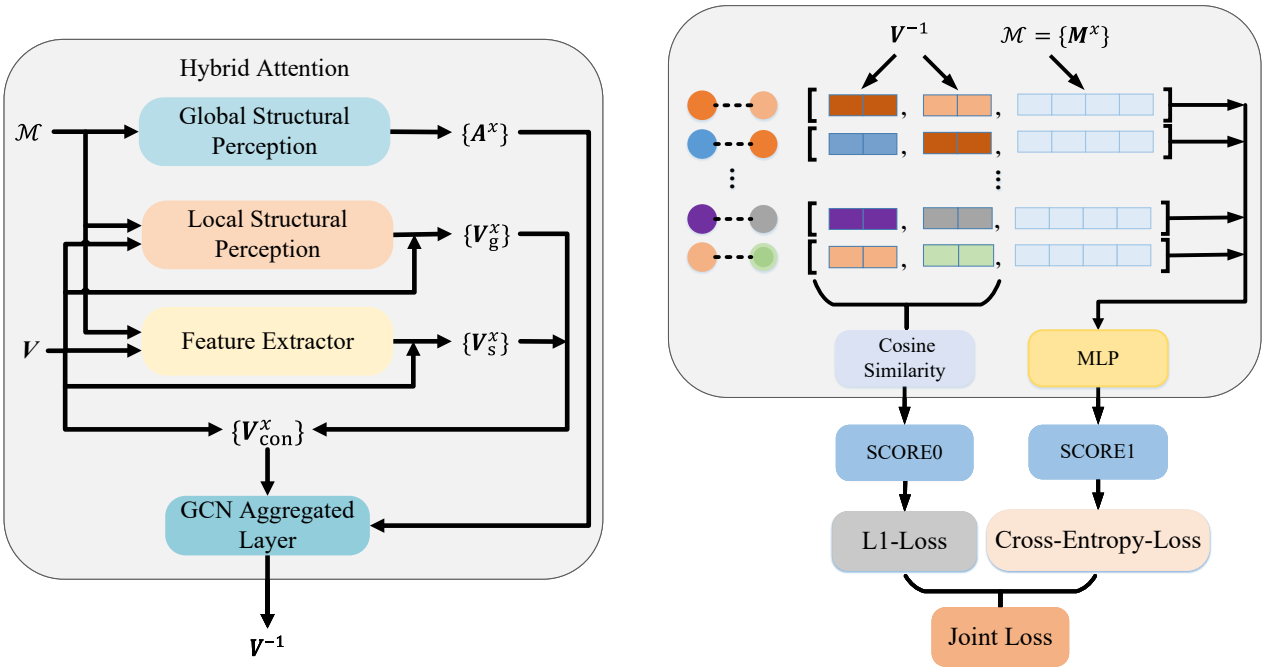
Following the construction of the fusion feature graph, the random walk[24], which can capture the neighborhood information of nodes, is utilized to construct the walks on the fusion feature graph. By refining these walks, we obtain the nodes' embeddings that represent the local structural information.

Specifically, we denote the embeddings of nodes $D = \{D_i\}$ as $\boldsymbol{V} = (\boldsymbol{V}_i)$ and the k-th neighborhood of node D_i as D_{ik}, where $\boldsymbol{V}_i$ is the embedding of i-th node D_i. In detail, for a given fusion feature graph $G_f = (D, E)$, we first generate a set of random walks $\Omega_i = \{\omega_j^i\}$ for each node D_i, where the j-th walk ω_j^i consists of some neighborhood nodes around the node D_i, i.e., $\omega_j^i = \{D_{ik}\}$. Then, for the node D_i and its j-th random walk $\omega_j^i = \{D_{ik}\}$, this walk is redefined by using the weight of the edges between D_i and the nodes in this walk. Finally, this redefined walk is employed to embed the node D_i, i.e., $\hat{\boldsymbol{V}}_i = (\hat{V}_{ik})$, $\hat{V}_{ik} = E_{ik}$. An example of this process is shown in Fig.3. Moreover, we normalize the embeddings of nodes to obtain $\boldsymbol{V}$, and the normalization process is as follows:

$$V_{ik} = \frac{\hat{V}_{ik}}{\sum_{m=0}^{|\hat{\boldsymbol{V}}_i|-1} \hat{V}_{im}}.$$

3.4 Feature Extraction Model

We introduce the feature extraction model (EX), which aggregates multiple feature information, captures the influence between multiple types of feature information, takes full advantage of the contributions of each raw document feature, and extracts the key information needed for solving the AND problem. EX is composed of three hybrid attention mechanism layers, and each layer contains three key modules, i.e., a local structural perception, a global structural perception, and a feature



(a) (b)

Fig. 4. Details of (a) feature extraction model (EX) and (b) decision model (DI).

extractor. Fig.4(a) illustrates the detailed architecture of EX, and Fig.5 showcases each module of the hybrid attention mechanism.

3.4.1 *Local Structural Perception*

Local structural perception consists of five GAT (graph attention network) layers, which are designed to assign varying weights to different neighboring nodes in a node's neighborhood and explore the contributions of each raw document feature. Formally, it can be described as follows:

$$\boldsymbol{H} = \boldsymbol{V}\boldsymbol{W},$$

$$\boldsymbol{a} = g(Con_{\mathrm{c}}(Con_{\mathrm{r}}(\boldsymbol{H}_i, \boldsymbol{H}_j))\boldsymbol{W}_{\mathrm{a}}), i, j \in \{0, \ldots, N\},$$

$$\boldsymbol{a}^x_{\mathrm{mask}} = mask_{\mathrm{r}}(\boldsymbol{a}, \boldsymbol{M}^x),$$

$$\boldsymbol{a}^x_{\mathrm{norm}} = \mathrm{softmax}(\boldsymbol{a}^x_{\mathrm{mask}}),$$

$$\boldsymbol{V}^x_{\mathrm{g}} = \boldsymbol{a}^x_{\mathrm{norm}}\boldsymbol{V} + \boldsymbol{H},$$

where $\boldsymbol{W}$ is a shared linear transformation weight to obtain sufficient expressive power; $Con_{\mathrm{c}}(\cdot)$ and $Con_{\mathrm{r}}(\cdot)$ are concatenation operations along each column and row, respectively; $\boldsymbol{W}_{\mathrm{a}}$ is the weight vector of attention coefficient $\boldsymbol{a}$; $g(\cdot)$ is an activation function, i.e., LeakyReLU; $mask_{\mathrm{r}}(\cdot)$ is the operation injecting each raw document feature graph structure into the mechanism to achieve $\boldsymbol{a}^x_{\mathrm{mask}}$; softmax$(\cdot)$ is the softmax function; $\boldsymbol{a}^x_{\mathrm{norm}}$ is the final attention coefficient; and $\boldsymbol{V}^x_{\mathrm{g}}$ is the final output feature embeddings for all nodes in one raw document feature (i.e., x) graph.

3.4.2 *Global Structural Perception*

Global structural perception is designed to investigate which nodes in the graph are important to the current node, leveraging the information from edges that correspond to a specific raw document feature x. We formalize it as follows:

$$\boldsymbol{a}^x = \mathrm{softmax}(\boldsymbol{M}^x),$$

$$\boldsymbol{A}^x = \boldsymbol{a}^x \odot \boldsymbol{M}^x + \boldsymbol{M}^x,$$

where $\boldsymbol{a}^x$ is the probability coefficient of $\boldsymbol{M}^x$, $\odot$ is the element-wise (Hadamard) product, and $\boldsymbol{A}^x$ is a comprehensive embedding for all edges with attention.

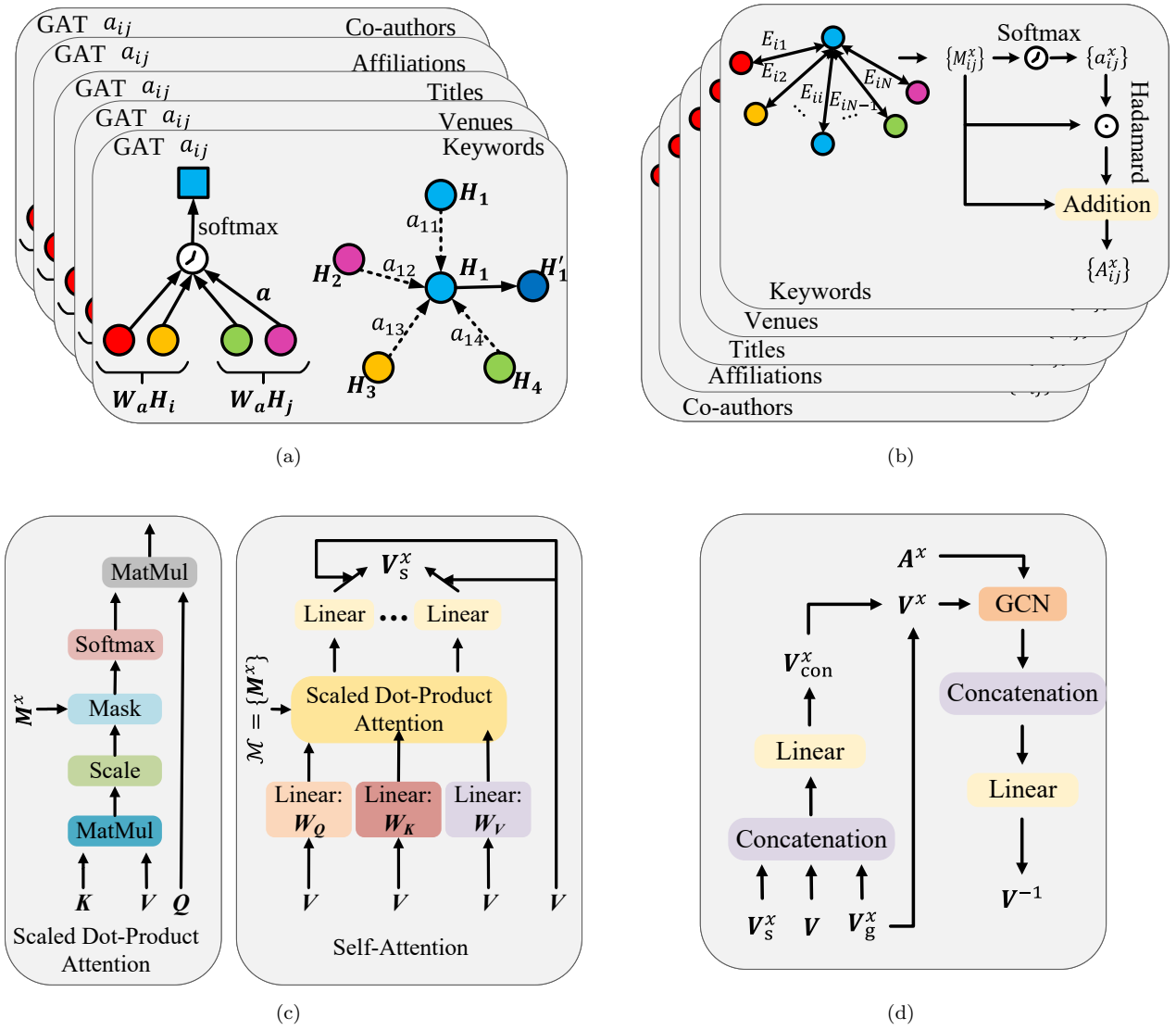


Fig. 5. Each module of the hybrid attention mechanism. (a) Local structural perception. (b) Global structural perception. (c) Feature extractor. (d) GCN aggregated layer.

3.4.3 *Feature Extractor*

In this module, we utilize a self-attention mechanism to extract the information of each raw document feature from a hybrid embedding, thereby generating the key information corresponding to each raw document feature. Through this process, our framework can explore which raw document features are important for the AND problem. We formalize it as follows:

$$\boldsymbol{Q} = \boldsymbol{V}\boldsymbol{W}_{\boldsymbol{Q}}, \boldsymbol{K} = \boldsymbol{V}\boldsymbol{W}_{\boldsymbol{K}}, \boldsymbol{V} = \boldsymbol{V}\boldsymbol{W}_{\boldsymbol{V}},$$

$$\boldsymbol{S} = \text{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d}}),$$

$$\boldsymbol{S}^x_{\text{mask}} = mask_{\text{f}}(\boldsymbol{S}, \sigma(\boldsymbol{M}^x)),$$

$$\boldsymbol{S}^x_{\text{att}} = \boldsymbol{S}^x_{\text{mask}}\boldsymbol{V},$$

$$\boldsymbol{V}^x_{\text{s}} = \rho(\boldsymbol{S}^x_{\text{att}}\boldsymbol{W}_{\text{s}}) + \boldsymbol{V},$$

where $\boldsymbol{W}_{\boldsymbol{Q}}$, $\boldsymbol{W}_{\boldsymbol{K}}$, $\boldsymbol{W}_{\boldsymbol{V}}$, and $\boldsymbol{W}_{\text{s}}$ are projection matrices; d is the latent dimensionality, and $\sqrt{d}$ is the scale factor used to avoid overly large inner product values, especially when the d is high; $mask_{\text{f}}(\cdot)$ is an operation that extracts the information of each raw document feature from the hybrid embedding $\boldsymbol{S}$ based on $\boldsymbol{M}^x$; $\boldsymbol{S}^x_{\text{att}}$ represents the output embeddings of the self-attention

mechanism; $\sigma(\cdot)$ is a binary function that converts values to either 0 or 1; $\rho(\cdot)$ is an activation function, and RELU is used; and $\boldsymbol{V}_{\mathrm{s}}^{x}$ is an embedding matrix that contains all publications with the author name α and the raw document feature x.

3.4.4 GCN Aggregated Layer

After acquiring all the above information, we aggregate them using graph convolutional networks (GCNs) and a single-layer feedforward neural network.

$$\boldsymbol{V}_{\mathrm{Con}}^{x} = g(Con(\boldsymbol{V}_{\mathrm{g}}^{x}, \boldsymbol{V}_{\mathrm{s}}^{x}, \boldsymbol{V})\boldsymbol{W}_{\mathrm{Con}}),$$

$$\boldsymbol{V}^{x} = \boldsymbol{V}_{\mathrm{Con}}^{x} + \boldsymbol{V}_{\mathrm{g}}^{x},$$

$$V_{\mathrm{Conx}}(\boldsymbol{V}) = \rho(Con_{x}(GCN(\boldsymbol{A}^{x}, \boldsymbol{V}^{x}))\boldsymbol{W}_{\mathrm{Conx}}),$$

where $Con(\cdot)$ is a concatenation operation aggregating the information about the three key modules of the hybrid attention mechanism; $Con_{x}(\cdot)$ is also a concatenation operation that can aggregate the information about five raw document features; $GCN(\cdot)$ is a GCN layer; $\boldsymbol{W}_{\mathrm{Con}}$ and $\boldsymbol{W}_{\mathrm{Conx}}$ are the weight of concatenation operations $Con(\cdot)$ and $Con_{x}(\cdot)$, respectively; and $V_{\mathrm{Conx}}(\cdot)$ represents a hybrid attention mechanism layer.

Additionally, to prevent information loss during the training process, we introduce a residual block in EX based on the idea from [25]. Specifically, EX consists of three hybrid attention mechanism layers with a residual block, and it takes the nodes' embeddings $\boldsymbol{V}$ and the set of feature adjacency matrices $\mathcal{M} = \{\boldsymbol{M}^{x}\}$ as input. The residual formulation is as follows:

$$\boldsymbol{V}^{k+1} = V_{\mathrm{Conx}}(Norm(\boldsymbol{V}^{k} + \sigma(\boldsymbol{V}^{k-1}))),$$

where $\boldsymbol{V}^{k}$ denotes the k-th hybrid attention mechanism layer's output vector, and $Norm(\cdot)$ is a normalization operation. Note that, when k is 0, we use the embeddings of nodes $\boldsymbol{V}$ introduced in Subsection 3.3.3 to fill the $\boldsymbol{V}^{0}$, i.e., $\boldsymbol{V}^{0} = \boldsymbol{V}$.

3.5 Decision Model

In what follows, we design the decision model (DI) to convert the AND problem into a binary classification task between each publication pair. Specifically, we first employ the embeddings of pairwise nodes $\boldsymbol{V}_{i}^{-1}$ and $\boldsymbol{V}_{j}^{-1}$ along with their corresponding edges' embeddings $Con_{x}(M_{ij}^{x})$ to construct a triplet $\boldsymbol{T}_{m}$. Then, we consider the structural information loss in the vector space of pairwise nodes' embeddings $\boldsymbol{V}_{i}^{-1}$ and $\boldsymbol{V}_{j}^{-1}$. The similarity between the vectors of pairwise nodes' embeddings is measured using cosine similarity, generating a matrix $\boldsymbol{Cos}$ to represent the structural information in the vector space. Finally, the generated triplets are fed into an MLP classifier for disambiguation. The details of DI are illustrated in Fig.4(b). Formally, it is described as:

$$\boldsymbol{T}_{m} = Con(\boldsymbol{V}_{i}^{-1}, \boldsymbol{V}_{j}^{-1}, Con_{x}(M_{ij}^{x})),$$

$$Cos_{m} = \cos(\boldsymbol{V}_{i}^{-1}, \boldsymbol{V}_{j}^{-1}),$$

$$\boldsymbol{\mathcal{P}_{r}} = MLP(\boldsymbol{T}),$$

where $\boldsymbol{V}^{-1}$ represents the final output layer of the feature extraction model (EX); $\boldsymbol{T}$ is the concatenated triplets matrix, and $m \in \{0, N^{2}\}$; $MLP(\cdot)$ is a multilayer perceptron layer; $\cos(\cdot)$ is used to calculate the cosine similarity between $\boldsymbol{V}_{i}^{-1}$ and $\boldsymbol{V}_{j}^{-1}$; Cos_{m} is the value of the cosine similarity between $\boldsymbol{V}_{i}^{-1}$ and $\boldsymbol{V}_{j}^{-1}$, and $\boldsymbol{Cos} = (Cos_{m})$; and $\boldsymbol{\mathcal{P}_{r}}$ is a N^{2} vector of predicted probabilities used to determine whether pairwise publications belong to a unique author or not.

3.6 Training

In this section, we introduce the details about the ground truth and the loss function. The ground truth in our approach is represented as a graph, following a similar approach as presented in [23]. To optimize our framework, we adopt a joint loss function that includes the log-likelihood loss function (Cross-Entropy-Loss) and the mean absolute error function (L1-Loss). Here, L1-Loss is applied to introduce the structural information loss of the vector space, which encourages the vectors of positive pairs to be closer than negative pairs in the vector space. The details of them are as follows.

3.6.1 Ground Truth Graph

For the convenience of calculation, we define a graph $G_{c} = (D, E)$ as the ground truth based on the given document set $P = \{p_{i}\}$ associated with each author name

and its annotation results $C = \{c_1, c_2, \ldots, c_k\}$[15, 23]. Based on the annotation results C, we first generate positive edge set E_p and negative edge set E_n, i.e., $E_p = \{E_{ij} = 1, \forall(p_i, p_j) \in c_k \times c_k, c_k \in C\}, E_n = \{E_{ij} = 0, \forall(p_i, p_j) \in c_k \times c_{k'}, k \neq k'\}$. The graph G_c is then composed of $E_p \cup E_n = E$ and $D = \{D_i\}$, where each D_i denotes the publication p_i. Next, the $N \times N$ adjacent matrix of the ground truth graph G_c is compressed into N^2 vector $\boldsymbol{\mathcal{Q}_r}$ as labels. It is important to note that each element in the adjacent matrix of G_c is represented by $E_{ij} \in E$, and E_{ij} can be 0 or 1.

3.6.2 Joint Loss Function

The framework is trained by minimizing the joint loss function, which contains the negative log-likelihood loss function (Cross-Entropy-Loss) and the mean absolute error function (L1-Loss). Given an author name α, the joint loss function of the framework is defined as:

$$L = -\frac{1}{N^2}\sum_{m=0}^{N^2-1}(\mathcal{Q}_{r_m}\log(\mathcal{P}_{r_m}) + (1-\mathcal{Q}_{r_m})\log(1-\mathcal{P}_{r_m})) + \frac{1}{N^2}\sum_{m=0}^{N^2-1}\text{abs}(Cos_m - \mathcal{Q}_{r_m}),$$

where $\mathcal{P}_{r_m}$ represents the m-th predicted probability in the vector $\boldsymbol{\mathcal{P}_r}$, $\mathcal{Q}_{r_m}$ denotes the m-th binary label in the vector $\boldsymbol{\mathcal{Q}_r}$, and abs$(\cdot)$ calculates the absolute value of the specified number.

4 Experiment

4.1 Datasets

4.1.1 Illustration of Datasets

To evaluate the performance of our proposed framework EAND, we conduct experiments on two publicly available real world datasets, as described below.

• OAG-WhoisWho[2]. This dataset contains 608,363 documents and 57,138 distinct authors with 642 equivocal author names[26]. For our experiments, we sample 320 author names from the OAG-WhoisWho dataset to construct the our experimental dataset, which contains 341,457 publications. We employ 200 author names for training, 60 author names for validation, and 60 author names for testing[23]. In addition, each publication in the dataset is associated with six features: co-authors, affiliations, venues, years, titles, and keywords.

• AD-AND[3]. To illustrate the effectiveness of our framework in addressing the issues: MRDF and SRDF, we construct a new dataset called AD-AND. The AD-AND dataset is a small-scale dataset collected from AMiner[11] and DBLP. Compared with the OAG-WhoisWho dataset, we expand the training set to facilitate more robust training. In detail, the AD-AND dataset contains 380 author names and includes 130,300 publications. We split the dataset into 260 author names for training, 60 author names for validation, and 60 author names for testing. Each publication in the AD-AND dataset has seven features, including co-authors, affiliations, venues, years, titles, abstracts, and keywords.

4.1.2 Missing of Raw Document Features

To clearly illustrate the MRDF phenomenon, we analyze the experimental datasets, and the results are presented in two tables: Table 2 and Table 3. In particular, Table 2 shows the number and percentage (relative to the total number of publications) of publications missing a specific raw document feature (i.e., "Co-author", "Affiliation", "Venue" or "Keyword") on two datasets. Table 3 presents the number and percentage (relative to the total number of publications) of publications missing one, two, and three raw document features on both datasets. Notably, in Table 3, we use "NMF" to represent the number of missing raw document features for a publication.

Table 2. Statistics of Publications Missing a Specific Raw Document Feature on Two Datasets

Missing Feature	OAG-WhoisWho		AD-AND	
	Number	Percentage(%)	Number	Percentage(%)
Co-author	295	0.09	-	-
Affiliation	112048	32.81	14976	11.49
Venue	25375	7.43	1258	0.97
Keyword	140522	41.15	22876	17.56

[2]https://www.aminer.cn/billboard/whoiswho, May 2021.

[3]https://github.com/wx-qzhou/EAND/ADAND, June 2022.

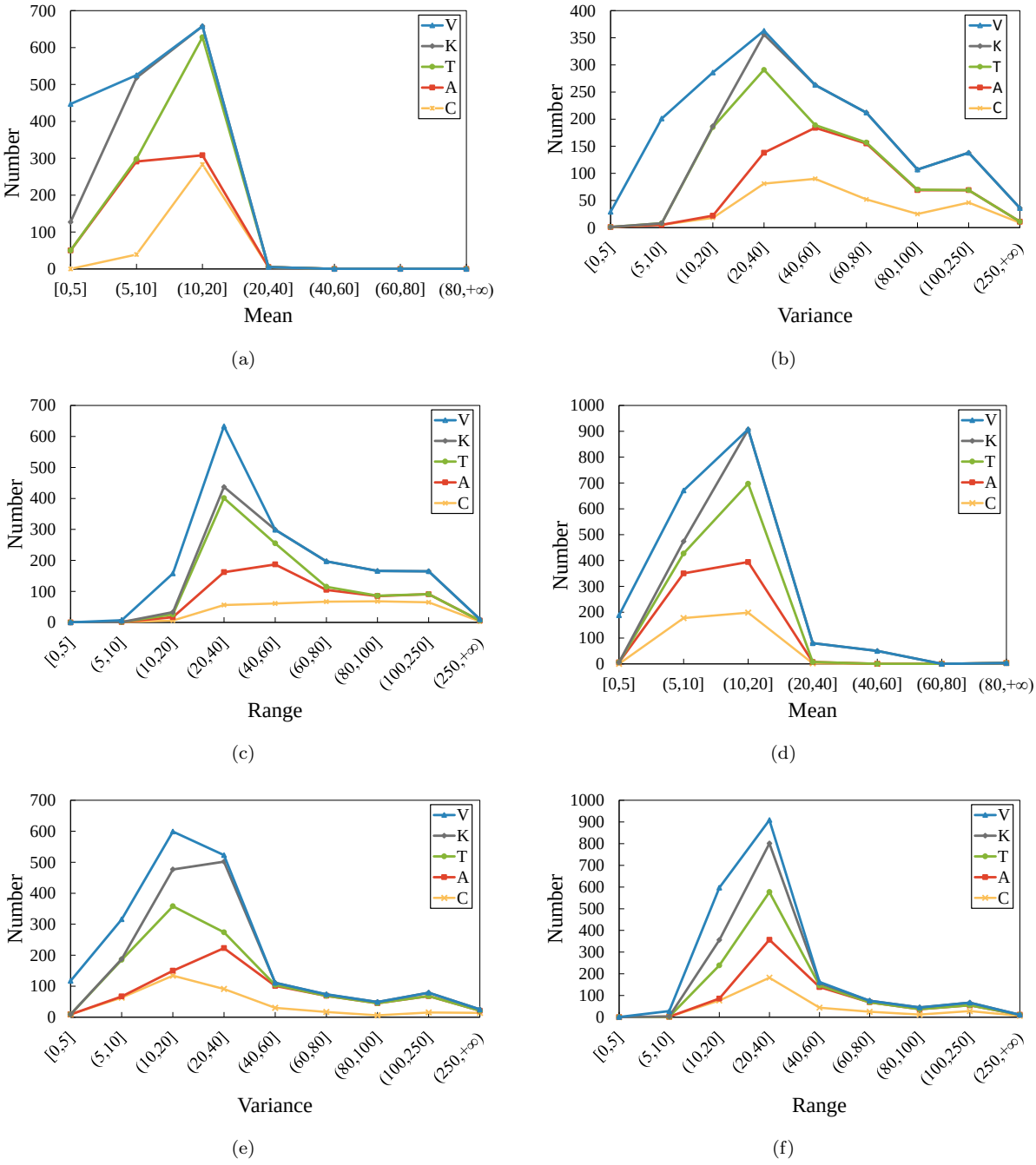


Fig. 6. Statistics of SRDF on two datasets. (a) Mean of LWS on OAG-WhoisWho. (b) Variance of LWS on OAG-WhoisWho. (c) Range of LWS on OAG-WhoisWho. (d) Mean of LWS on AD-AND. (e) Variance of LWS on AD-AND. (f) Range of LWS on AD-AND.

Observed from Table 2, it can be observed that 0.09%, 32.81%, 7.43%, and 41.15% of publications on the OAG-WhoisWho dataset miss co-authors, affiliations, venues, and keywords, respectively. On the AD-AND dataset, 11.49% of publications lack affiliations, 0.97% lack venues, and 17.56% lack keywords. All publications on the OAG-WhoisWho dataset contain titles and years without any missing values. Similarly, on the AD-AND dataset, no publications lack co-authors, titles, or years. It is important to note that abstracts are not included in any publication of the OAG-WhoisWho dataset, and thus this feature is not considered in the analysis. More-

over, Table 3 shows that 32.38% of publications lose one raw document feature, 18.46% lose two, and 4.06% lose three on the OAG-WhoisWho dataset. On the AD-AND dataset, the percentages of publications losing one, two, and three raw document features are 24.02%, 2.41%, and 0.39%, respectively, of the total number of publications. The maximum number of missing raw document features for arbitrary single publication on both datasets is three. Therefore, we present statistics on the number of publications missing one, two, and three raw document features.

Table 3. Statistics of Publications Missing One, Two, or Three Raw Document Features on Two Datasets

NMF	OAG-WhoisWho		AD-AND	
	Number	Percentage(%)	Number	Percentage(%)
1	110566	32.38	31300	24.02
2	63038	18.46	3145	2.41
3	13866	4.06	507	0.39

According to the statistics in Table 2 and Table 3, 54.9% of the total number of publications on the OAG-WhoisWho dataset and 26.82% of the total number of publications on the AD-AND dataset are missing the raw document features. This highlights the gravity of the MRDF issue. Consequently, it is essential to efficiently utilize the multiple feature information available in raw publications.

4.1.3 *Scale-Difference of Raw Document Features*

SRDF is a common phenomenon that some publications have many words in one raw document feature, while others have only a few words. For instance, given the raw document feature: "Co-author", a publication p_i has three words related to this raw document feature, and another publication p_j has fourteen words. We assume that both p_i and p_j are written by a unique author, and p_j includes all the words in p_i. Each word's IDF for both p_i and p_j is 0.1. We estimate the pairwise publication similarity using Jaccard, IDF, improved Jaccard, and improved IDF. Table 4 shows the results, where the improved Jaccard-based result achieves 1, outperforming the traditional Jaccard result (0.214). Similarly, the improved IDF-based result achieves 0.463, which is more precise than the traditional IDF result (0.214).

We conducted further analysis on the datasets to demonstrate the significance of considering this phenomenon when quantifying the similarity of pairwise publications. For each author name α, we have a set of publications $P^\alpha = \{p_1^\alpha, \ldots, p_N^\alpha\}$, and each publication p_i^α is associated with a word set Z_i^x corresponding to the raw document feature x. In all, for the author name α with N publications, we obtain a set of word sets $\{Z_1^x, \ldots, Z_N^x\}$. To describe the SRDF phenomenon, we analyze the quantity distribution of the mean, variance, and range (i.e., the difference between maximum and minimum) of the word sets' lengths (i.e., $|Z_i^x|$) about all author names on each dataset. Formally, for each author name α, we calculate $dist(\{|Z_1^x|, \ldots, |Z_N^x|\})$, where $dist(\cdot)$ can represent the mean, variance, or range operator. Subsequently, we analyze the distribution of these quantities about all author names on each dataset. Note that, if a publication p_i^α lacks this raw document feature x, the length of the corresponding word set is zero, that is, $|Z_i^x| = 0$.

Table 4. Illustrative Results of Pairwise Publication Similarity Using Traditional and Improved Similarity Coefficients

Pairwise Publication	Similarity			
	Jaccard	Improved Jaccard	IDF	Improved IDF
p_i & p_i	1	1	0.214	0.463
p_i & p_j	0.214	1	0.214	0.463
p_j & p_j	1	1	1	1

The statistics of the datasets are presented in Fig.6. Notably, in the following tables and figures, "C", "A", "T", "V", "K", and "Y" are used to represent "Co-author"," Affiliation", "Title", "Venue", "Keyword", and "Year", respectively, if there is no ambiguity. To simplify descriptions, the length of word sets (associated with each raw document feature) of the publications related to each author name is denoted as LWS. On each dataset, the mean of LWS about the most author names is between 5 and 20, the variance of LWS ranges between 10 and 60, and the range of LWS is between 20 and 60. These statistics demonstrate the severity of the SRDF issue. Therefore, improving certain similar-

Table 5. Precision of Author Name Disambiguation on Sampled Author Names from OAG-WhoisWho

Author Name	Size	EAND	MFAND	MA-PairRNN	GANAND	AMiner	Beard	AGAND
A. Kobayashi	229	81.09	81.91	83.01	68.98	72.89	85.94	94.77
Y. Shimada	307	98.11	97.96	72.09	74.32	76	90.47	97.09
Xiaoming Xie	479	94.07	88.18	85.3	74.3	87.77	84.87	90.17
Suqin Liu	518	98.57	92.76	70	84.87	95.19	89.88	80.91
Junyi Li	611	100	99.92	89.2	73.64	98.17	85.41	94.41
Feng Deng	703	99.98	99.29	100	74.62	99.2	97.84	82.28
Xiaodong He	895	99.68	92.67	88.15	74.33	94.66	86.43	81.22
Xiaohua Liu	1335	95.65	98.09	87.19	68.02	95.3	98.51	82.87
Weimin Liu	1485	98.1	76.32	83	72.31	98.33	77.43	91.36
Min Yang	2245	77.02	68.96	84.07	66.14	56.08	76.26	74.41
Avg.	-	94.23	89.61	75.79	73.15	87.36	87.3	86.95

Table 6. Recall of Author Name Disambiguation on Sampled Author Names from OAG-WhoisWho

Author Name	Size	EAND	MFAND	MA-PairRNN	GANAND	AMiner	Beard	AGAND
A. Kobayashi	229	92.44	79.87	83.11	53.12	67.96	73.63	90.69
Y. Shimada	307	96.75	94.35	77.11	72.5	62.89	85.54	40.53
Xiaoming Xie	479	80.75	90.51	78.04	57.84	93.58	74.58	17.48
Suqin Liu	518	99.71	99.68	77	49.59	60.54	55.75	17.9
Junyi Li	611	97.74	92.85	82.01	26.15	85.65	99.2	19.57
Feng Deng	703	98.06	99.11	100	48.86	50.48	77.29	11.54
Xiaodong He	895	93.98	90.74	85.06	57.62	56.82	59.41	19.22
Xiaohua Liu	1335	92.71	98.4	79.99	35	91.3	69.38	13.58
Weimin Liu	1485	97.58	96.17	82	41.29	44.41	86.08	13.3
Min Yang	2245	86.21	68.06	78.1	18.76	19.91	32.14	24.12
Avg.	-	93.59	90.97	74.43	46.07	63.35	71.3	26.79

Table 7. F1 of Author Name Disambiguation on Sampled Author Names from OAG-WhoisWho

Author Name	Size	EAND	MFAND	MA-PairRNN	GANAND	AMiner	Beard	AGAND
A. Kobayashi	229	86.39	80.88	83.06	60.02	70.34	79.31	92.68
Y. Shimada	307	97.42	96.12	74.52	73.4	68.83	87.94	57.19
Xiaoming Xie	479	86.91	89.33	81.51	65.04	90.59	79.4	29.28
Suqin Liu	518	99.13	96.09	73.33	62.6	74	68.82	29.32
Junyi Li	611	98.86	96.26	85.45	38.59	91.48	91.79	32.42
Feng Deng	703	99.01	99.2	100	59.05	66.91	86.36	20.24
Xiaodong He	895	96.75	91.69	86.58	64.92	71.01	70.42	31.09
Xiaohua Liu	1335	94.15	98.25	83.43	46.22	93.26	81.42	23.34
Weimin Liu	1485	97.84	85.1	82.5	52.56	61.18	81.53	23.22
Min Yang	2245	81.36	68.51	80.98	29.22	29.39	45.22	36.43
Avg.	-	93.91	90.28	75.11	56.54	73.44	78.49	40.96

ity coefficients is crucial for achieving a more precise quantification of pairwise publication similarity.

4.2 Baselines

To validate the performance of our proposed framework, we conduct a comprehensive comparison with six state-of-the-art author name disambiguation methods.

• Beard[15]. This model utilizes a well-designed set of similarity features, including author names, titles, etc., to train a distance function for measuring the similarity between each pair of papers. To determine clusters, a semi-supervised algorithm called HAC is employed.

• AGAND[16]. This method constructs three graphs based on document similarity, co-author relationship, and triplets. The final result is generated through agglomerative hierarchical clustering.

• AMiner[11]. The approach consists of two stages: a supervised global stage that fine-tunes the word2vec results and an unsupervised local stage that leverages the first stage and the local linkage graph to improve the global embeddings.

• GANAND[17]. The method builds a generative adversarial framework with two modules. The discriminative module distinguishes whether two papers are from the same author, while the generative module selects possibly homogeneous papers from the heterogeneous information network.

• MA-PairRNN[18]. The method is a novel pairwise node sequence classification framework for name disambiguation. It designs a multi-view graph embedding layer to generate node representation inductively, and employs a Pseudo-Siamese recurrent neural network to learn sequence pair similarity.

• MFAND[19]. The model, presented in our previous work, utilizes the R3JG encoder and employs a binary classifier for disambiguation. R3JG integrates and reconstructs various information, including raw document features, the fusion feature, the local structural information, and the global structural information.

Table 8. Properties of Compared Methods

	GSP	LSP	FE	JLF
GSPJ	√	×	×	√
GLPJ	√	√	×	√
GLAJ	√	√	√	√
GLAN	√	√	√	×

To explore the advantages of global structural perception (GSP), local structural perception (LSP), feature extractor (FE), and the use of the joint loss function (JLF) or only the negative log-likelihood loss function, we design the following comparison methods: GSPJ, GLPJ, GLAJ, and GLAN. The properties of these methods are summarized in Table 8.

Table 9. Results of Author Name Disambiguation on OAG-WhoisWho

Method	Macro			Micro		
	Precision	Recall	F1	Precision	Recall	F1
AGAND	78.52	35.56	48.95	75.97	22.19	34.35
Beard	78.59	56.66	65.84	79.82	56.65	66.27
AMiner	78.39	60.69	68.41	80.83	21.05	33.4
GANAND	69.43	42.69	52.88	67.16	36.34	47.16
MA-PairRNN	79	64.59	70.71	77.95	68.83	73.11
MFAND	74.91	76.39	75.64	71.95	75.58	73.72
EAND	77.95	82.64	80.23	77.59	86.81	81.94

4.3 Experimental Results

We evaluate our proposed method by calculating pairwise precision, recall, and F1 for each sampled author name, as well as micro-precision, micro-recall, micro-F1, macro-precision, macro-recall, and macro-F1 over the entire testing set. Tables 5 – 7 show the performance of different AND methods on sampled author names of different sizes, sampled from the OAG-WhoisWho dataset. As a result of leveraging multiple feature information and the feature extraction model (EX), our proposed method consistently outperforms other state-of-the-art methods in most cases. The average pairwise F1 of 10 samples in our method surpasses other methods by at least +3.63%.

Table 10. Results of Author Name Disambiguation on AD-AND

Method	Macro			Micro		
	Precision	Recall	F1	Precision	Recall	F1
AGAND	78.08	56.74	65.72	70.86	50.75	59.14
Beard	79.89	76.66	78.53	75.92	74.12	75.01
AMiner	84.26	72.68	78.04	80.84	68.51	74.17
GANAND	94.97	67.74	79.08	96.46	66.71	78.87
MA-PairRNN	80.02	83.03	81.5	67.6	75.37	71.27
MFAND	85.51	78.07	81.62	68.87	80.53	74.25
EAND	86.88	83.07	84.93	79.99	83.3	81.61

To further validate the superiority of our proposed method over state-of-the-art approaches, we present the experimental results obtained from two real-world datasets. Detailedly, the results in Table 9 and Table 10 demonstrate that our proposed method outperforms other baselines by at least +8.22% in micro-F1 and +4.59% in macro-F1 on the entire OAG-WhoisWho dataset, and by at least +2.74% in micro-F1 and +3.31% in macro-F1 on the entire AD-AND dataset. Specifi-

cally, on the OAG-WhoisWho dataset, our proposed method significantly outperforms the baselines in terms of macro-F1 (+31.28% over AGAND, +14.39% over Beard, +11.82% over AMiner, +27.35% over GANAND, +9.52% over MA-PairRNN, and +4.59% over MFAND relatively). Our proposed method also outperforms the baselines in terms of macro-F1 (+19.21% over AGAND, +6.4% over Beard, +6.89% over AMiner, +5.85% over GANAND, +3.43% over MA-PairRNN, and +3.31% over MFAND relatively) on the AD-AND dataset.

Table 11. Contributions of Each Component Based on Macro

Module	OAG-WhoisWho			AD-AND		
	Precision	Recall	F1	Precision	Recall	F1
GSPJ	73.32	84.4	78.47	76.54	85.6	80.82
GLPJ	76.55	81.91	79.14	85.01	81.49	83.21
GLAJ	77.95	82.64	80.23	86.88	83.07	84.93

Table 12. Contributions of Each Component

Module	OAG-WhoisWho		AD-AND	
	Micro-F1	macro-F1	Micro-F1	Macro-F1
GLAN	79.18	80.07	78.41	84.24
GLAJ	81.94	80.23	81.61	84.93

Taking a deeper dive into the experimental results, our framework showcases a superior performance in the F1 score compared to all state-of-the-art methods, owing to its emphasis on achieving a balance between precision and recall. As depicted in Table 9 and Table 10, we find that our model achieves better recall scores than other methods. This is attributed to the effective utilization of global structural information and raw document features, which most methods do not incorporate. Moreover, we observe that GANAND and MA-PairRNN can achieve better performance on the AD-AND dataset compared to the OAG-WhoisWho dataset. This is because the OAG-WhoisWho dataset faces the severe MRDF issue compared to the AD-AND dataset. Furthermore, on the OAG-WhoisWho dataset, MA-PairRNN achieves better precision than GANAND. This is because MA-PairRNN applies the multi-view attention mechanism to explore the importance of the meta-path and divides publications into small blocks based on affiliations. Nevertheless, while MA-PairRNN primarily captures the importance of the meta-path, which represents local structural information, it overlooks crucial information from the fusion feature and the global structural information. This limitation in MA-PairRNN's approach significantly constrains its overall performance. From Table 10, GANAND exhibits better precision than other models. This is because GANAND uses adversarial representation learning, which generates high-quality samples. However, GANAND cannot achieve better precision than other models as observed from Table 9 because the OAG-WhoisWho dataset suffers from the serious MRDF issue. Notably, except for EAND, MFAND performs better than other models. This is because MFAND considers multiple feature information and addresses the MRDF issue to a certain extent. Nonetheless, compared with MFAND, our proposed method achieves better performances in the precision, recall, and F1 scores, as observed from Table 9 and Table 10. This is because we employ the hybrid attention mechanism to take full advantage of the contributions of each raw document feature while also considering the following issues: MRDF and SRDF. Overall, our proposed framework EAND achieves the best performance in almost all evaluation metrics, showcasing its effectiveness and superiority.

4.4 Ablation Analysis

To evaluate the performance of each module, we present the results at different stages in Table 11 and Table 12. From Table 11, it can be seen that GSP can achieve better recall than other modules. This aligns with the findings from the MFAND method, which highlighted the importance of global structural information for improved recall. By incorporating the attention mechanism, we further improve recall scores by exploring the importance of nodes in the graph relative to the current node, leveraging the information of edges and the corresponding raw document feature. Additionally, when LSP is added to GSPJ, the model achieves better precision. This is because LSP is designed to assign varying weights to different neighboring nodes in a node's neighborhood and explore the contributions of each raw

document feature under conditions of local structure. However, this module may result in a reduction of recall scores. FE effectively extracts the information of each raw document feature from a hybrid embedding to reconstruct the information of each raw document feature. As evident from Table 11, our framework achieves a better performance in both macro-F1 and micro-F1 when FE is added to GLPJ. In Table 12, we present the impact of L1-Loss inclusion or exclusion. It can be seen that L1-Loss has a more substantial influence on micro-F1 than macro-F1. This observation is expected since L1-Loss encourages the vectors of positive pairs to be closer than negative pairs in the vector space.

4.5 Estimation of Parameters

In this section, we investigate how the nodes' neighborhood parameters (length and number of walks) affect the performance of our method on the datasets: OAG-WhoisWho and AD-AND. As observed from Fig.7, our proposed method achieves better results in macro-F1 when the length of walks reaches around 16. Similarly, as seen in Fig.8, our proposed method achieves better performances in macro-F1 when the number of walks is set to 8 and 12. This is because the noise is introduced when the length and number of walks are set to large values. Conversely, when the length and number of walks are set to small values, the nodes will contain less useful information. Both these parameters have a relative impact on the performance of our proposed method, but the performance differences are not significant, especially for the AD-AND dataset.

4.6 Selection of Raw Document Features

Each raw document feature exhibits different discriminative capability[19, 20] and makes different contributions to the AND problem. In light of this, we explore which raw document features can be selected by our proposed method. Notably, we take the MRDF issue into consideration to investigate the importance of each raw document feature used in our approach. As highlighted by [27] and [20], the co-author relationship of an author is considered a strong discriminative feature. Our observations from Fig.9 provide further support to this finding, as we notice that co-authors and affiliations are the most influential discriminative features. Additionally, titles, venues, and keywords have some influence on the AND problem, while years have the least influence.

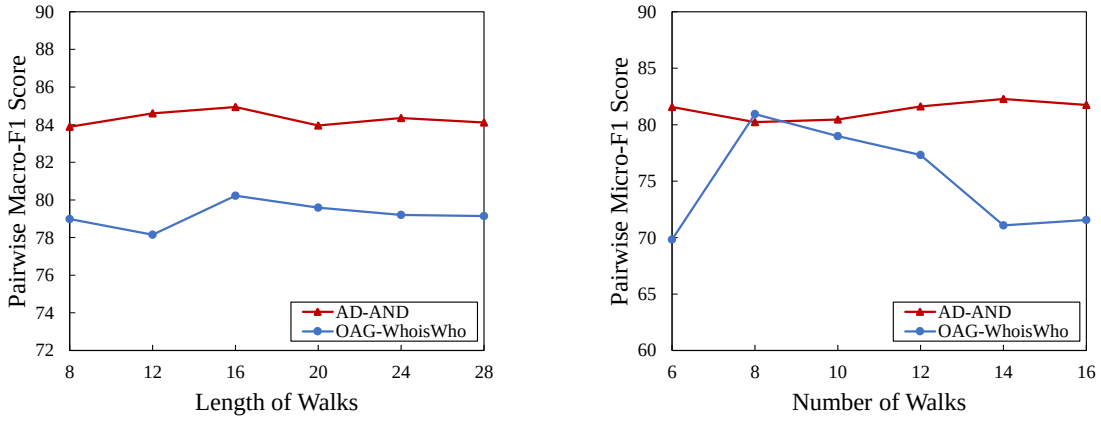
To investigate how different raw document features affect the performance of our proposed approach, we add the features one by one, starting with co-authors and affiliations. Observed from Fig.10, when several raw document features are considered, including co-authors, affiliations, titles, venues, and keywords, our proposed approach achieves better performances on both datasets. However, the inclusion of years has a negative effect on our proposed method, which is expected. This is because co-authors, affiliations, titles, venues, and keywords are highly relevant to the author's identity, whereas years are not. Generally, authors with the same name may publish articles in the same year, while authors with a unique identity may publish articles in different years. As a result, years do not contribute significantly to our proposed method.

5 Related Work

Author name disambiguation has been extensively studied using various methods[16, 17, 19]. However, existing methods only exploit a part of different types of information, such as raw document features, the fusion feature, the local structural information, and the global structural information and fail to fully leverage the contributions of each raw document feature. The state-of-the-art methods for author name disambiguation can be categorized into two groups: context-based and graph-based methods.

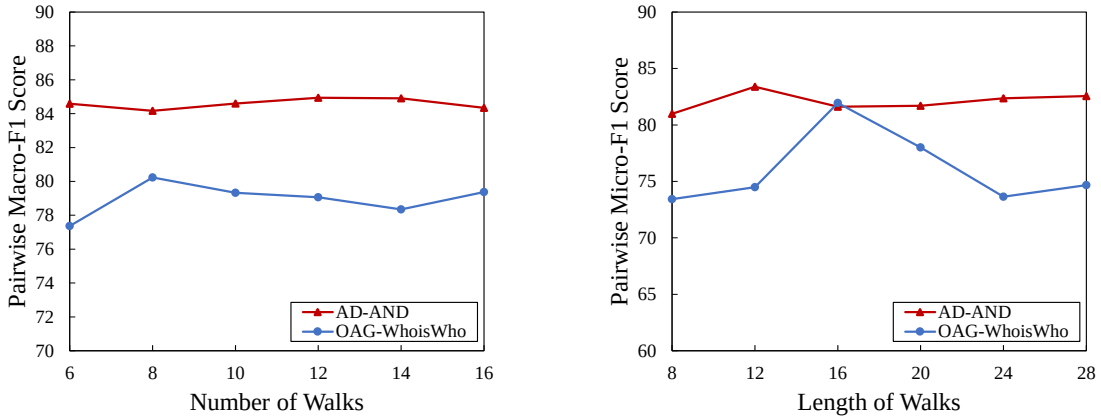
5.1 Context-Based Methods

Context-based methods consider all raw document features to generate the context, represented by the feature vectors. They leverage supervised learning methods to learn a pairwise function between publications based on these feature vectors. Subsequently, the learned



(a) (b)

Fig. 7. Effect of different lengths of random walks on the AND results. (a) Macro-F1 score. (b) Micro-F1 score.



(a) (b)

Fig. 8. Effect of different numbers of random walks on the AND results. (a) Macro-F1 score. (b) Micro-F1 score.

pairwise function is used to predict whether two publications written by authors with the same name belong to a unique identity. For instance, Han *et al.*[28] utilized naive bayes and SVM (supported vector machine) to address the AND problem. Han *et al.*[14] employed TF-IDF and NTF to define similarity functions and calculate document similarity, while using k-way spectral clustering for disambiguation. Yoshida *et al.*[29] proposed a two-stage clustering method to learn better feature representations via the first clustering step. Müller[30] employed a deep neural network to solve the AND problem. Kim *et al.*[12] introduced a hybrid method that extracts structure-aware features and global features, using gradient boosted trees (GBT) and deep neural network (DNN) to experiment with the disambiguation results, respectively. Jhawar *et al.*[13] conducted experiments with two ensemble-based classification algorithms, namely, random forest and gradient boosted decision trees, on a publicly available corpus of manually disambiguated author names from PubMed.

5.2 Graph-Based Methods

Graph-based methods tackle the AND problem by utilizing graphical models that capture information from neighbors in the graph topology. For example, Fan *et al.*[31] constructed a graph by collapsing all co-authors

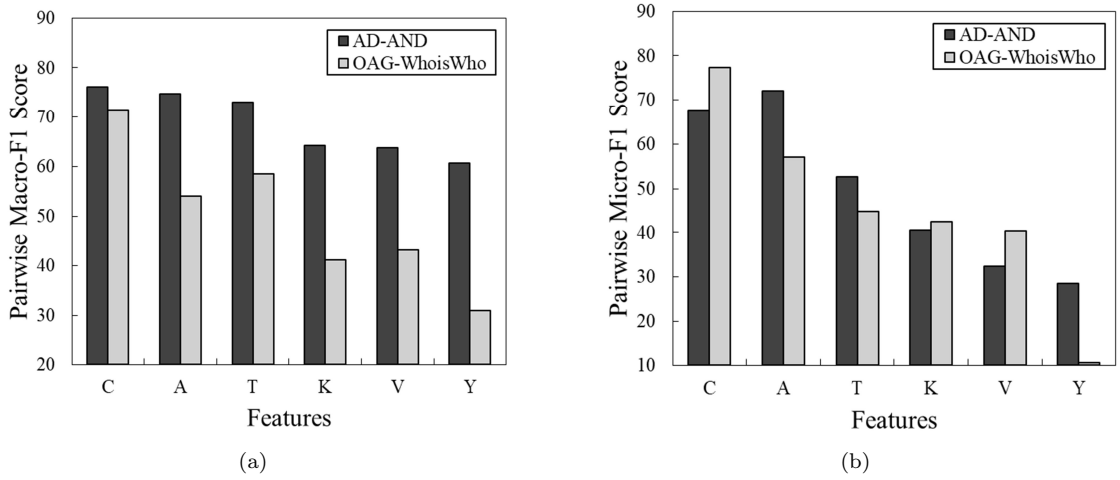


Fig. 9. Effect of different raw document features on the AND results. (a) Macro-F1 score. (b) Micro-F1 score.

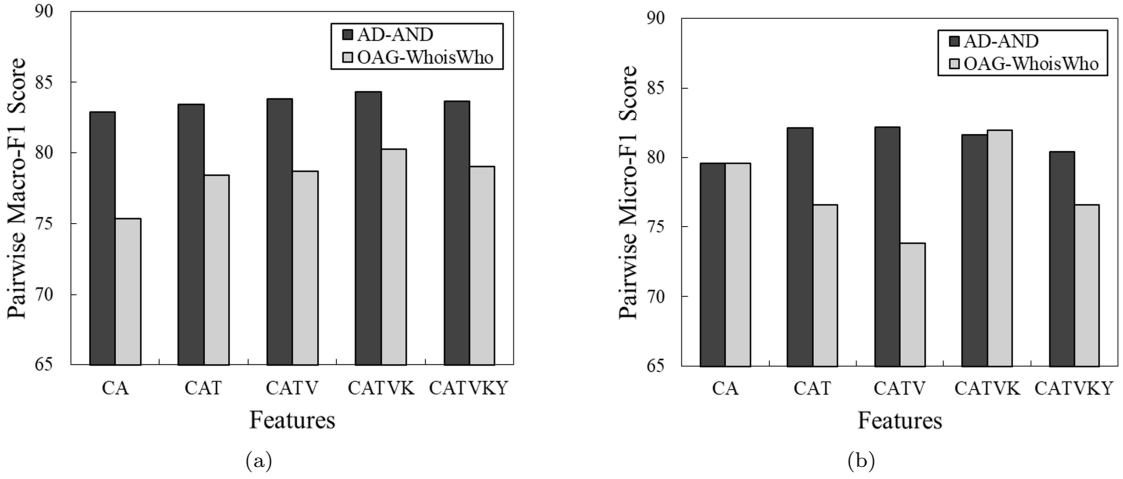


Fig. 10. Effect of selecting different raw document features by adding the features one by one on the AND results. (a) Macro-F1 score. (b) Micro-F1 score.

with identical names to a single node and measured pairwise node distance based on the number of valid paths. Tang *et al.*[32] employed hidden markov random fields (HMRF) to model node and edge features in a unified probabilistic framework. Zhang and Hasan [16] constructed three graphs based on document similarity and co-author relationships, and learned graph embeddings from them. Zhang *et al.*[11] designed a supervised global stage to fine-tune word2vec results and applied an unsupervised local stage based on the first stage and the local linkage graph to improve global embeddings. Wang *et al.*[17] proposed a generative adversarial framework, where the discriminative module distinguishes whether two papers are from the same author, and the generative module selects possibly homogeneous papers directly from the heterogeneous information network. Sun *et al.*[18] introduced a novel pairwise node sequence classification framework based on the multi-view graph embedding layer and Pseudo-Siamese recurrent neural network for name disambiguation. Zhou *et al.*[19] used an encoder called R3JG to integrate and reconstruct information (i.e., raw document features, the fusion feature, the local structural information, and the global structural information) while they used a binary classifier for disambiguation.

Despite the extensive research conducted in this do-

main, to the best of our knowledge, there has been no work considering all the above-mentioned information for the AND problem so far or fully utilizing the contributions of each raw document feature, while also addressing the MRDF and SRDF issues. To overcome these limitations, we propose a unified framework named EAND. Specifically, our framework aims to mitigate the effect of MRDF and SRDF, take full advantage of the contributions of each raw document feature and effectively address the AND problem.

6 Conclusions

In this paper, we proposed EAND, a unified framework that considers diverse information types, including raw document features, the fusion feature, the local structural information, and the global structural information. Our framework introduces a novel feature extraction model that captures the influence between multiple types of feature information, fully leveraging the contributions of each raw document feature to effectively tackle the AND problem. To mitigate the MRDF issue, we designed a novel generating strategy that extracts local structure information from the fusion feature and global structural information from raw document features. Additionally, we addressed the SRDF issue by utilizing improved similarity coefficients for quantifying the similarity of pairwise publications. The L1-Loss function is used to encourage positive pairs to be closer than negative pairs in the vector space by introducing structural information loss. Meanwhile, different pruning strategies are employed for feature graphs to effectively remove noise. Extensive experimental results demonstrated our proposed method's superiority over state-of-the-art methods, with an improvement of at least +2.74% in the micro-F1 score and +3.31% in the macro-F1 score. In the future, we will extend the proposed framework to recommender systems.

References

[1] Gupta S, Duhan N, Bansal P. An approach for focused crawler to harvest digital academic documents in online digital libraries. *International Journal of Information Retrieval Research*, 2019, 9(3):23–47. DOI: 10.4018/IJIRR.2019070103.

[2] Chikazawa Y, Katsurai M, Ohmukai I. Multilingual author matching across different academic databases: A case study on KAKEN, DBLP, and PubMed. *Scientometrics*, 2021, 126(3):2311–2327. DOI: 10.1007/s11192-020-03861-3.

[3] Tang J, Zhang J, Yao L, Li J, Zhang L, Su Z. ArnetMiner: Extraction and mining of academic social networks. In *Proc. the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2008, pp. 990–998. DOI: 10.1145/1401890.1402008.

[4] Ferreira A A, Gonçalves M A, Laender A H F. Automatic Disambiguation of Author Names in Bibliographic Repositories. Morgan and Claypool Publishers, 2020. DOI: 10.2200/S01011ED1V01Y202005ICR070.

[5] Martín-Martín A, Thelwall M, Orduña-Malea E, López-Cózar E D. Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: A multidisciplinary comparison of coverage via citations. *Scientometrics*, 2021, 126(1):871–906. DOI: 10.1007/s11192-020-03690-4.

[6] Yin X, Han J, Yu P S. Object Distinction: Distinguishing objects with identical names. In *Proc. the 23rd International Conference on Data Engineering, ICDE*, April 2007, pp. 1242–1246. DOI: 10.1109/ICDE.2007.368983.

[7] Li X, Morie P, Roth D. Identification and tracing of ambiguous names: Discriminative and generative approaches. In *Proc. the Nineteenth National Conference on Artificial Intelligence, Sixteenth Conference on Innovative Applications of Artificial Intelligence*, July 2004, pp. 419–424. DOI: 10.5555/1597148.1597217.

[8] Pooja K M, Mondal S, Chandra J. A graph combination with edge pruning-based approach for author name disambiguation. *Journal of the Association for Information Science and Technology*, 2020, 71(1):69–83. DOI: 10.1002/asi.24212.

[9] Ma Y, Wu Y, Lu C. A graph-based author name disambiguation method and analysis via information theory. *Entropy*, 2020, 22(4):416–433. DOI: 10.3390/e22040416.

[10] Zhang L, Ban Z. Author name disambiguation based on rule and graph model. In *Natural Language Processing and Chinese Computing - 9th CCF International Conference, NLPCC*, October 2020, pp. 617–628. DOI: 10.1007/978-3-030-60450-9_49.

[11] Zhang Y, Zhang F, Yao P, Tang J. Name disambiguation in AMiner: Clustering, maintenance, and human in the loop. In *Proc. the 24th ACM SIGKDD International Conference*

on Knowledge Discovery & Data Mining, KDD, July 2018, pp. 1002–1011. DOI: 10.1145/3219819.3219859.

[12] Kim K, Rohatgi S, Giles C L. Hybrid deep pairwise classification for author name disambiguation. In *Proc. the 28th ACM International Conference on Information and Knowledge Management, CIKM*, November 2019, pp. 2369–2372. DOI: 10.1145/3357384.3358153.

[13] Jhawar K, Sanyal D K, Chattopadhyay S, Bhowmick P K, Das P P. Author name disambiguation in PubMed using ensemble-based classification algorithms. In *Proc. the ACM/IEEE Joint Conference on Digital Libraries, JCDL*, August 2020, pp. 469–470. DOI: 10.1145/3383583.3398568.

[14] Han H, Zha H, Giles C L. Name disambiguation in author citations using a k-way spectral clustering method. In *ACM/IEEE Joint Conference on Digital Libraries, JCDL*, June 2005, pp. 334–343. DOI: 10.1145/1065385.1065462.

[15] Louppe G, Al-Natsheh H T, Susik M, Maguire E J. Ethnicity sensitive author disambiguation using semi-supervised learning. In *Knowledge Engineering and Semantic Web - 7th International Conference, KESW*, September 2016, pp. 272–287. DOI: 10.1007/978-3-319-45880-9_21.

[16] Zhang B, Hasan M A. Name disambiguation in anonymized graphs using network embedding. In *Proc. the 2017 ACM on Conference on Information and Knowledge Management, CIKM*, November 2017, pp. 1239–1248. DOI: 10.1145/3132847.3132873.

[17] Wang H, Wang R, Wen C, Li S, Jia Y, Zhang W, Wang X. Author name disambiguation on heterogeneous information network with adversarial representation learning. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI*, June 2020, pp. 238–245. DOI: 10.1609/aaai.v34i01.5356.

[18] Sun Q, Peng H, Li J, Wang S, Dong X, Zhao L, Yu P S, He L. Pairwise learning for name disambiguation in large-scale heterogeneous academic networks. In *20th IEEE International Conference on Data Mining, ICDM*, November 2020, pp. 511–520. DOI: 10.1109/ICDM50108.2020.00060.

[19] Zhou Q, Chen W, Wang W, Xu J, Zhao L. Multiple features driven author name disambiguation. In *IEEE International Conference on Web Services, ICWS*, September 2021, pp. 506–515. DOI: 10.1109/ICWS53863.2021.00071.

[20] Santana A F, Gonçalves M A, Laender A H F, Ferreira A A. On the combination of domain-specific heuristics for author name disambiguation: The nearest cluster method. *International Journal on Digital Libraries*, 2015, 16(3):229–246. DOI: 10.1007/s00799-015-0158-y.

[21] Kim J, Owen-Smith J. Orcid-linked labeled data for evaluating author name disambiguation at scale. *Scientometrics*, 2021, 126(3):2057–2083. DOI: 10.1007/s11192-020-03826-6.

[22] Godoi T A, Silva Torres R, Carvalho A M B R, Gonçalves M A, Ferreira A A, Fan W, Fox E A. A relevance feedback approach for the author name disambiguation problem. In *13th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL*, July 2013, pp. 209–218. DOI: 10.1145/2467696.2467709.

[23] Xiao Z, Zhang Y, Chen B, Liu X, Tang J. A framework for constructing a huge name disambiguation dataset: Algorithms, visualization and human collaboration. *arXiv:2007.02086*, 2020, https://arxiv.org/abs/2007.02086, July 2020. DOI: abs/2007.02086.

[24] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD*, August 2014, pp. 701–710. DOI: 10.1145/2623330.2623732.

[25] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, June 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[26] Chen B, Zhang J, Tang J, Cai L, Wang Z, Zhao S, Chen H, Li C. CONNA: Addressing name disambiguation on the fly. *IEEE Transactions on Knowledge and Data Engineering*, 2022, 34(7):3139–3152. DOI: 10.1109/TKDE.2020.3021256.

[27] Cota R G, Ferreira A A, Nascimento C, Gonçalves M A, Laender A H F. An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations. *Journal of the Association for Information Science and Technology*, 2010, 61(9):1853–1870. DOI: 10.1002/asi.21363.

[28] Han H, Giles C L, Zha H, Li C, Tsioutsiouliklis K. Two supervised learning approaches for name disambiguation in author citations. In *ACM/IEEE Joint Conference on Digital Libraries, JCDL*, June 2004, pp. 296–305. DOI: 10.1145/996350.996419.

[29] Yoshida M, Ikeda M, Ono S, Sato I, Nakagawa H. Person name disambiguation by bootstrapping. In *Proc. the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR*, July 2010, pp. 10–17. DOI: 10.1145/1835449.1835454.

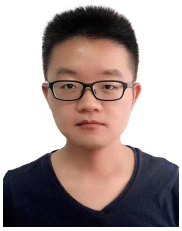
[30] Müller M. Semantic author name disambiguation with word embeddings. In *Research and Advanced Technology for Digital Libraries - 21st International Conference on Theory and Practice of Digital Libraries, TPDL*, September 2017, pp. 300–311. DOI: 10.1007/978-3-319-67008-9_24.

[31] Fan X, Wang J, Pu X, Zhou L, Lv B. On graph-based name

disambiguation. *ACM Journal of Data and Information Quality*, 2011, 2(2):1–23. DOI: 10.1145/1891879.1891883.

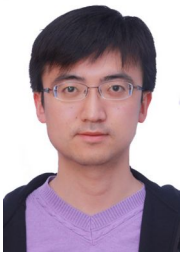
[32] Tang J, Fong A C M, Wang B, Zhang J. A unified probabilistic framework for name disambiguation in digital library. *IEEE Transactions on Knowledge and Data Engineering*, 2012, 24(6):975–987. DOI: 10.1109/TKDE.2011.13.

Qian Zhou received his M.S. degree in computer science and technology from Soochow University, Suzhou, in 2022. Currently, he is a research assistant at the School of Computer Science and Technology, Soochow University, Suzhou. His current research interests mainly include data mining, deep learning, and natural language processing.



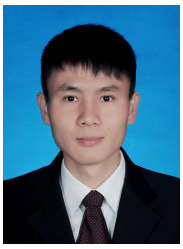
Wei Chen is currently an associate professor in the School of Computer Science and Technology at Soochow University, Suzhou. He received his Ph.D. degree in computer science from Soochow University, Suzhou, in 2018. His research interests include heterogeneous information network analysis, cross-platform linkage and recommendation, spatio-temporal database, and knowledge graph embedding and refinement.

Peng-Peng Zhao received his Ph.D. degree in computer science from Soochow University, Suzhou, in 2008. He is a professor at the School of Computer Science and Technology at Soochow University, Suzhou. From 2016 to 2017, he was a visiting scholar, working at the Data Mining and Business Analysis Laboratory at Rutgers University, New Jersey. He has published more than 60 papers in prestigious international conferences and journals, including ACM MM, AAAI, IJCAI, ICDM, CIKM, DASFAA, and ICME. He was a program committee member of some international conferences, such as AAAI, IJCAI, CIKM, and PAKDD. His current research interests include data mining, deep learning, big data analysis, and recommender systems.



An Liu is a professor at the School of Computer Science and Technology, Soochow University, Suzhou. Prior to that in 2014, he was a senior research associate at the Joint Research Center of City University of Hong Kong (CityU), Hong Kong, and the University of Science and Technology of China (USTC), Hefei. He received his Ph.D. degree in computer science from both CityU, Hong Kong and USTC, Hefei, in 2009. His research interests include security, privacy, and trust in emerging applications, cloud computing, and services computing.

Jia-Jie Xu is an associate professor at the School of Computer Science and Technology, Soochow University, Suzhou. He got his Ph.D. and M.S. degrees from the Swinburne University of Technology, Melbourne and the University of Queensland, Brisbane in 2011 and 2006 respectively. Before joining Soochow University in 2013, he worked as an assistant professor at the Institute of Software, Chinese Academy of Sciences. His research interests mainly include spatio-temporal database systems, big data analytics, and workflow systems.



Jian-Feng Qu is a lecturer at the School of Computer Science and Technology, Soochow University, Suzhou. He received his B.S., M.S., and Ph.D. degrees in Computer Science from Jilin University, Changchun, in 2013, 2016 and 2019 respectively. His research interests include information extraction, data mining, natural language processing, and deep learning.

Lei Zhao is a professor at the School of Computer Science and Technology, Soochow University, Suzhou. He received his Ph.D. degree in computer science from Soochow University, Suzhou, in 2006. His recent research is to analyze large graph databases in an effective, efficient, and secure way. He has published over 150 papers including more than 40 published in well-known journals and conferences such as ICDE, DASFAA, WISE, JCST.