Knowledge and Information Systems
https://doi.org/10.1007/s10115-019-01354-1

REGULAR PAPER

HEEL: exploratory entity linking for heterogeneous information networks

Chengyu Wang[1] · Xiaofeng He[1] · Aoying Zhou[2]

Received: 11 December 2017 / Revised: 7 March 2019 / Accepted: 8 March 2019

Abstract
A heterogeneous information network (HIN) is a ubiquitous data model, consisting of multiple types of entities and relations. Names of entities in HINs are inherently ambiguous, making it difficult to fully disambiguate a HIN. In this paper, we introduce the task of exploratory entity linking for HINs. Given a partially disambiguated HIN, we aim at linking ambiguous names to disambiguated entities in the HIN if their referent entities are present. We also try to "explore" other alternatives by discovering new entities and adding them to the HIN. A partial classification EM-based approach is proposed to address this task. We present a constrained probability propagation model to link surface names to entities in the HIN. New entity detection process is modeled as a maximum edge weight clique problem. Experiments illustrate that our method outperforms state-of-the-art methods for entity linking with HINs and author name disambiguation.

Keywords Heterogeneous information network · Exploratory entity linking · Partial classification EM · Author name disambiguation

1 Introduction

A *heterogeneous information network* (**HIN**) is a semantic network containing multiple types of entities and relations. Typical HINs include social networks, bibliographic networks and domain-specific knowledge bases [13]. The mining of HINs improves the performance of applications such as object classification, link prediction.

✉ Xiaofeng He
xfhe@sei.ecnu.edu.cn

Chengyu Wang
chywang2013@gmail.com

Aoying Zhou
ayzhou@dase.ecnu.edu.cn

[1] School of Computer Science and Software Engineering, East China Normal University, Shanghai 200062, China

[2] School of Data Science and Engineering, East China Normal University, Shanghai 200062, China

Published online: 01 April 2019

Names of entities in HINs are inherently ambiguous [20]. For instance, different authors in bibliographic networks can have identical or similar names. In DBLP, there are 2370 highly ambiguous author names with disambiguation pages. For instance, the name *Yang Liu* refers to 33 distinct researchers, each linking to their own papers. A total of 735 papers authored by *Yang Liu* have not been assigned to any of the 33 researchers. When new papers are published, the authors are added DBLP without the *incremental disambiguation* process [34]. This *partial disambiguation* phenomenon of HINs harms the linking quality of entities, consequently harming the performance of other tasks which take existing HINs as knowledge sources.

To address this issue, a considerable amount of research work has been conducted under the framework of name disambiguation (**ND**) [12,18,31,33] and Entity Linking (**EL**) [10,22,23,32]. ND groups identical or similar surface names into clusters where each cluster represents the same underlying entity. The connections between surface name clusters and entities in existing HINs are not directly modeled. EL links a surface name to its referent entity in knowledge bases. But there is few work addressing EL with HINs [20]. Additionally, HINs tend to be incomplete [31]. Current methods in EL fail to discover new entities.

This paper addresses the task of *Exploratory Entity Linking for HINs* (**HEEL**). Given a *partially disambiguated* HIN, we first construct a *Fully Disambiguated Subgraph* (**FDS**). For each ambiguous name, HEEL tries to link it to either an existing disambiguated entity in the FDS, or a surface name cluster that represents a new entity that is not in the FDS yet, otherwise assigns it a value NIL. Take the DBLP case as an example. HEEL can link the name *Yang Liu* in a publication record to (i) an existing researcher named *Yang Liu* in DBLP, (ii) an author named *Yang Liu* that are not in DBLP, together with a list of his/her most possible publications or (iii) NIL, meaning that he/she is not an existing author in DBLP and we cannot detect a new author with a publication list with high confidence.

We propose a partial classification expectation maximum (**PC-EM**) framework to solve the HEEL problem. It consists of three iterative steps: E-step, PC-step and M-step. In E-step, given a publication record with an ambiguous author name, a constrained probability propagation (**CPP**) model is employed to estimate the probability distribution of referent authors. In PC-step, we first link a part of author names in publication records to their referent authors if the model prediction has high confidence. For the rest of the author names, we try to discover a new author in each iteration, represented by a collection of author names and the corresponding publication records. This problem is modeled as the maximum edge weight clique problem (**MEWC**). In M-step, parameters of the CPP model are updated via a constrained gradient accent algorithm. After the iterative process ends, we assign the remaining author names the value NIL, meaning no referent authors can be linked to or detected.

In summary, we make the following major contributions:

- We introduce the HEEL task to disambiguate entities in HINs. A PC-EM framework is proposed to solve this task.
- We propose a CPP model to estimate the probability of an author name being linked to an author in a HIN. A partial classification technique and a MEWC detection algorithm are presented to discover new authors.
- Extensive experiments over multiple bibliographic datasets illustrate that the proposed approach are effective over three tasks: (i) author name linking with HINs, (ii) author name disambiguation and (iii) new author discovery.

The rest of this paper is organized as follows. Section 2 summarizes the related work. We introduce our task in Sect. 3. Our approach is described in Sect. 4 with experiments presented in Sect. 5. We conclude our paper in Sect. 6.

2 Related work

The research on HEEL is inspired by ND (especially author name disambiguation), EL and HIN mining. We overview the related work from the three aspects.

2.1 Author name disambiguation

ND deals with the situation where different entities share identical or similar surface names. The key step is to learn the semantics of entities such that identical or similar surface names that refer to different entities can be distinguished, such as the BoW model [2], the author-topic-community model [14]. With the popularity of online encyclopedias, the knowledge of entities can be automatically mined. Bunescu and Pasca [3] design an SVM kernel based on entity descriptions in Wikipedia. Recently, by using deep learning techniques, the features of surface names can be represented as low-dimensional dense vectors. Zwicklbauer et al. [35] employ semantic embeddings to represent entities for entity disambiguation.

Due to the prevalence of author name ambiguity, a lot of methods have been proposed to disambiguate authors over bibliographic datasets. DISTINCT [33] distinguishes different objects with identical names based on set resemblance and random walk. Li et al. [15] cluster author names in temporal records by considering temporal association between publication records. Wang et al. [31] introduce an active learning approach for disambiguating person names through a pairwise factor graph model. Qian et al. [19] combine machine learning models with human judgment to improve the performance of author disambiguation. For online name disambiguation, Zhang et al. [34] present a Bayesian classification model to capture the temporal dynamics of record streams. Additionally, CSLR [16] employs a categorical distribution similarity measure to disambiguate authors. This task is also addressed in the data challenge of SIGIR'14 [4] and the KDD Cup 2013 [26]. For a comprehensive overview of author name disambiguation, please refer to the survey paper [9]. Real-life applications include Google Scholar, Microsoft Academic Search, AMiner [29], etc. These systems create profiles for each researcher and support author search functionality. For example, in [18], the system ALIAS is presented to provide semantic service for duplicated author name search and top-k similar author search. Chiang et al. [6] support multiple types of academic search based on random walk with restart.

2.2 Entity linking

With the development of large-scale knowledge bases, surface names from raw data sources can be directly linked to a certain entity in the knowledge base by EL. A recent survey is presented in [21]. In the literature, Ganea et al. [10] introduce a probabilistic Bag-Of-Hyperlinks model to link all the entities in a document collectively. Shen et al. [23] employ the YAGO taxonomy as an additional knowledge source to improve linking performance. Li et al. [17] present a generative topic model to link surface names to entities with linkless knowledge bases. Wang et al. [30] employ a pairwise linking technique to detect linking errors in Wikipedia.

Besides single name-entity matching, Han et al. [11] propose to link a collection of names to entities based on collective decision. Sil and Florian [25] design a general framework for language-independent EL, which models contexts of surface names and entities uniformly and performs EL by probabilistic inference. However, none of the prior work has the capacity for "exploratory" linking. In contrast, our method tries to discover new entities and thus has the potential for knowledge base or HIN population. Many studies also focus on linking surface names in other data sources, including Web lists [22], tweets [32], queries [7], etc.

2.3 HIN mining

A third thread of related work is mining HINs. The concept of HIN is first proposed in [28]. It has a strong expressive power to integrate information extracted from different data sources. In some research work, HINs are also categorized as domain-specific knowledge bases [13, 20]. The challenge of mining HINs is structural analysis on rich semantics embedded in multiple types of entities and links. In a HIN, the complex relations of entities are usually modeled as meta-paths [13,24,27]. For example, the "Author–Paper–Author" (A–P–A) meta-path expresses the co-author relation between authors. The work most related to ours is [20], which links surface names in documents to entities in HINs based on meta-path constrained random walk. In contrast, we aim to link ambiguous names to disambiguated entities in order to turn a partially disambiguated HIN into a fully disambiguated one. Another issue is the NIL problem. In [20], surface names are linked to the most probable entity without addressing the NIL situation. In our work, we also discover new entities and add them to the HIN.

3 Preliminaries and task description

In this section, we present the HEEL task and introduce the FDS construction based on DBLP. We first review the definition of HIN [28]:

Definition 1 (*HIN*) A HIN is a directed graph $G = (V, L)$ where node set V and edge set L are the collections of entities and relations, respectively. Each entity $v \in V$ belongs to one of a multiple number of entity types, and each edge $(v_i, v_j) \in L$ belongs to one of a multiple number of relation types.

In DBLP, a publication record contains the title of the paper (modeled as a collection of terms), the authors, the publication venue and the year, with other noisy attributes filtered. Denote M as the collection of ambiguous author names that have their respective disambiguation pages in DBLP. Let r_m be a publication record with an author name $m \in M$ which has not been linked to a specific author yet. R_m be the collection of all publication records with an "un-disambiguated" author name $m \in M$. We use all publication records that are not in $\bigcup_{m \in M} R_m$ to construct a HIN, called an FDS $G = (V, L)$. We can see that all author names in the FDS have been disambiguated.

Following [20], we extract five types of entities (i.e., authors, papers, venues, terms and years). Note that a term is a word in the paper title that is not a stopword. The star schema of the FDS of DBLP and the number of entities of each type is illustrated in Fig. 1. We assume all relations are symmetric. For example, there is a *writes* relation from an author to his/her paper, and a *writes*$^{-1}$ relation reversely. For each $m \in M$, denote E_m and E^*_m as the collections of referent authors in DBLP and new authors, respectively. The goal of HEEL is to learn a linking function $f : R_m \rightarrow E_m \cup E^*_m \cup \text{NIL}$:

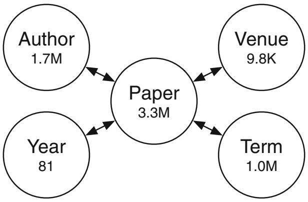


Fig. 1 Star schema of DBLP network. The numbers refer to the numbers of different entities in DBLP network

- $f(r_m) = e_m \in E_m$, if m in r_m refers to an existing author $e_m \in E_m$.
- $f(r_m) = e_m \in E_m^*$, if m in r_m refers to a new author e_m added into E_m^*.
- $f(r_m) =$NIL, if m in r_m does not refer to any authors in $E_m \cup E_m^*$.

We summarize the HEEL task as follows:

Definition 2 (*HEEL*) Given an FDS $G = (V, L)$ and a collection of publication records R_m with the same ambiguous author name $m \in M$, the goal is to learn the linking function f in order to discover the new author collection E_m^* and link each author name m with the publication record $r_m \in R_m$ to $f(r_m)$.

The reason that we introduce NIL is discussed below. If a new author is "discovered," it is added to E_m^* when the author names can not be linked to any existing authors. As more "un-disambiguated" records are discovered, this method can perform incremental linking by iteratively detecting new authors from either "un-disambiguated" records or these with referent authors marked as NIL.

Important notations are summarized in Table 1.

4 The partial classification EM approach

This section begins by introducing PC-EM. Next, we describe the CPP model, the partial classification method and the new author detection algorithm.

Table 1 Important notations

Notation	Description
$G = (V, L)$	An FDS constructed from DBLP
M	Collection of ambiguous author names
$m \in M$	An ambiguous author name in M
r_m	A publication record with an ambiguous author name m
$R_m = \{r_m\}$	Collection of publication records with author name m
E_m	Collection of existing referent authors for m
E_m^*	Collection of new referent authors for m
e_m	The referent author of m in r_m
$\Pr(e\|r_m)$	Prob. of m being linked to referent author e given r_m
$K = (\xi, C)$	An ECMP with meta-path ξ and constraints C
$\Pr(e\|r_m, K_i)$	Prob. of m being linked to e given r_m using the ECMP K_i only
$\mathbf{K} = \{K_i\}$	Collections of ECMPs used in this paper
$\mathbf{W}$	Weight vector of the CPP model
G_m	The HAG w.r.t. author name m

4.1 General algorithm of PC-EM

PC-EM is an iterative process consisting of E-step, PC-step and M-step after model initialization. The high-level procedure is illustrated in Algorithm 1. The process stops if (i) no new authors can be detected; or (ii) the link assignments and model parameters are stabilized.

Algorithm 1 General Framework of PC-EM

```
// Initialization
Initialize $E_m^* = \emptyset$;
Learn parameters $\mathbf{W}$ of CPP model based on FDS $G$;
while not converge do
    // E-step
    for each $r_m \in R_m$ do
        for each $e \in E_m \cup E_m^*$ do
            Predict $\Pr(e|r_m)$ based on CPP model;
        end for
    end for
    // PC-step
    for each $r_m \in R_m$ do
        $e_m = \operatorname{argmax}_{e \in E_m \cup E_m^*} \Pr(e|r_m)$;
        if Prediction $f(r_m) = e_m$ is confident then
            Link $m$ to $e_m$, remove $r_m$ from $R_m$ and add $r_m$ to $G$;
        end if
    end for
    Construct a HAG $G_m$ based on $R_m$;
    Detect a Maximum Edge Weight Clique $R_m^{'} \subseteq R_m$ from HAG $G_m$;
    if $|R_m^{'}| > 1$ then
        Add a new author $e_m^*$ to $E_m^*$;
        for each $r_m \in R_m^{'}$ do
            Link $m$ to $e_m^*$, remove $r_m$ from $R_m$ and add $r_m$ to $G$;
        end for
    end if
    // M-step
    Update parameters $\mathbf{W}$ of CPP model based on FDS $G$;
end while
// Post-processing
for each $r_m \in R_m$ do
    Link $m$ to NIL;
end for
```

Initialization In PC-EM, the constrained probability propagation (CPP) model is the major component that predicts referent authors given publication records with ambiguous author names. In the initial stage, we set $E_m^* = \emptyset$. Model parameters $\mathbf{W}$ are learned based on the FDS G via distant supervision.

E-step For each $r_m \in R_m$, we predict the referent author probability $\Pr(e|r_m)$ based on the CPP model where $e \in E_m \cup E_m^*$.

PC-step If the CPP model prediction is confident, $f(r_m) = \operatorname{argmax}_{e \in E_m \cup E_m^*} \Pr(e|r_m)$. Then, we remove r_m from R_m and add five types of entities and the corresponding relations in r_m to G. For the remaining publication records in R_m, because no existing authors in $E_m \cup E_m^*$ are suitable to be the referent author, it is likely that a new author is discovered. We try to find a collection of publication records $R_m^{'} \subseteq R_m$ such that all the author names m in $R_m^{'}$ have a large probability to refer to the same new author. The publication records $R_m^{'}$ are

detected by solving the maximum edge weight clique (MEWC) problem over a graph named homogeneous affinity graph (HAG) G_m constructed from R_m. The new author e^*_m is added to E^*_m. For each $r_m \in R^{'}_m$, let $f(r_m) = e^*_m$. Next, we remove $R^{'}_m$ from R_m and add entities and relations in $R^{'}_m$ to the FDS G. In this way, we perform partial classification by linking author names in part of R_m to authors in $E_m \cup E^*_m$.
M-step After new authors are added in E^*_m and linking assignments are changed, the model parameters $\mathbf{W}$ are updated based on the enlarged FDS G.
Post-processing When this iterative process stops, we link each author name m in $r_m \in R_m$ to NIL, meaning no referent authors are detected.

This PC-EM process converges after a limited number of iterations, with the reasons stated as follows. In each iteration, the model tries to add a new author to the system and updates model parameters and linking assignments. Assume the kth iteration is the first iteration that the algorithm cannot detect a new author. The linking assignments are calculated in the E-step of the kth iteration. Because the number of authors does not change, the CPP model is the same as the one trained in the $(k-1)$th iteration and thus does not need to be re-trained again. Therefore, the PC-EM process converges after a finite number of iterations.

4.2 Constrained probability propagation model

The CPP model is a random walk based model that generates conditional probabilities $\Pr(e|r_m)$ $(e \in E_m \cup E^*_m)$ given a publication record r_m with author name $m \in M$. It is used in the E-step for probability prediction and trained in the M-step. We describe the model in detail and also illustrate how parameters are learned via constrained gradient ascent.

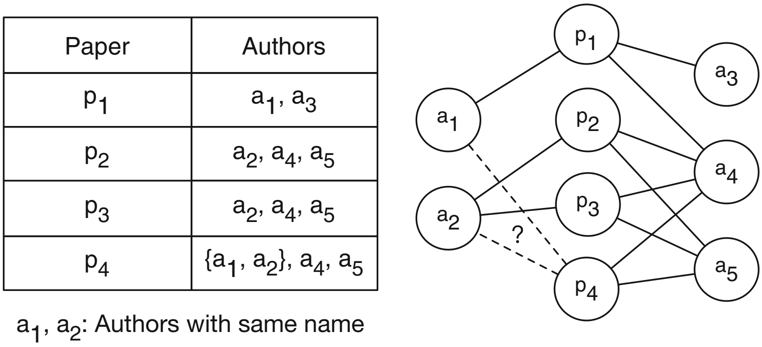
4.2.1 Model description

The distributions of random walkers over meta-paths can generate a probability distribution based on the link structure of the HIN. Based on the previous research [13,27], we present the definition of meta-path:

Definition 3 (*Meta-Path*) A meta-path of length n is a path in the form of $\Phi_1 - \Phi_2 - \cdots - \Phi_{n+1}$, where each Φ_i is an entity type.

Consider the toy example in Fig. 2. For simplicity, we only consider two entity types: authors A and papers P here. Given three papers p_1, p_2 and p_3, together with their authors, we wish to predict which one of the two authors with the same name (a_1 and a_2) writes p_4 on condition that a_4 and a_5 are p_4's authors. From the network structure, we can see that a_2 co-authors with a_4 and a_5 frequently. Thus, it is highly possible that p_4 should link to a_2 rather than a_1. Here, the meta-path "P–A–P–A" expresses the relation between a paper and the collaborator of the authors of the paper.[1] However, it is not straightforward to calculate the author distribution by meta-path constrained random walk in previous study (see [13]). For example, random walkers may go from p_4 to a_3 (e.g., $p_4 - a_4 - p_1 - a_3$), but this path is not useful for author prediction.

[1] In meta-path description, we use P, A, T, V and Y to represent any nodes (i.e., entities) in the FDS with the type of paper, author, term, venue and year, respectively.



Paper	Authors
p1	a1, a3
p2	a2, a4, a5
p3	a2, a4, a5
p4	{a1, a2}, a4, a5

a1, a2: Authors with same name

Fig. 2 Example of author prediction. An author of paper p_4 can refer to either a_1 or a_2, which share the same name

We further propose the concept of *entity-constrained meta-path* (**ECMP**) as follows:

Definition 4 (*Entity-Constrained Meta-Path*) An ECMP is a pair $K = (\xi, C)$ where $\xi = \Phi_1 - \cdots - \Phi_{n+1}$ is a meta-path of length n and C is a set of constraints on Φ_i $(i = 1, \ldots, n+1)$.

For example, in Fig. 2, the relation between paper p_4 and authors a_1 and a_2 is modeled by the ECMP $\{p_4\} - A - P - \{a_1, a_2\}$, which is a "P–A–P–A" meta-path that has two constraints: (i) $\Phi_1 = \{p_4\}$ and (ii) $\Phi_4 = \{a_1, a_2\}$.

For publication record r_m, the ECMPs have the following characteristics:

1. They start with the paper node p w.r.t. r_m and end in an author node in $E_m \cup E_m^*$.
2. They have length $n > 1$ (because a length-one path from p to $E_m \cup E_m^*$ is not much meaningful for author prediction).

Based on the star schema in Fig. 1, we use four ECMPs in our approach: (i) $\{p\} - A - P - E_m \cup E_m^*$, (ii) $\{p\} - T - P - E_m \cup E_m^*$, (iii) $\{p\} - V - P - E_m \cup E_m^*$ and (iv) $\{p\} - Y - P - E_m \cup E_m^*$. Long paths can be also applied in this task, but as shown in [27], these paths may not carry rich semantic meanings. Given a collection of ECMPs $\mathbf{K}$, the CPP model $\Pr(e|r_m)$ is defined as a linear combination of probabilities:

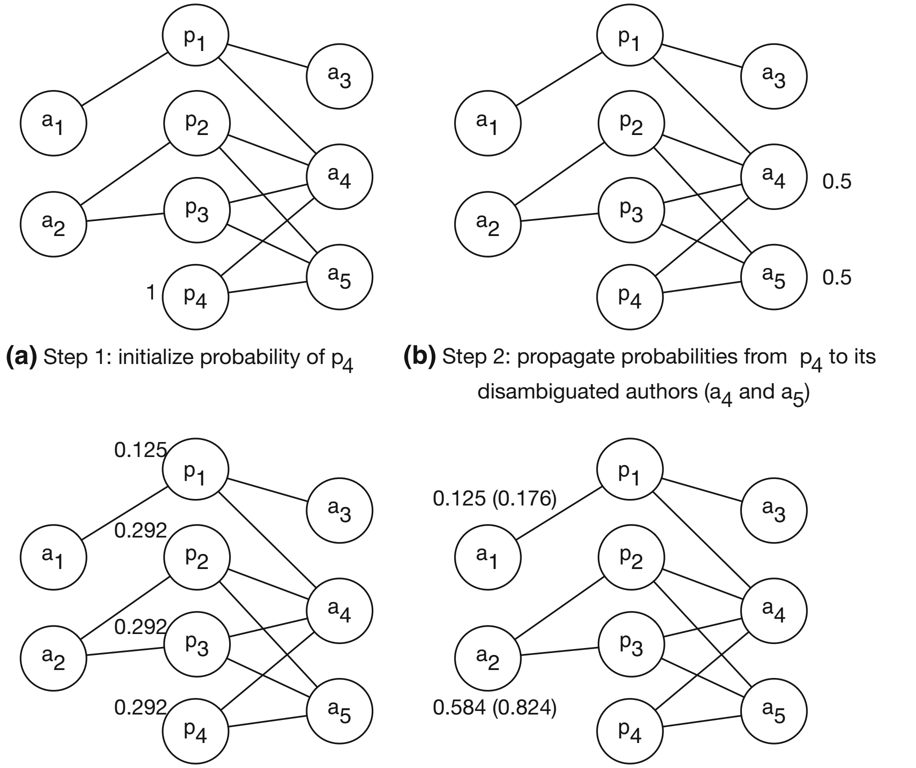
$$\Pr(e|r_m) = \sum_{K_i \in \mathbf{K}} w_i \Pr(e|r_m, K_i) \tag{1}$$

where $\sum_{K_i \in \mathbf{K}} w_i = 1$. $\Pr(e|r_m, K_i)$ is the probability of author name m being linked to referent author e in r_m along ECMP K_i.

To approximate $\Pr(e|r_m, K_i)$, we compute $\Pr(p \rightarrow e|K_i)$ in the FDS, which is the random walk probability from the paper node p w.r.t. r_m to author e via ECMP K_i. Inspired by Lao and Cohen [13], we define the general version of the ECMP-constrained random walk process as follows. Let u, v and $u^{'}$ be arbitrary nodes in G. For an empty ECMP K_i (i.e., length $n_i = 0$), we set $\Pr(v \rightarrow u|K_i) = 1$ if $u = v$ and constraints C_i are satisfied; otherwise $\Pr(v \rightarrow u|K_i) = 0$.

For (i) a non-empty ECMP K_i with $\xi_i = \Phi_1 - \Phi_2 - \cdots - \Phi_{n+1}$ and constraints C_i and (ii) a shorter ECMP $K_i^{'}$ with $\xi_i^{'} = \Phi_1 - \Phi_2 - \cdots - \Phi_n$ and constraints C_i, $\Pr(v \rightarrow u|K_i)$ is defined recursively as:

$$\Pr(v \rightarrow u|K_i) = \sum_{u^{'} \in \Phi_n} \Pr(v \rightarrow u^{'}|, K_i^{'}) \frac{I(u, u^{'}, \Phi_{n+1}, C_i)}{N(u^{'}, \Phi_{n+1}, C_i)} \tag{2}$$



(a) Step 1: initialize probability of p_4

(b) Step 2: propagate probabilities from p_4 to its disambiguated authors (a_4 and a_5)

(c) Step 3: propagate probabilities from authors to their papers

(d) Step 4: propagate probabilities from papers to a_1 and a_2 only, then re-normalize probabilities

Fig. 3 Example of the ECMP-constrained random walk process. We consider the ECMP $\{p_4\}-A-P-\{a_1, a_2\}$ only

where $I(u, u^{'}, \Phi_{n+1}, C_i) = 1$ if u is connected to $u^{'}$ of type Φ_{n+1} in G; otherwise, we set $I(u, u^{'}, \Phi_{n+1}, C_i) = 0$. $N(u^{'}, \Phi_{n+1}, C_i)$ is the number of entities of type Φ_{n+1} that are connected with $u^{'}$ and satisfy constraints C_i.

Finally, the probability $\Pr(p \rightarrow e|K_i)$ is re-normalized to generate $\Pr(e|r_m, K_i)$:

$$\Pr(e|r_m, K_i) = \frac{\Pr(p \rightarrow e|K_i)}{\sum_{e^{'} \in E_m \cup E_m^*} \Pr(p \rightarrow e^{'}|K_i)} \tag{3}$$

Readers can refer to Fig. 3 for the probability propagation process of the toy example in Fig. 2. We can see that the ECMP-constrained random walk process employed in the CPP model gives different weights to candidate authors based on the link structure of the HIN. We do not use any prior knowledge (similar to [20]) in the model because the distributions of "un-disambiguated" authors are not proportional to the prestige levels of actual authors. In fact, in the DBLP dataset, the pages of famous authors are usually well maintained. So papers with "un-disambiguated" authors may actually link to junior researchers, rather than famous authors.

4.2.2 Model training

In the training phase, we randomly sample a collection of referent author–publication record pairs $D = \{(e_m, r_m)\}$ from the FDS G. For each pair $(e_m, r_m) \in D$, we remove the author node e_m temporarily from G and calculate $\Pr(e_m|r_m, K_i)$ for each $K_i \in \mathbf{K}$. The optimization objective is:

$$\max \ J(\mathbf{W}) = \sum_{(e_m,r_m)\in D} \ln \sum_{K_i\in\mathbf{K}} w_i \Pr(e_m|r_m, K_i) \text{ s.t. } \sum_{K_i\in\mathbf{K}} w_i = 1 \tag{4}$$

A constrained gradient ascent algorithm is employed to solve the optimization problem. In the tth iteration, each weight $w_i^{(t)}$ is updated as:

$$\frac{\partial J(\mathbf{W})}{\partial w_i} = \sum_{(e_m,r_m)\in D} \frac{\Pr(e_m|r_m, K_i)}{\sum_{K_j\in\mathbf{K}} w_j \Pr(e_m|r_m, K_j)} \tag{5}$$

$$w_i^{(t+1)} = w_i^{(t)} + \eta \cdot \frac{\partial J(\mathbf{W})}{\partial w_i}|_{w_i=w_i^{(t)}} \tag{6}$$

After updating the value of $\mathbf{W}$, the weights are re-normalized to satisfy the constraint $\sum_{K_i\in\mathbf{K}} w_i = 1$. This process iterates until the weight vector $\mathbf{W}$ converges. Therefore, our model is distantly supervised and does not require human-labeled training data.

4.2.3 Model prediction

In the prediction phase, given a publication record r_m with an ambiguous author name m, we extract all five types of entities except the ambiguous author m from r_m and insert the entities and corresponding relations into G temporarily. The CPP model calculates the probability $\Pr(e|r_m)$ using random walk probabilities and weights $\mathbf{W}$.

4.3 Partial classification

The partial classification technique is used in the PC-step to determine whether the prediction of the CPP model is confident. For each $r_m \in R_m$, it links the author name m to its referent entity in $E_m \cup E_m^*$.

Given the publication record r_m, the initial prediction of the CPP model is: $f(r_m) = \text{argmax}_{e\in E_m\cup E_m^*} \Pr(e|r_m)$. Because the PC-EM approach is self-supervised, if the prediction is incorrect, the error will propagate in the next iteration. Dalvi et al. [8] propose two criteria: Jensen–Shannon Divergence (JSD) and Max–Min, to indicate that the prediction of a classifier is not confident enough to make the prediction. This implies new, unknown classes may exist. For our task, the JSD criterion is implemented as:

$$\text{JSD}(\mathbf{Pr}(E_m \cup E_m^*|r_m)\|\mathbf{u}) > \frac{1}{|E_m \cup E_m^*|} \tag{7}$$

where $\mathbf{Pr}(E_m \cup E_m^*|r_m)$ is the $|E_m \cup E_m^*|$-dimensional vector where each element is $\Pr(e|r_m)$ $(e \in E_m \cup E_m^*)$. $\mathbf{u}$ is the $|E_m \cup E_m^*|$-dimensional uniform distribution vector: $\mathbf{u} = (\frac{1}{|E_m\cup E_m^*|}, \frac{1}{|E_m\cup E_m^*|}, \ldots, \frac{1}{|E_m\cup E_m^*|})$. The Max–Min criterion is:

$$\frac{\max_{e\in E_m\cup E_m^*} \Pr(e|r_m)}{\min_{e\in E_m\cup E_m^*} \Pr(e|r_m)} > 2 \tag{8}$$

For the author linking task, if one or two criteria hold, it is confident to predict that the author name m in the publication record r_m refers to the author e_m. However, the experiments show that either criterion is not effective. Please refer to Sect. 5 for detailed explanation.

We design two new criteria: (i) threshold-based cut and (ii) Max-Second Max constraint. In threshold-based cut, the model prediction $f(r_m) = \max_{e \in E_m \cup E_m^*} \Pr(e|r_m)$ is confident if

$$\max_{e \in E_m \cup E_m^*} \Pr(e|r_m) > \tau_1 \tag{9}$$

where $\tau_1 \in (0, 1)$.

The Max-Second Max constraint is a variant of the Max–Min criterion. We observe that no matter if the prediction is confident or not, $\min_{e \in E_m \cup E_m^*} \Pr(e|r_m)$ tends to be very small (e.g., 10^{-4}). Thus, the max–min probability ratio is very large even if $\max_{e \in E_m \cup E_m^*} \Pr(e|r_m)$ is small. Denote $\text{secmax}_{e \in E_m \cup E_m^*} \Pr(e|r_m)$ as the second largest probability. The experiments show that it is effective to determine the confidence level of model prediction by the Max-Second Max constraint:

$$\frac{\max_{e \in E_m \cup E_m^*} \Pr(e|r_m)}{\text{secmax}_{e \in E_m \cup E_m^*} \Pr(e|r_m)} > \tau_2 \tag{10}$$

where $\tau_2 > 1$. In this paper, the prediction $f(r_m) = e_m$ is confident if both threshold-based cut and Max-Second Max constraint hold.

For each publication record $r_m \in R_m$, if author name m can be linked to the referent author in $E_m \cup E_m^*$ based on Eqs. (9) and (10) or the new author e_m^* (see Sect. 4.4), we remove r_m from R_m and add five types of entities and the corresponding relations in r_m to the FDS G. In this way, the PC-EM approach supports self-supervised learning where model parameters and linking assignments are iteratively updated.

The improvement of PC-EM compared to the traditional classification EM-based approach [5] is twofold: (i) we link an author name m to existing referent authors in $E_m \cup E_m^*$ only if the model has high confidence; or (ii) we discover a new author e_m^* in each iteration to support exploratory linking.

4.4 New author detection

For publication records R_m with author name m that cannot be linked to any known authors in $E_m \cup E_m^*$, we aim at finding a subset of R_m (i.e., $R_m^{'} \subseteq R_m$) such that all the author names m in $R_m^{'}$ have a large probability to refer to the same new author, denoted as e_m^*. Here, we model the new author detection problem as MEWC over a graph model HAG.

A HAG is an undirected graph $G_m = (R_m, L_m, W_m)$, where R_m is the node set in G_m, L_m is the edge set and W_m gives weights to each edge. Let $Au(r_m)$ be the collection of co-authors of m in the publication record r_m and $Te(r_m)$ be the collection of terms in the paper title of r_m. $\forall r_m, r_m^{'} \in R_m$, the edge $(r_m, r_m^{'}) \in L_m$ if $Au(r_m) \cap Au(r_m^{'}) \neq \emptyset$. The weight $w(r_m, r_m^{'})$ of the edge $(r_m, r_m^{'})$ is defined as:

$$w(r_m, r_m^{'}) = \alpha \frac{|Au(r_m) \cap Au(r_m^{'})|}{|Au(r_m) \cup Au(r_m^{'})|} + (1 - \alpha) \frac{|Te(r_m) \cap Te(r_m^{'})|}{|Te(r_m) \cup Te(r_m^{'})|} \tag{11}$$

where $\alpha \in (0, 1)$ is a tuning parameter.

A part of author names m in R_m are likely to refer to the same underlying author if the paper terms and co-authors are most similar. Therefore, we are looking for a clique in the HAG with maximum sum of edge weights instead of maximum size. Consider the example

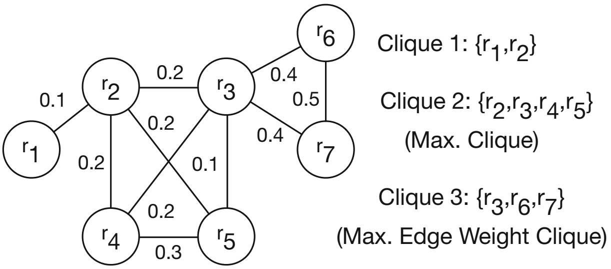


Fig. 4 A simple graph with edge weights

in Fig. 4. In the graph, among three maximal cliques (i.e., Cliques 1–3), Clique 2 is the maximum clique, but Clique 3 is the maximum edge weight clique that we wish to detect.

Hence, the goal of MEWC is to find a subgraph $G_m^{'} = (R_m^{'}, L_m^{'})$ from G_m such that $R_m^{'}$ is a clique with maximum sum of edge weights. We define the objective function as follows:

$$\max \sum_{(r_m, r_m^{'}) \in L_m^{'}} w(r_m, r_m^{'}) \text{ s.t. } L_m^{'} \subseteq L_m \quad \text{and} \quad \forall r_m, \forall r_m^{'} \in R_m^{'}, (r_m, r_m^{'}) \in L_m^{'} \tag{12}$$

A number of algorithms have been proposed to solve MEWC, such as unconstrained quadratic programming, branch-and-cut algorithm. [1]. However, they aim at computing exact results which suffer from high complexity due to the NP-hardness of MEWC [1]. Here, we present a Monte Carlo-based algorithm to solve MEWC approximately, as shown in Algorithm 2.

In each iteration, it selects an edge $(r_m, r_m^{'})$ from L_m with probability proportional to its weight $w(r_m, r_m^{'})$. After adding $(r_m, r_m^{'})$ to $L_m^{'}$, it removes $(r_m, r_m^{'})$ and edges that does not connect with any nodes in $R_m^{'}$ from L_m. It repeats until no more edges in L_m can be added to $L_m^{'}$. After that, a clique $R_m^{'}$ is selected. Because Algorithm 2 can only produce approximate results, we run it k times and select the clique with largest edge weights as output. Thus, the new author named m is found with a list of publications $R_m^{'}$. The worst-case time complexity of the algorithm is $O(|L_m|^2)$. The entire time complexity of this step is $O(k|L_m|^2)$. Therefore, we approximately solve the NP-hard problem in quadratic time.

Note that our new author detection approach imposes strong constraints on two publication records that are assigned to the same author. Because with no labeled data in this step, our approach requires very high precision in a completely unsupervised learning process. After this step, the basic characteristics of the new author can be learned in the M-step, given a handful of typical publication records that we obtain here as "seeds." More linking assignments to the new author are done in the next E-step.

5 Experiments

In this section, we conduct experiments on multiple datasets to evaluate the effectiveness of HEEL. We report the performance and compare it with state-of-the-art methods. Specifically, we aim at answering the following three research questions:

RQ1 Is HEEL effective to link ambiguous names to a HIN?

Algorithm 2 Maximum Edge Weight Clique Detection Algorithm

```
Input: HAG $G_m = (R_m, L_m, W_m)$.
Output: Clique $R_m^{'}$.
Initialize $G_m^{'} = (R_m^{'}, L_m^{'})$ with $R_m^{'} = \emptyset$ and $L_m^{'} = \emptyset$;
while $L_m \neq \emptyset$ do
    Sample $(r_m, r_m^{'})$ from $L_m$ with prob. $\propto w(r_m, r_m^{'})$;
    $R_m = R_m \setminus \{r_m, r_m^{'}\}$, $R_m^{'} = R_m^{'} \cup \{r_m, r_m^{'}\}$;
    $L_m = L_m \setminus \{(r_m, r_m^{'})\}$, $L_m^{'} = L_m^{'} \cup \{(r_m, r_m^{'})\}$;
    for each $(r_m, r_m^{'}) \in L_m$ do
        if $r_m \notin R_m^{'}$ and $r_m^{'} \notin R_m^{'}$ then
            $R_m = R_m \setminus \{r_m, r_m^{'}\}$, $L_m = L_m \setminus \{(r_m, r_m^{'})\}$;
        end if
    end for
end while
return Clique $R_m^{'}$;
```

RQ2 Is HEEL effective to disambiguate a collection of author names with the corresponding publication records?

RQ3 Can HEEL discover new authors and turn the partially disambiguated HIN to a fully disambiguated one?

As seen, RQ1, RQ2 and RQ3 correspond to the three tasks: author name linking with HINs, author name disambiguation and new author discovery.

5.1 Task 1: author name linking with HINs

5.1.1 Experimental data and settings

For EL, the only prior work that considers EL with HINs is [20]. They aim to link author names in plain texts to DBLP and their test set is not suitable for evaluating our task. In this paper, we use two publicly available benchmark datasets for author name disambiguation as our test sets. The first one is a classical dataset and is the same as that used in [16,33] and many others, which is a subset of DBLP. However, the size of this dataset is relatively small (588 records). We also use another dataset is created by Li et al. [16], which is larger in size and has bigger ambiguity (2050 records). The statistics of test sets are summarized in Table 2. To evaluate the performance of EL, we ask human annotators to link each cluster of authors in both datasets to an actual author in DBLP or NIL if not present. Consequently, all the target author names of these publication records are annotated with either their referent authors or NIL. In the experiments, we randomly sample 30% of the records from Dataset 1 as the development set to tune the parameters and compare our method against others over the rest of the test sets.

We download the June 2016 version DBLP data dump[2] as the knowledge source for the HIN. To avoid overfitting, we extract all the disambiguated publication records that are not in the two test sets to construct the FDS. The FDS contains 3.3M papers, 1.7M authors, 1.0M terms, 9.8K venues and 81 years. The terms are filtered by a stopword dictionary and processed by a Porter stemmer.[3] We follow the HIN construction method introduced in [20]

[2] http://dblp.dagstuhl.de/xml/release/.

[3] https://tartarus.org/martin/PorterStemmer/.

Table 2 Test dataset summarization

Author name	Dataset 1 (from [33])		Dataset 2 (from [16])	
	#Records	#Authors	#Records	#Authors
Hui Fang	9	3	45	8
Ajay Gupta	16	4	25	8
Joseph Hellerstein	151	2	234	2
Rakesh Kumar	36	2	104	8
Michael Wagner	29	5	61	16
Bing Liu	89	6	192	23
Jim Smith	19	3	54	5
Lei Wang	55	13	400	144
Wei Wang	140	14	833	216
Bin Yu	44	5	102	18
Total	588	−−	2050	−−

and the schema in Fig. 1 to create the FDS. For parameter learning, we randomly sample 2K publication records with ambiguous author names as automatically generated training data. The hyper-parameter settings are $\eta = 0.001$ and $\alpha = 0.5$, fine-tuned based on human inspection.

5.1.2 Evaluation metrics

Following previous EL research [20,22,23], we employ *Accuracy* as the evaluation metric. In this paper, because we pay special attention to the NIL linking issue, we report three linking accuracy values. The first two are linkable and unlinkable accuracy, calculated as:

$$\text{Linkable Accu. (LA)} = \frac{\text{\#Correctly linked author names}}{\text{\#Total linkable author names}} \tag{13}$$

$$\text{Unlinkable Accu. (UA)} = \frac{\text{\#Author names correctly predicted as NIL}}{\text{\#Total unlinkable author names}} \tag{14}$$

The third metric is overall accuracy (OA), which is the weighted average of linkable and unlinkable accuracy where the weight is proportional to the number of linkable/unlinkable author names.

5.1.3 Performance of PC-EM

In PC-EM, we first tune the thresholds τ_1 and τ_2 in the PC-step. Empirically, we only require τ_1 to be a very small number to achieve a high performance due to the large number of possible referent authors. We fix τ_2 to be different values and tune τ_1. In Fig. 5, we present the overall linking accuracy under the different values of τ_1 from 0.05 to 0.25 with $\tau_2 = 1.5$. When the value of τ_1 is small, the model tends to predict the most probable author even if the confidence level is low. When τ_1 is large, the model is likely to give the NIL prediction in EL. Therefore, the different choice of τ_1 reflects the trade-off between linkable accuracy and unlinkable accuracy. In Fig. 5, we can see a clear trend from rise to decline. The peak arrives when $\tau_1 = 0.1$ with the highest overall linking accuracy of 84.6%.

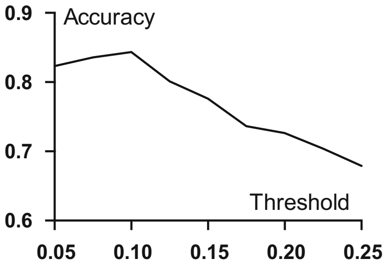


Fig. 5 Parameter analysis of τ_1

Table 3 Comparison of author linking results using different criteria in the PC-step over the development set (%)

Criterion	OA	LA	UA
JSD [8]	66.9	71.2	38.8
Max–Min [8]	77.8	**89.0**	5.5
Thres.	83.8	84.5	79.5
Thres. & Max-SecMax	**84.6**	85.2	**79.6**

Bold values indicate the highest scores among all methods

We further investigate the overall linking performance using different criteria in the PC-step over the development set, as shown in Table 3. The JSD and Max–Min criteria proposed in [8] do not achieve satisfactory results based on our experimental results. The possible causes are twofold: (i) The dimensionality of the author distribution vector varies from author to author, and thus, it is not easy to determine the confidence level only based on JSD; and (ii) the probabilities of authors being linked to some author names are extremely small and the Max–Min ratio is not strongly related to the prediction confidence. The reason why the unlinkable accuracy with Max–Min criterion is extremely low is that the Max–Min ratio is very high in most cases whether the prediction is confident or not. Compared with previous methods, our threshold-based criterion considers the characteristics of author distributions in the HEEL task and has high linking performance, with overall linking accuracy of 83.8%. We combine the threshold-based criterion with the Max-SecMax criterion to improve the accuracy by 0.8%.

5.1.4 Comparison with baselines

To the best of our knowledge, SHINE [20] is the only work and the state-of-the-art approach that addresses EL with HINs. In their work, they also introduce two simple methods (i.e., POP and VSim) as baselines. In this part, we present the comparative study on these methods and the proposed approach (i.e., PC-EM). POP is the entity popularity model in [20] based on a PageRank-like algorithm over the HIN. VSim is the vector space model which selects the referent author with highest VSM similarity. SHINE [20] links mentions to entities in HINs based on meta-path-constrained random walk process.

We observe that existing baselines cannot deal with the NIL linking problem.[4] To address the NIL issue, we add a threshold-based filter as a post-processing step in POP, VSim and SHINE. If the prediction score is below a threshold (tuned on the development set), we set the prediction result to NIL. The results are illustrated in Table 4. From the experiments, we

[4] To our knowledge, there exist some other EL methods that consider the NIL issue such as [23]. But their task is to link mentions in the plain texts to entities in the knowledge bases and it is not easy to modify them for EL with HINs.

Table 4 Comparison of author linking results of different approaches over two test sets (%)

Method	Dataset 1			Dataset 2		
	OA	LA	UA	OA	LA	UA
POP	42.9	42.9	42.4	28.6	26.4	34.1
VSim	79.4	80.5	72.2	75.6	76.1	74.6
SHINE [20]	81.1	**85.6**	51.8	76.7	81.0	67.5
PC-EM	**83.1**	83.6	**79.3**	**79.5**	**81.6**	**74.8**

Bold values indicate the highest scores among all methods

can see that the baseline POP has very low linking accuracy. This is because unlike [20], a lot of publication records that require to be linked involve junior researchers rather than popular ones. The experimental results agree with our observations well. VSim considers the contextual relatedness between an author name and all possible referent authors. However, these methods perform shallow matching only, without modeling the semantics between different types of entities appeared in the contexts. SHINE is the most competitive method, with overall linking accuracies of 81.1% and 76.7%. PC-EM improves the overall accuracy by 2% and 2.8% over two datasets, respectively.

5.2 Task 2: author name disambiguation

5.2.1 How our method works

PC-EM is basically a linking model and does not generate author clusters directly. To predict whether two authors named m in two publication records r_m and $r_m^{'}$ refer to the same underlying author, we define $f_{ND}(r_m, r_m^{'})$:

$$f_{ND}(r_m, r_m^{'}) = \begin{cases} 1 & f(r_m) = f(r_m^{'}),\ f(r_m) \in E_m \cup E_m^{*} \\ -1 & \text{Otherwise} \end{cases} \tag{15}$$

$f_{ND}(r_m, r_m^{'}) = 1$ means the author name m appeared in r_m and $r_m^{'}$ refers to the same author; otherwise, $f_{ND}(r_m, r_m^{'}) = -1$. Note that if $f(r_m) = \text{NIL}$ and $f(r_m^{'}) = \text{NIL}$, our model still outputs negative. This is because if the target authors in r_m and $r_m^{'}$ were in fact the same, r_m and $r_m^{'}$ would have been clustered into a clique in the PC-step in a large probability.

5.2.2 Evaluation results

We present the results under the evaluation of author name disambiguation. The evaluation metrics that we employ are pairwise *precision*, *recall* and *F-measure*. We use the same settings for PC-EM as those in the previous experiments.

For comparative study, we obtain the source codes, original data and experimental results over both datasets from [16], which provides a benchmark for evaluation. In total, we have five baselines, i.e., Jac, DISTINCT [33], Arnetminer [31], CSLR [16] and BNCE [34]. Jac is the simple method that uses the Jaccard similarity between two publication records to determine whether the two author names refer to the same person. DISTINCT, Arnetminer and CSLR are treated as baselines for author name disambiguation due to the convincing results and high citation counts. For details, please refer to [16]. However, they do not consider the

Table 5 Result comparison for author name disambiguation (%)

Method	Dataset 1			Dataset 2		
	Pre	Rec	F-1	Pre	Rec	F-1
Jac	88.2	82.8	84.1	78.0	65.6	66.5
DISTINCT [33]	76.9	**90.8**	80.2	68.3	**87.4**	73.5
Arnetminer [31]	81.5	88.4	80.2	63.2	69.7	60.2
CSLR [16]	**95.0**	80.5	86.3	**92.9**	69.2	77.8
BNCE [34]	85.3	84.5	84.9	79.1	75.8	77.4
PC-EM	88.6	86.3	**87.4**	82.6	79.2	**80.9**

Bold values indicate the highest scores among all methods

Table 6 Detailed disambiguation results for 10 names

Name	Pre	Rec	F-1	Name	Pre	Rec	F-1
Hui Fang	1.00	1.00	1.00	Bing Liu	0.92	0.86	0.89
Ajay Gupta	0.93	0.93	0.93	Jim Smith	0.93	0.86	0.89
Joseph Hellerstein	0.78	0.87	0.83	Lei Wang	0.83	0.71	0.77
Rakesh Kumar	1.00	1.00	1.00	Wei Wang	0.81	0.70	0.75
Michael Wagner	0.81	0.80	0.80	Bin Yu	0.85	0.90	0.87

emergence of new authors. To our knowledge, BNCE [34] is the most recent and the state-of-the-art method to address this issue by Bayesian non-exhaustive classification. We implement this method by taking the disambiguated records for each author name as training records. Table 5 illustrates the experimental results of different methods. Overall, our method achieves an F-measure of 87.4% and 80.9%, respectively. We can see that the proposed approach PC-EM outperforms the most competitive method CSLR by 1.1% and 3.1%. Our method also outperforms [34] mostly because the PC-EM process generates new authors effectively.

In Table 6, we present the detailed disambiguation results for 10 author names over Dataset 1. It shows that the overall results are generally satisfactory, even with 100% accuracy for a few author names (e.g., Hui Fang, Rakesh Kumar). However, we have to admit that the performance is not sufficiently high for a few author names such as Lei Wang and Wei Wang. The most cause is that a lot of Chinese names have the same spelling in English alphabet, causing the large number of referent authors [16]. We also notice that some of the authors only have very few (< 3) papers in DBLP. The characteristics of these authors are not well captured.

5.3 Task 3: new author discovery

To the best of our knowledge, there are few studies that focus on discovering new authors outside existing bibliographic systems. Without standard evaluation frameworks available, we conduct an experiments to compare the number of new authors generated by our approach and the ground truth. Based on human annotation, only around 10% of the records in Dataset 1 cannot be linked to existing authors and thus this dataset is not suitable for Task 3. Following the experimental settings in [33], we remove ambiguous authors who have only one paper

Table 7 Comparison of #new authors generated by PC-EM and the ground truth

Name	PC-EM	Truth	Name	PC-EM	Truth
Hui Fang	2	1	Bing Liu	2	1
Ajay Gupta	6	8	Lei Wang	28	36
Rakesh Kumar	3	4	Wei Wang	32	51
Michael Wagner	3	3	Bin Yu	5	8

Table 8 Examples of new authors named *Wei Wang* outside DBLP, each annotated with the affiliation and one paper

Wei Wang 0085	Institute of Microelectronics, Peking University
A Novel 3D Flexible Parylene-Metal Structure Fabrication Technique	
Wei Wang 0086	School of Computer Science and Technology, HUST
GPU-based Multifrontal Optimizing Method in Sparse Cholesky Factorization	
Wei Wang 0087	School of Computer Science and Engineering, Southeast University
Stochastic modeling of dynamic right-sizing for energy-efficiency in cloud data centers	
Wei Wang 0088	College of Educational Science, Nanjing Normal University
XAR-Miner: Efficient Association Rules Mining for XML Data	

from Dataset 2 and calculate the number of new authors. We report the results for eight author names, illustrated in Table 7.[5]

We can see that the estimated numbers are close to the actual numbers for most of the cases, which means the method is capable of detecting new authors automatically. However, this task is far from being completely solved. This is because unlike very famous researchers, new, undiscovered authors are usually junior researchers or students with very few papers. Based on our experience, it is even challenging for human experts to find out the profiles of these authors. Thus, the performance of our method is likely to drop. In summary, the proposed approach provides a relatively effective solution, while this task is still an open challenge for the research community.

Additionally, we present a case study of new authors detected outside DBLP. Take the most ambiguous name *Wei Wang* as an example. In DBLP, 84 authors named *Wei Wang* are represented in the format of “Name+ID” (from “Wei Wang 0001” to “Wei Wang 0084”). Due to space limitation, we only list five new authors named *Wei Wang* in Table 8 based on the new clusters generated by PC-EM. We present the affiliation information and the title of one paper for each author. Thus, our approach has the potential to bring richer semantics to DBLP, distinguishing similar authors more clearly.

5.4 Application

Based on our work, we implement an application for exploratory author name linking for DBLP. The general framework is presented in Fig. 6. The system has two modes: (i) default mode and (ii) on-the-fly mode. The default mode works in the way introduced in this paper, using our trained model and the FDS as the underlying HIN. The on-the-fly mode can be viewed as a light-weight version of our method. It simply takes all the publication records

[5] There are no unlinkable records for the remaining two author names.

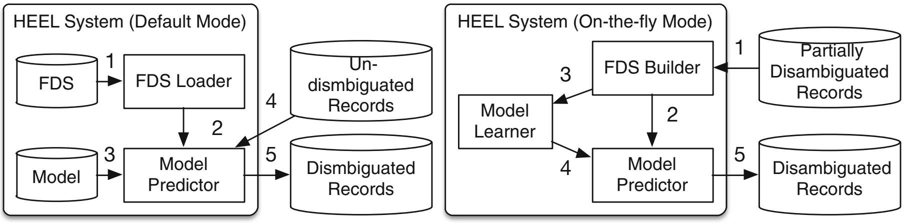


Fig. 6 Application framework of the HEEL system

Table 9 Cases of linking errors in the DBLP network with original and corrected author IDs

Original author	Corrected author
Zhaoyang Zhang 0001	Zhaoyang Zhang 0003
Paper: Performance Analysis of Media Cloud-Based Multimedia Systems With Retrying Fault-Tolerance Technique	
Hao Chen 0002	Hao Chen 0011
Paper: Automatic Detection of Cerebral Microbleeds From MR Images via 3D Convolutional Neural Networks	
Nan Tang 0002	Nan Tang 0004
Paper: A Novel Algorithm for Detecting Air Holes in Steel Pipe Welding Based on Hopfield Neural Network	
Wei Wang 0036	Wei Wang 0060
Paper: Cooperative fuzzy adaptive output feedback control for synchronisation of nonlinear multi-agent systems under directed graphs	

with respective to an ambiguous author name as input and constructs the FDS using part of the records in which the names have been disambiguated. The system trains the model automatically based on the FDS and makes prediction over the rest of the records.

5.5 Other issues to be considered

In the previous research, we assume the linking structure of the FDS derived from DBLP is correct. In this preliminary study, we try to detect the errors in the FDS. Each time, we take one disambiguated record out of the FDS as input, mask the author ID and link it to the rest part of the FDS. If the linking result and the original author ID are different, it is possible that the linking result in DBLP is incorrect. In Table 9, we present a couple of cases of linking errors and the corrected author IDs predicted by our approach. Due to space limitation, we only list the authors that are incorrectly linked in DBLP for each record. As seen, our method has the potential to improve linking quality of HINs by correcting errors. Note that the development of a complete algorithm for automatic error discovery for HINs can be more complicated and is beyond the scope of this paper. It is also interesting to estimate the error rates of large-scale HINs that are frequently used in the research community and real-world applications in the future.

6 Conclusion and future work

In this paper, we propose the task of HEEL to address the name ambiguity issue for HINs. A PC-EM-based framework is introduced to address this task without human intervention. We propose a CPP model to predict the distribution of referent authors given an ambiguous author name in a publication record. To tackle the challenge of "new author" problem, we present a partial classification technique and a MEWC algorithm. Experiments show that our method outperforms previous methods. Currently, our work only focuses on name ambiguity in HINs. Future work includes: (i) improving the performance of new author discovery and the linking quality of HINs, (ii) designing exploratory linking algorithms for surface names in other data sources to HINs and (iii) studying how our PC-EM approach can be used for other classification applications.

Acknowledgements This work is supported by the National Key Research and Development Program of China under Grant No. 2016YFB1000904. Chengyu Wang is partially supported by the Outstanding Doctoral Dissertation Cultivation Plan of Action under Grant No. YB2016040.

References

1. Alidaee B, Glover F, Kochenberger GA, Wang H (2007) Solving the maximum edge weight clique problem via unconstrained quadratic programming. Eur J Oper Res 181(2):592–597
2. Bagga A, Baldwin B (1998) Entity-based cross-document coreferencing using the vector space model. In: ACL-COLING, pp 79–85
3. Bunescu RC, Pasca M (2006) Using encyclopedic knowledge for named entity disambiguation. In: EACL
4. Carmel D, Chang M-W, Gabrilovich E, Hsu B-JP, Wang K (2014) Erd'14: entity recognition and disambiguation challenge. In: SIGIR Forum vol 48, no 2, pp 63–77
5. Celeux G, Govaert G (1992) A classification EM algorithm for clustering and two stochastic versions. Comput Stat Data Anal 14(3):315–332
6. Chiang M-F, Liou J-J, Wang J-L, Peng W-C, Shan M-K (2013) Exploring heterogeneous information networks and random walk with restart for academic search. Knowl Inf Syst 36(1):59–82
7. Cornolti M, Ferragina P, Ciaramita M, Rüd S, Schütze H (2016) A piggyback system for joint entity mention detection and linking in web queries. In: WWW, pp 567–578
8. Dalvi BB, Cohen WW, Callan J (2013) Exploratory learning. In: ECML-PKDD, pp 128–143
9. Ferreira AA, Gonçalves MA, Laender AHF (2012) A brief survey of automatic methods for author name disambiguation. In: SIGMOD Record, vol 41, no 2, pp 15–26
10. Ganea O-E, Ganea M, Lucchi A, Eickhoff C, Hofmann T (2016) Probabilistic bag-of-hyperlinks model for entity linking. In: WWW, pp 927–938
11. Han X, Sun L, Zhao J (2011) Collective entity linking in web text: a graph-based method. In: SIGIR, pp 765–774
12. Kanani PH, McCallum A, Chris P (2007) Improving author coreference by resource-bounded information gathering from the web. In: IJCAI, pp 429–434
13. Lao N, Cohen WW (2010) Relational retrieval using a combination of path-constrained random walks. Mach Learn 81(1):53–67
14. Li C, Cheung WK, Ye Y, Zhang X, Chu D-H, Li X (2015) The author-topic-community model for author interest profiling and community discovery. Knowl Inf Syst 44(2):359–383
15. Pei L, Luna DX, Andrea M, Divesh S (2011) Linking temporal records. In: PVLDB, vol 4, no 11, pp 956–967
16. Li S, Cong G, Miao C (2012) Author name disambiguation using a new categorical distribution similarity. In: ECML-PKDD, pp 569–584
17. Li Y, Tan S, Sun H, Han J, Dan R, Yan X (2016) Entity disambiguation with linkless knowledge bases. In: WWW, pp 1261–1270
18. Pitts M, Savvana S, Roy SB, Mandava V (2014) ALIAS: author disambiguation in Microsoft academic search engine dataset. In: EDBT, pp 648–651
19. Qian Y, Hu Y, Cui J, Zheng Q, Nie Z (2011) Combining machine learning and human judgment in author disambiguation. In: CIKM, pp 1241–1246

20. Shen W, Han J, Wang J (2014) A probabilistic model for linking named entities in web text with heterogeneous information networks. In: SIGMOD, pp 1199–1210
21. Shen W, Wang J, Han J (2015) Entity linking with a knowledge base: issues, techniques, and solutions. TKDE 27(2):443–460
22. Shen W, Wang J, Luo P, Wang M (2012) LIEGE: link entities in web lists with knowledge base. In: KDD, pp 1424–1432
23. Shen W, Wang J, Luo P, Wang M (2012) LINDEN: linking named entities with knowledge base via semantic knowledge. In: WWW
24. Shi C, Li Y, Yu PS, Bin W (2016) Constrained-meta-path-based ranking in heterogeneous information network. Knowl Inf Syst 49(2):719–747
25. Sil A, Florian R (2016) One for all: towards language independent named entity linking. In: ACL, pp 2255–2264
26. Solecki B, Silva L, Efimov D (2013) KDD cup 2013: author disambiguation. In: KDD Cup 2013 workshop, pp 9:1–9:3
27. Sun Y, Han J, Yan X, Yu PS, Tianyi W (2011) Pathsim: meta path-based top-k similarity search in heterogeneous information networks. In: PVLDB, vol 4, no 11, pp 992–1003
28. Sun Y, Han J, Zhao P, Yin Z, Cheng H, Wu T (2009) Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In: EDBT, pp 565–576
29. Tang J (2016) Aminer: toward understanding big scholar data. In: WSDM, p 467
30. Wang C, Zhang R, He X, Zhou A (2016) Error link detection and correction in Wikipedia. In: CIKM, pp 307–316
31. Wang X, Tang J , Cheng H, Yu PS (2011) ADANA: active name disambiguation. In: ICDM, pp 794–803
32. Yang Y, Chang M-W (2015) S-MART: novel tree-based structured learning algorithms applied to tweet entity linking. In: ACL-IJCNLP, pp 504–513
33. Yin X, Han J, Yu PS (2007) Object distinction: distinguishing objects with identical names. In: ICDE, pp 1242–1246
34. Zhang B, Dundar M, Al Hasan M (2016) Bayesian non-exhaustive classification. A case study: online name disambiguation using temporal record streams. In: CIKM, pp 1341–1350
35. Zwicklbauer S, Seifert C, Granitzer M (2016) Robust and collective entity disambiguation through semantic embeddings. In: SIGIR, pp 425–434

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Chengyu Wang is a Ph.D. candidate in School of Computer Science and Software Engineering, East China Normal University (ECNU), China. He received his B.E. degree in Software Engineering from ECNU in 2015. His research interests include web data mining, information extraction and natural language processing. He is working on the construction and application of large-scale knowledge graphs.

Xiaofeng He is a Professor in Computer Science at School of Computer Science and Software Engineering, East China Normal University, China. He obtained his Ph.D. degree from Pennsylvania State University, USA. His research interests include machine learning, data mining and information retrieval. Prior to joining ECNU, he worked at Microsoft, Yahoo Labs and Lawrence Berkeley National Laboratory.

Aoying Zhou is a Professor in Computer Science at East China Normal University (ECNU), where he is heading School of Data Science and Engineering. Before joining ECNU in 2008, Aoying worked for Fudan University at the Computer Science Department for 15 years. He is the winner of the National Science Fund for Distinguished Young Scholars supported by NSFC and the professorship appointment under Changjiang Scholars Program of Ministry of Education. He is now acting as a vice-director of ACM SIGMOD China and Database Technology Committee of China Computer Federation. He is serving as a member of the editorial boards VLDB Journal, WWW Journal, and etc. His research interests include data management, memory cluster computing, big data benchmarking and performance optimization.