arXiv:2302.11848v2 [cs.IR] 6 Jun 2023

网络规模学术姓名消歧：WhoIsWho 基准、排行榜和工具包

Bo Chen 清华大学
cb21
@mails.tsinghua.edu.cn

Jing Zhang*中国人民大学
zhang-jing@ruc.edu.cn

Fanjin Zhang 清华大学
zfj17@mails.tsinghua.edu.cn

Tianyi Han 智谱AI
tianyi.han@aminer.cn

Yuqing Cheng
智谱AI
yuqing.cheng@aminer.cn

Xiaoyan Li
智谱AI
xinyan.li@aminer.cn

Yuxiao Dong 清华大学
yuxiaod
@tsinghua.edu.cn

Jie Tang*
清华大学
jietang@tsinghua.edu.cn

摘要

姓名消歧是在线学术系统中的一个基础问题，随着研究论文数量不断增长，如今正面临着更大的挑战。例如，在在线学术搜索平台 AMiner 上，约有 10% 的姓名对应着 100 位以上的作者。由于现有研究所使用的数据集规模较小或质量较低，这些现实世界中的棘手案例尚未得到有效解决。基于多种不同数据集设计的各类任务和评估协议，也进一步阻碍了有效算法的开发。为此，我们推出了 WhoIsWho：一个大规模基准测试，包含超过 1,000,000 篇论文，采用交互式标注流程构建；同时提供涵盖综合任务的排行榜，以及一个易于使用的工具包，封装了完整流程，并提供应对这些任务所需的最强大特征和基线模型。我们开发的强基线模型已部署于 AMiner 系统中，用于每日分配 arXiv 论文[1]。

CCS 概念

• **信息系统** → **数据管理系统**；**信息集成**；**实体解析**；

关键词

姓名消歧，基准

ACM 参考格式：

Bo Chen、Jing Zhang、Fanjin Zhang、Tianyi Han、Yuqing Cheng、Xiaoyan Li、Yuxiao Dong 和 Jie Tang。2023。网络规模学术姓名消歧：WhoIsWho 基准、排行榜和工具包。载于《 *第 29 届 ACM SIGKDD 知识发现与数据挖掘会议（KDD '23），2023年8月6日至10日，美国加利福尼亚州长滩。ACM，美国纽约州纽约市，16 页。* https://doi.org/10.1145/3580305.3599930

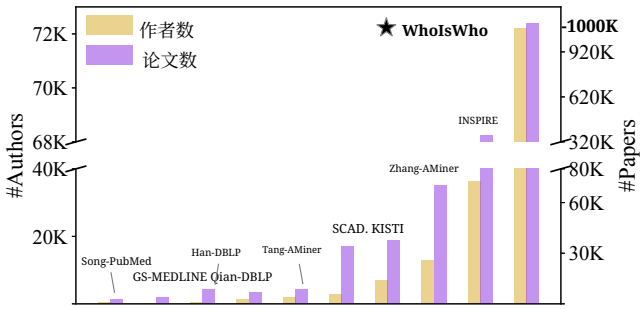


图1：当前主流姓名消歧基准数据集的规模。其中，WhoIsWho规模最大，包含1,000+ 个姓名、70,000+ 位作者和1,000,000+篇论文。

1 引言

姓名消歧旨在厘清“谁是谁”，是 Google Scholar2、Semantic Scholar3 和 AMiner[4] 等在线学术系统面临的基本问题之一。过去几十年来，各个科学领域的研究论文数量激增。例如，Google Scholar、Bing Academic Search 和 AMiner 均已索引约 3 亿篇论文 [10, 33, 36]。因此，在现代数字图书馆中，作者姓名歧义问题——同一作者使用不同姓名变体，或不同作者拥有完全相同的姓名（即同名者）——变得愈发复杂。例如，截至2023年1月，AMiner 上有超过10,000位名为“Yang Yang”的作者。其中三位显示在图2中。由于这三位作者都是计算机科学家，他们的论文之间存在复杂的关联。属于“Yang Yang(THU)”的论文 P_5被错误地归到了“Yang Yang(UND)”名下，因为这两位“Yang Yang”都与“孙逸洲”合著过论文，因而 P_5 之间出现了看似可靠的共同作者和共同关键词关系，以及

*Jing Zhang 和 Jie Tang 是通讯作者。[1]公开排行榜见 http://whoiswho.biendata.xyz/。工具包见 https://github.com/THUDM/WhoIsWho。每日 arXiv 论文分配的在线演示见 https://na-demo.aminer.cn/arxivpaper。

[2]https://scholar.google.com/ [3] https://www.semanticscholar.org/ [4] https://aminer.org

允许免费为个人或课堂用途制作本作品的部分或全部数字或纸质副本，但不得为营利或商业利益制作或分发副本，且副本必须附有此声明和完整引文

在首页。必须尊重本作品中第三方组件的版权。其他所有用途，请联系所有者/作者。*KDD '23，2023年8月6日至10日，美国加利福尼亚州长滩*© 2023 版权所有者/作者所有。ACM ISBN 979-8-4007-0103-0/23/08。
https://doi.org/10.1145/3580305.3599930

“Yang Yang(UND)”的正确论文 P_4 。此外，“Yang Yang(THU)”和“Yang Yang(ZJU)”其实是同一个人，但由于毕业后机构变动，被分成了两位不同的作者。这一真实案例展现了在线学术系统中姓名消歧面临的巨大挑战；然而，由于基准规模较小、质量较低，且任务设计与评估设置不统一，现有研究 [3, 18,20, 21, 32, 35, 38, 47–49], 无法应对这些挑战。

具体而言，尽管若干姓名消歧基准（如 PubMed [40, 45], MAG [46], DBLP [14], 等 [17, 37],）直接从现有数字图书馆中采集，但其中不可避免地存在虚假信息和归属错误，如图2所示，这些问题不利于构建有效的算法 [4, 44]。鉴于此，另一些研究尝试从在线噪声数据中人工标注少量高质量数据，以减少这些噪声 [11, 30, 34]带来的负面影响。然而，如图1所示，其中大多数基准的实例数量不足。此外，基于这些基准，先前的研究定义了各种任务和评估协议，使我们无法公平地比较不同方法，从而阻碍姓名消歧社区的发展。

本工作。我们提出了 WhoIsWho，这是一个基准、一个排行榜，以及一个用于网络规模学术姓名消歧的工具包。具体而言，WhoIsWho 具有以下特点：

- **交互式大规模基准构建**。为了创建具有挑战性的基准，我们设计了一个交互式标注流程，借助所开发的可视化工具，对同一姓名下高度歧义的论文作者隶属关系进行标注。10+名专业标注人员受聘开展标注任务，每人耗时约 24 人月。迄今为止，我们已经发布了一个大规模、高质量且具有挑战性的基准，其中包含超过 1,000 个姓名、70,000 名作者和 1,000,000 篇论文。图1显示，WhoIsWho 基准的规模比现有人工标注数据集大出几个数量级。
- **包含综合任务的竞赛排行榜**。为了公平比较各种名称消歧方法，我们举办了包含两条赛道的竞赛：第一条是从头开始的姓名消歧（SND）旨在将同一作者的论文归为一组，以满足从零开始创建原始学术系统的需求。另一项任务是实时姓名消歧（RND），也称为增量姓名消歧，目标是将新到达的论文分配给现有的已明确作者。RND 任务对于维护现有在线学术系统中论文的常规分配至关重要，因为这类系统拥有大量已明确的作者档案。除此之外，我们还定义了错误分配检测（IND），旨在纠正在线论文与作者之间的归属错误，以确保学术系统的可靠性。截至目前，前两项任务已举办三轮竞赛，吸引了超过 3,000 名研究人员。此外，我们还定期维护排行榜，以跟踪最新进展。IND 任务的竞赛正在积极筹备中。
- **易于使用的工具包**。为了帮助研究人员快速入门姓名消歧领域，我们总结了研究成果，并构建了一个端到端流程，以规范整个姓名消歧过程，包括数据加载、特征创建、模型构建和评估。我们全面研究了竞赛获胜方法，汇集了最有效的特征和模型，并将其封装到工具包中。最终用户可以自由调用基线方法和封装好的特征，开发自己的算法。

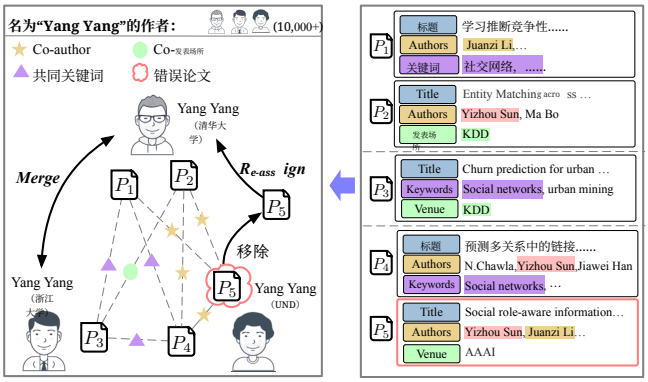


图2：说明如何为姓名为“Yang Yang”的作者进行标注所面临的挑战。论文 p_5 由于与同一位第三方的共同署名而被错误分配。两位作者因机构变动而被错误地分开。

我们深入分析了竞赛获胜者所用方法采用的特征，发现融合多模态特征——即涉及论文属性的语义特征，以及由共同作者、共同机构和共同发表场所关联构成的关系特征——对姓名消歧方法的性能提升最大。基于这些发现，我们提供了简单而有效的基线方法（RND/SND-all），其表现与顶尖竞赛方法不相上下。尤其是，RND-all 已部署在 AMiner 上，用于每日分配 arXiv 论文。

总而言之，WhoIsWho 是一个持续开展、由社区推动的开源项目。我们计划随着时间推移更新排行榜，并提供新的数据集和方法。我们也欢迎大家通过 oagwhoiswho@gmail.com 贡献力量。

2 WHOISWHO 基准

本节首先介绍用于构建大规模高质量基准的交互式标注流程，然后展示该基准的内在分布。

2.1 交互式基准构建

我们将交互式基准构建流程形式化为两个子模块：数据收集和数据标注。

2.1.1 数据集收集。实际上，我们从 AMiner [33] 收集原始数据。为获取噪声更少、歧义更高的姓名消歧数据，我们采用以下规则：

按 H 指数选择作者。对于 AMiner 中的每位作者，我们计算其 H 指数 [12], 这一用于衡量专家影响力的指标，然后保留 H 指数较高的作者。一般认为，知名度越高的作者，其个人资料中的噪声越少，因为他们可能已经在学术平台上完成了自我澄清。具体而言，我们根据作者的 H 指数值按降序排列，并筛除 H 指数低于 5 的作者。这一阈值是文献中用于识别在其研究领域具有显著影响力的作者的广泛

表 1：数据标注流程。对于由三名标注人员执行的操作，采用多数投票来解决冲突。Anno. 是 annotators（标注人员）的缩写。

步骤	#标注人员数	操作
清理（大致）	1	**1.** 删除不属于相关作者的论文； **2.** 将合并过度的作者档案拆分为多个作者。
验证	3	**1.** 与"清理"步骤相同，用于处理更难的情况错误论文。
Add	3	**1**。将未分配论文添加到特定作者名下。
合并	3	**1**。将不同的作者档案合并为同一位作者。

文献中用于识别在其研究领域具有显著影响力的作者的公认标准。

选择歧义较大的姓名。我们统计 AMiner 中同名作者的数量。"同名"指通过姓名阻塞方式统一姓名，例如将姓氏移至名字前，或保留除姓氏外的所有姓名首字母 [2, 14]。例如，"Jing Zhang"的变体包括"Zhang Jing"、"J 张"和"Z 静"。使用该姓名的作者越多，姓名的歧义就越大。我们筛除作者数量低于某一阈值的姓名，使 WhoIsWho 更具挑战性[5]。

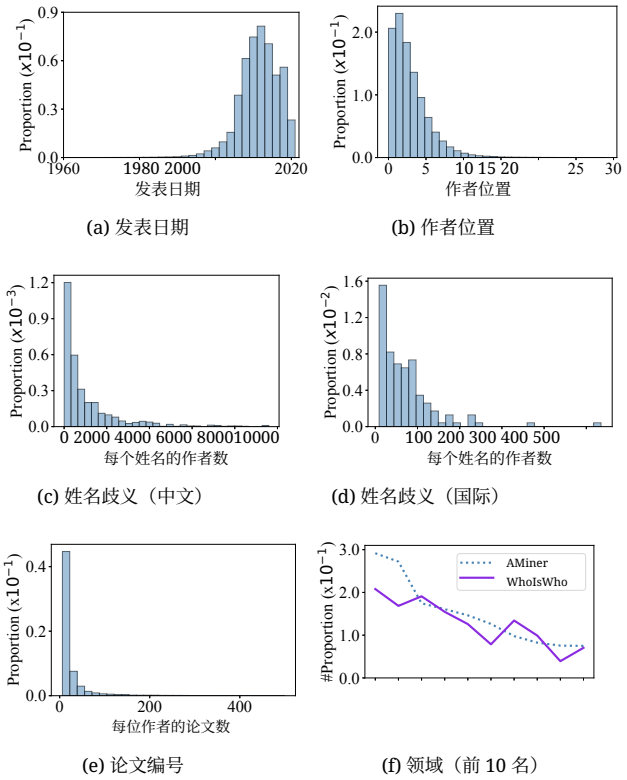
获得歧义较大的姓名及其对应作者后，我们收集每位作者的论文。具体而言，我们收集所有作者的论文标题、作者姓名、机构、关键词、摘要、出版年份和发表场所（会议或期刊），作为论文的属性。此外，还有大量论文尚未分配给任何作者。为了增加基准的挑战性，我们也收集这些论文，并将其称为未分配论文；这些论文的作者与基准中的作者同名，因此在数据标注流程中可能会被分配给基准中的作者。

2.1.2 数据集标注。图2展示了一些现实世界中姓名消歧的棘手案例。由于论文之间的关系错综复杂，标注人员很难为这些案例贴上标签。鉴于此，我们设计了一款交互式标注工具[6]，改编自 [29]，不仅提供有关论文和作者的详细信息，还提供多种实用的原子操作，帮助标注人员执行任意操作。图12展示了一个示例。该工具允许标注人员

进行交互式标注，因为每次执行操作后，作者档案都会更新并展示给标注人员。

借助该工具，我们制定了四个标准化的标注步骤（详见表 1），以确保人工标注流程能够合理开展。总体而言，标注人员有权移除错误论文、添加未分配论文、将一位作者拆分为两位作者，以及合并两位作者。具体来说，第一步**"清理"**允许标注人员从目标作者中移除或拆分明显错误的论文。这类论文与目标作者涉及不同主题。随后，**"验证"**步骤允许标注人员对难以识别的错误论文执行同样的"清理"功能。这类论文与目标作者的主题相关。之后，**"添加"**步骤

[5]我们将 WhoIsWho 中的阈值设为 6。[6] https://www.aminer.cn/billboard/id:5e42777f530c70f19522863e



(a) 发表日期　(b) 作者位置

(c) 姓名歧义（中文）　(d) 姓名歧义（国际）

(e) 论文编号　(f) 领域（前 10 名）

图 3：WhoIsWho 基准统计

使标注人员能够将未分配论文添加给相关作者。最后，**"合并"**步骤允许标注人员将两位作者的论文合并为一位作者的论文。由于后三个步骤比第一步更具挑战性，要求三名标注人员对同一姓名进行标注，并通过多数表决汇总其结果。值得注意的是，标注人员每次都会将同名作者的所有论文一并标注。为防止他们随意移除论文，标注人员必须为每位作者保留至少80%的论文。

总而言之，一方面，所设计的交互式标注流程为论文提供了丰富的信息，充分支持标注人员有效地标注数据集。另一方面，每篇论文至少由 10 位熟练的标注人员审核，这进一步保证了 WhoIsWho 的质量。

2.2 WhoIsWho 基准的统计信息

我们对 WhoIsWho 基准进行了全面分析，以展示其在多个维度上的优势，如图 3 所示。

标注的作者身份的准确性。我们首先检查人工标注的作者身份的准确性。为此，我们从基准中随机抽取 1,000 篇论文，并人工核验每篇论文属于哪位作者。每篇论文均由三名熟练的标注人员通过多数投票进行核验。最终准确率为 99.6%，仅出现四个归属错误，这表明该基准提供了大量高质量实例。

发表日期分布。 图 3(a) 展示了论文发表日期的分布。在此之前，科学文献数量很少

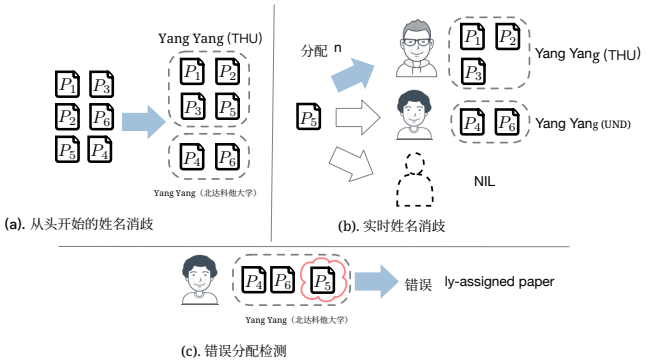


图4：三名称消歧任务。

于 2000 年，因为当时管理数字图书馆仍是一项相对较新的技术。随着 2020 年后互联网的快速发展，数字记录的数量增长得更快。然而，2022 年前后的记录少于 2010 年前后，这表明在线姓名消歧系统可能无法及时分配最新论文。

作者位置分布。 若干数据集专注于消歧论文中特定位置上的作者。例如，Song-PubMed [30]旨在对第一作者进行消歧，这会给姓名消歧方法带来偏向特定作者位置的信息。相反，WhoIsWho 基准平等考虑所有作者位置，如图3(b)所示的合理长尾曲线。

姓名歧义分布。 不同族群的作者姓名通常具有不同程度的歧义。例如，与其他国籍的作者相比，中国作者更难消歧 [9, 15, 16]。图3(c)和3(d)分别展示了 AMiner 中每个中文姓名和国际姓名对应的已澄清作者档案分布，表明中文姓名比国际姓名更具歧义。由于我们的目标是构建一个高歧义基准，以促进姓名消歧方法的发展，因此数据集中收录的中文姓名多于国际姓名，约占作者姓名的87%。

论文数量分布。 我们还展示了基准中每位作者的论文数量分布，如图 3(e) 所示。长尾分布表明，大多数情况下论文数量适中，只有少数著名科学家拥有数百篇出版物。

领域分布。 与仅涵盖偏向性领域的若干数据集相比，例如基于 PubMed [30, 45]、主要聚焦医学领域的数据集，WhoIsWho 对一般学科有着广泛的覆盖。为验证这一点，我们随机抽取了 100,000 篇论文，并采用 Scopus 中 SCI mago 期刊排名（SJR）[7] 的分类体系来获取论文所属领域。频率最高的前 10 个领域如图 3(f) 所示，这表明该基准不仅覆盖了多种领域，而且能够代表 AMiner 中的整体领域分布。

[7]http://www.scimagojr.com

3 WHOISWHO 任务与竞赛

本节首先介绍三个具有标准化评估协议的姓名消歧任务。随后，我们回顾三轮历史竞赛，以及基于不同发布版本中所定义任务构建的常规排行榜。

3.1 任务构建与评估协议

这里，我们形式化定义了三项任务，即从零开始的姓名消歧、实时姓名消歧和错误分配检测，并介绍相应的评估指标，如图 4 所示。

定义 1。 论文。一篇论文 p 关联着多个属性字段，即 $p = \{x_1, \cdots x_F\}$，其中 $x_f \in p$ 表示第 f 个属性。F 是属性的数量。

定义 2. 作者。作者 a 由一组论文组成，即 $a = \{p_1、 \cdots p_n\}$，其中每篇论文为 $p_i = \{x_1 \cdots x_F\}$，n 是作者 a 所发表的论文数量。

定义 3. 候选论文。给定一个以 na 表示的人名，$\mathcal{P}^{na} = \{p_1^{na}、 \ldots\ldots、 p_N^{na}\}$ 是由任何姓名为 na 的作者撰写的候选论文集合。

定义 4. 候选作者。给定一个以 na 表示的人名，$\mathcal{A}^{na} = \{a_1^{na}、 \ldots\ldots、 a_M^{na}\}$ 是具有相同姓名 na 的候选作者集合。“相同姓名”指的是使用姓名阻塞技术 *[2, 14]* 来统一姓名的方式。

3.1.1 从头开始的姓名消歧。 在构建数字图书馆之初，我们需要将大量已发表的论文划分为若干组，每组代表属于同一个人的论文。为此，我们将从头开始的姓名消歧形式化为一个聚类问题。

问题 1. 从头开始的姓名消歧（***SND***）。给定一组候选论文 $\mathcal{P}^{na}$，*SND* 旨在找到一个函数 Φ，将 $\mathcal{P}^{na}$ 划分为一组互不相交的簇 C^{na}，即，

$$\Phi(\mathcal{P}^{na}) \rightarrow C^{na}, where\ C^{na} = \{C_1^{na}, C_2^{na}, \cdots, C_K^{na}\},$$

其中，每个簇都由同一作者撰写的论文组成，即$\mathbb{I}(p_i^{na}) = \mathbb{I}(p_j^{na}) \ \forall (p_i^{na}, p_j^{na}) \in C_k^{na} \times C_k^{na}$；不同簇则包含不同作者撰写的论文，即 $\mathbb{I}\ p_i^a \neq \mathbb{I}\ p_j^a \ \forall\ p_i^a\ p_j^a \in ()\ ()\ (,)$，$C_k^a \times C_{k'}^a,\ k \neq k\ \ \mathbb{I}\ p_i^{na}$ p_i^{na} ′。 () 是论文 的作者标识。

评估协议。我们采用宏平均成对 F1 来评估相关社交网络分析方法的性能，该指标已被许多社交网络消歧方法广泛采用 [18, 28, 31, 47, 49]。

3.1.2 实时姓名消歧。在当前阶段，将新论文分配给现有作者对于在线数字图书馆至关重要。例如，AMiner 每月接收超过 500,000 篇新论文。为此，我们将实时姓名消歧形式化为一个分类问题。

问题 2. 实时姓名消歧（***RND***）。给定一篇论文 p^{na}（即其中待消歧的作者姓名为 na 的论文）以及候选作者集合 $\mathcal{A}^{na}$，正确作者 a^* 可以是 $\mathcal{A}^{na}$ 中的真实作者，也可以是不存在的作者档案，即空值。我们的目标是学习一个函数，将论文 p^{na} 分配给 a^*，即，

$$\Psi(p^{na}, A^{na}) \rightarrow a^*$$

在线学术平台上经常会出现空值情况。假设本科生在会议或期刊上发表他们的第一篇论文，但当前数据库尚未为他们建立作者档案，那么就无法将论文分配给任何作者。鉴于此，我们在 RND任务中纳入了 NULL 情景。已有研究 [3] 也考虑了空值情况，但他们创建的是合成的空值标签，而非纳入实际的 NIL 案例。据我们所知，我们是首个在 WhoIsWho 基准中考虑空值情况，并使用人工标注的真实 NIL 案例的研究。

评估协议。 我们提出加权 F1，用于评估解决 RND 问题的方法。对于待消歧的作者 a，我们按如下方式计算指标：

$$Precision^a = \frac{\#PapersCorrectlyAssignedToTheAuthor}{\#PapersAssingedToTheAuthor},$$

$$Recall^a = \frac{\#PapersCorrectlyAssignedToTheAuthor}{\#PapersOfTheAuthorTobeAssinged},$$

$$Weight^a = \frac{\#PapersOfTheAuthorTobeAssigned}{\#TotalPapersTobeAssigned},$$

其中，精确率衡量被预测归属于a的论文的正确性，召回率衡量 a的实际论文中有多少篇能够被正确分配给 a。随后，我们根据每位作者的精确率和召回率计算其 F1 分数。之后，再按照每位作者的权重对 F1 分数取平均，该权重由其将被分配的论文所占比例决定。我们采用加权平均策略，以减轻某些极端情况带来的负面影响，例如仅有一篇论文的作者。

3.1.3 错误分配检测。 由于 SND 和 RND 方法带来的不可避免的累积错误会严重影响后续分配的有效性，因此错误分配检测是一项至关重要的任务，用于检测并移除错误分配的论文。

问题 *3*：错误分配检测（*IND*）。给定一个由来自不同作者 K 的多篇论文组成的混淆作者实体 $a^* = \{p_i, \cdots p_j, p_a, \cdots p_b, \cdots p_m, \cdots p_n\}$，其中 $a_1 = \{p_i, \cdots p_j\}$、$a_2 = \{p_a, \cdots p_b\}$ 和 $a_K = \{p_m, \cdots p_n\}$。假设 a_1 在 a^* 中涵盖的论文比例最高，我们设定 $a^* = a_1$。因此，将归属于 $\{a_2, \cdots a_K\}$ 的论文定义为待检测的错误分配论文。

评估协议采用异常检测领域广泛使用的 ROC曲线下面积（AUC） [22] 以及平均准确率均值（MAP）作为评估指标。MAP 更关注错误案例的排名。

3.1.4 讨论。提出的三项姓名消歧任务揭示了在线数字图书馆中相关姓名消歧问题的生命周期。具体而言，SND 任务反映了数字图书馆建设早期阶段的需求；RND 任务对应当前在线平台的迫切需求；IND 任务则致力于纠正姓名消歧算法积累的错误，这对于维护姓名消歧系统的可靠性至关重要。此外，这三项任务可以作为其他任何复杂姓名

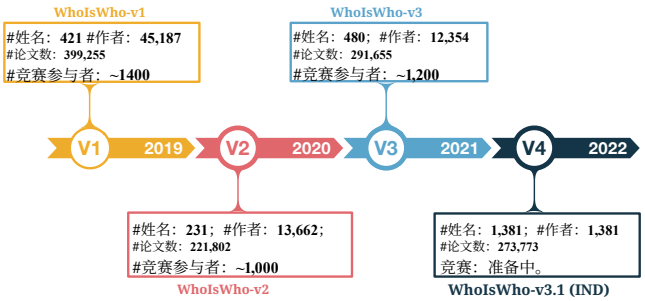


图 5：WhoIsWho 基准的发布时间及举办的竞赛。

消歧任务的骨干。我们认为，在这些任务上表现更好的姓名消歧方法，足以应对大多数姓名消歧场景。尽管 Zhang 和 Tang [44] 已经提出了类似类型的任务，但我们通过以下方式对其进行了改进：1）将 NIL问题纳入考虑，并将 RND 问题形式化为更一般的分类问题，而非排序问题；2）规范化三项任务的评估协议；3）为前两项任务组织竞赛，以推动其完成。

3.2 历史竞赛与常设排行榜

从 2019 年到 2022 年，WhoIsWho 定期发布了三个版本的基准。为推动社区发展，我们在 BienData[8]上赞助举办了三轮姓名消歧竞赛。基准的发布时间及相应竞赛的时间线如图 5 所示。迄今为止，全球已有 3,000 多人下载 WhoIsWho 基准，下载次数超过 10,000 次。WhoIsWho 已成为姓名消歧社区中最知名、最具代表性的基准之一。此外，为了帮助有意随时解决姓名消歧问题的研究人员，我们基于 WhoIsWho 最新发布的基准，维护了一个常设排行榜，并举办相应竞赛。

在接下来的部分中，我们简要回顾竞赛获胜者提出的方法，并以此为基础开展深入的实证分析，探究可能对姓名消歧方法性能产生重大影响的关键因素。

3.2.1 竞赛冠军的方法。鉴于截至目前表现最佳，我们回顾在 SND 和 RND 的前两项任务中竞赛获胜者所采用的方法。如何衡量论文与作者之间细粒度的相似度，对找到这两项任务的解决方案至关重要。因此，为了衡量这些相似度，我们需要构建作者与论文之间的交互关系，这一点需要优先探究。接下来的部分将略去一些技术细节，重点介绍量化论文与作者之间关联的策略。

从头开始的姓名消歧。SND 任务旨在将同一作者撰写的论文归为一组。竞赛冠军将论文间相似度分为两类。

[8]https://www.biendata.xyz/

语义方面。竞赛冠军将论文的标题、发表场所、作者所属机构、年份和关键词视为语义特征，并据此衡量论文之间的主题相似度。具体而言，他们首先基于 WhoIsWho 基准中所有论文的语义特征学习 word2vec [23] 词嵌入。然后，他们将一篇论文的语义特征映射到相应的词嵌入，并取平均值作为论文嵌入。最后，他们基于这些语义嵌入计算论文间的软语义相似度。

关系方面。竞赛冠军将作者姓名和机构作为论文的关系特征。例如，同一作者姓名同时出现在两篇论文中，反映了它们之间的关系。具体而言，他们将论文视为节点，将论文之间的关联视为边，从而构建关系图。如果两篇论文有相同姓名的合著者，则添加合著者边。如果相关作者在两篇论文中具有相同的机构，则添加共同机构边。之后，他们使用 metapath2vec [6] 获取论文的关系嵌入。最后，他们根据这些关系嵌入计算论文之间的关系相似度分数。

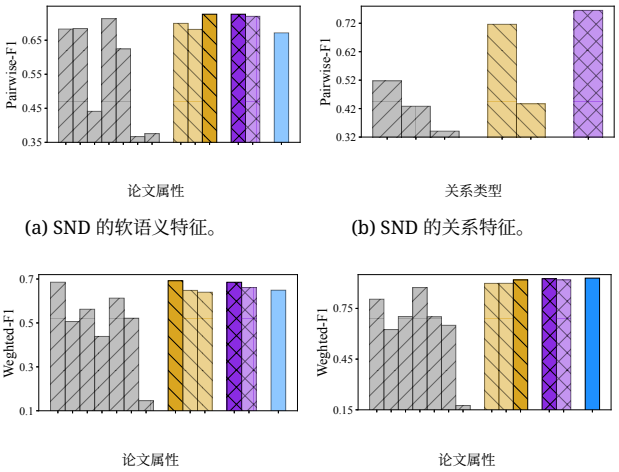
此外，竞赛冠军结合两种多模态相似度来估算论文之间的最终相似度，然后使用 DBSCAN [7] 获得聚类结果。

实时姓名消歧。RND任务侧重于衡量论文与每位候选作者的一组论文之间的关联。与 SND 任务相比，竞赛冠军能够以如下方式捕获未分配论文与候选作者之间更精确的语义特征。

语义方面。除了软语义特征（即通过嵌入技术测量的特征）之外，他们还考虑了特定语义特征（即通过手工构造特征测量的特征）。对于软语义特征，他们识别目标论文与候选作者的每篇论文之间的相似度，就像 SND 所做的那样。然后，他们采用聚合函数，获得目标论文与候选作者所有论文之间的总体相似度。对于特定语义特征，他们提出了36维人工设计特征，以显式捕捉

目标论文与候选作者之间的语义相关性。完整特征列于表5中。最后，他们将软语义特征与特定语义特征进行拼接，生成最终的相似度特征。然后，他们采用集成方法获得分类结果。

我们注意到，竞赛冠军的方法忽略了对关系属性的刻画，因此提出以下假设：1）与 SND 任务不同，SND 任务只需针对一个姓名构建一次论文关系图，而 RND 任务则需要针对每篇未分配论文，构建其与相应候选作者之间的图，这一过程耗时较长。2）某些临时特征可以在一定程度上捕捉关系相关性。例如，合著者共现特征统计目标论文与候选作者之间的合著者人数，可视为虚拟论文—作者图上的合著者边权重。然而，如何在 RND 任务中建模关系相关性，仍有待深入研究。



(a) SND 的软语义特征。 (b) SND 的关系特征。

(c) RND 的软语义特征。 (d) RND 的特定语义特征。

Figure6：SND 和 RND 任务。

错误分配检测。IND 任务旨在检测累积的错误论文，这对于保障学术系统的可靠性至关重要。然而，目前尚无可用的 IND 基准。为此，我们发布了专用于 IND 任务的 V3.1 数据，其中包含 1,000+ 位作者和 200,000+ 篇论文。据我们所知，我们是首个明确界定并发布相应 IND 基准的团队。此外，我们计划在未来几个月内基于已发布的 WhoIsWho-v3.1 基准举办一场 IND 任务竞赛。

3.2.2 讨论。总之，我们发现，要全面衡量论文之间的相关性，关键在于将多模态特征（即语义特征和关系特征）结合起来。竞赛结果表明，同时捕捉这两方面特征的方法取得了令人瞩目的成果。尽管第三项任务 IND 的竞赛尚未举行，但我们推测 IND 任务也会得出类似结果，因为该任务同样依赖于评估论文之间的一致性。

4 实证因素分析

我们进行了深入的消融研究，以了解各种因素对姓名消歧性能的影响。为确保比较公平，我们只调整关注的因素，其他因素均保持不变。我们采用 WhoIsWho 任务中定义的指标进行评估。每项实验均运行 5 次，并报告 WhoIsWho-v3 验证集上的平均结果。

4.1 语义特征重要性

我们研究了可用论文属性（即标题（T）、关键词（K）、摘要（A）、会议/期刊（V）、年份（Y）、作者姓名（N）和作者所属机构（O））对 SND 和 RND 任务的影响。

从头开始的姓名消歧。为了进行软语义特征分析，我们采用了与竞赛获胜方法类似的实现流程，同时探索不同的属性。

结果。结果如图6(a)所示。标题、关键词、作者姓名和机构字段比其他字段对消歧的影响更显著

表2：不同特征模态（语义或关系）及其组合的性能（%）。

任务	语义特征		关系特征		All
	Soft	临时	关系	Ego	
SND	72.64	-	76.52	-	**88.46**
RND	76.55	93.01	-	72.92	**93.40**

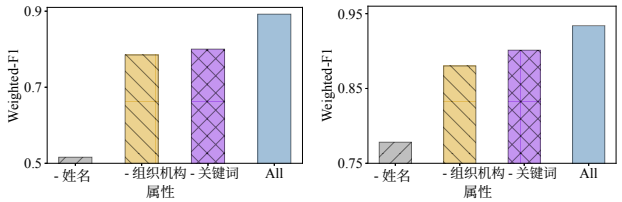
。摘要字段包含许多冗余词语和噪声。发表场所和年份也无法捕捉论文之间的相似度。**（1）结合一致的属性或许能更好地表达语义相关性**。 如黄色柱所示，我们将这四种有效的单一特征组合起来。标题与姓名的组合表现甚至不如其中任何一种单一属性。我们推测，与标题、关键词和机构相比，作者姓名具有更多语言学特征，而标题、关键词和机构之间则存在语义相关性。因此，组合两种差异较大的属性会导致性能下降。当标题与关键词或机构搭配时，其性能有所提升，这表明组合一致的属性可能更好地表达语义相关性。**（2）标题、关键词和机构的组合表现最佳**最后，由紫色柱表示的标题、关键词和机构组合，表现优于将所有属性混合在一起的方案（即蓝色柱）。这表明，未经校准便添加更多属性可能会引入噪声并降低性能。

实时姓名消歧。 我们还采用了 RND 竞赛获胜方法的实现流程。除了软语义特征分析之外，我们还利用表 5 中列出的手工构造特征，探究各种论文属性如何影响姓名消歧方法的性能。

结果。 结果如图6(c)和图6(d)所示。 **(1) 两项任务中的软语义特征呈现出相似的趋势**。关于软语义特征，图6(c)和图6(a)表明，两项任务呈现出共同趋势：1）标题、关键词和机构属性表现良好；2）标题、关键词和机构的组合优于简单混合所有考量的特征。这是意料之中的，因为两项任务都通过相同的软语义特征模态来衡量论文与作者之间的一致性。至于特定语义特征，如图6(d)所示，作者姓名是决定算法性能的最有效因素。**(2) 混合所有属性的效果最佳**。 令人意外的是，代表组合所有特征性能的蓝色柱优于其他组合模式，这表明尽管专用特征表征框架属于语义特征类别，但其潜在偏差与软特征框架不同。

4.2 关系特征重要性

从经验上看，作者姓名和发表场所字段在论文之间表现出更强的关系依赖。此外，机构字段兼具关系和语义特征。因此，我们在论文之间构建了三种关系边：共同作者，即两篇论文仅在共享相同作者姓名时才存在关系；共同机构，即两篇论文仅在共享相同所属机构[9]时才存在关系；共同发表 venue，即两篇论文仅在发表于同一发表场所或期刊时才存在关系。



（a）SND。 （b）RND。

图 7：真实案例分析。

从头开始的姓名消歧。 我们还遵循竞赛冠军方法的实现流程，在构建的关系图中获取关系论文嵌入，同时探究不同关系边的影响。

结果。 图 6（b）展示了使用不同关系类型时的性能。灰色柱状图表明，在单一关系类型中，CoAuthor 的表现最佳，这说明作者姓名所包含的关系信息比语义信息更为重要。CoVenue 的表现最差，这是因为来自不同领域的大量论文可能会发表在同一发表场所/期刊上。综合考虑黄色和紫色柱状图所代表的混合结果后，结合全部三种特征可获得最佳结果；这与第 4.1 节的实证发现一致，即一致的属性组合能够提升性能。

4.3 特征模态重要性

我们探究语义特征和关系特征如何影响消歧效果，并全面考察多模态特征的组合模式，以了解哪些组合表现最佳。对于 SND 任务，我们将标题、关键词和机构等论文属性作为软语义特征。对于关系特征，我们采用三种关系类型，即 CoAuthor、CoOrg 和 CoVenue。对于 RND 任务，除了第 4.1 节中使用的软语义特征和特定语义特征外，我们还为每一对目标论文和候选作者构建异构自我图，以添加关系特征。

结果。表2展示了单一特征模态及其组合的性能。 **（1）融合多模态特征表现最佳**。我们发现，单一模态（即语义特征或关系特征）的表现不如它们的组合模式（即 SND-all 和 RND-all），这表明语义特征和关系特征彼此互补。然而，对于 RND 任务，仅使用特定语义特征就能与其组合相抗衡。关系特征带来的提升微乎其微。这解释了为什么该任务中表现最佳的参赛方法没有利用关系特征。因此，如何有效融入关系方面仍是一个悬而未决的问题。

4.4 总体评估

在本节中，我们将提出的 SND-all 和 RND-all 框架分别与现有最先进的姓名消歧

[9]我们仅考虑待消歧作者所属的机构。

表3：从头开始进行姓名消歧的性能（%）。

模型	成对-精确率	成对召回率	成对F1
G/L-Emb	50.77	84.64	63.48
LAND	61.20	61.12	61.12
IUAD	58.82	65.22	61.63
竞赛冠军	82.72	**96.59**	89.14
SND-all	**83.06**	96.35	**89.22**

表4：实时姓名消歧性能（%）。

模型	加权-精确率	加权-召回率	加权-F1
IUAD	75.53	90.49	82.34
CONNA	90.54	89.22	89.64
CONNA+临时。	90.23	92.64	91.14
竞赛冠军	92.09	**94.95**	93.49
RND-all	**92.14**	94.94	**93.52**

方法进行比较，这些方法分别用于 SND 和 RND 任务10。实验结果基于 WhoIsWho-v3 测试集得出。

对比基线方法。此外，除了竞赛冠军的方法外，我们还比较了其他主流方法，

从头开始的姓名消歧。 **G/L-Emb** [47] 通过图自动编码，在全局论文-论文网络上学习论文嵌入，然后在为每个姓名构建的局部论文-论文网络上微调这些嵌入。**LAND** [28] 利用论文和作者构建异构知识图谱（KG），并运用知识图谱嵌入技术获取节点嵌入，再据此执行聚类方法。**IUAD** [18]通过重构协作网络来确定论文的作者归属，其中节点是作者，边表示合作者关系。**SND-all** 是我们根据第 4.3 节的实证研究提出的强基线方法。它结合软语义特征与异构关系图特征来执行 SND 任务。

实时姓名消歧。 我们采用以下基线方法：**IUAD** [18] 也用于执行 RND任务，方法是重构新到论文与现有作者之间的协作网络。**CONNA** [3] 是一种基于交互关系的模型。它首先在两个属性的词元嵌入之间构建基本交互关系，然后将不同属性矩阵聚合为论文级交互关系，最后将论文级矩阵聚合为作者级交互关系；而**CONNA+临时**。也是一种组合方法，它将手工构造特征融入 [3] 中介绍的 CONNA框架。为确保比较公平，我们采用表 5 中使用的特征。**RND-all** 也是我们根据第 4.3 节的发现提出的方法。它采用第 4.1 节中使用的软语义特征和特定语义特征，并构建异构自我图作为关系特征。两类特征结合起来进行预测。

其他常用方法，如 Louppe 等人 [20], Zhang 等人 [41], Camel [43], 等，在实证中表现不如所采用的基线方法，因此未纳入实验。

[10]我们仅考虑已发布代码的基线方法。

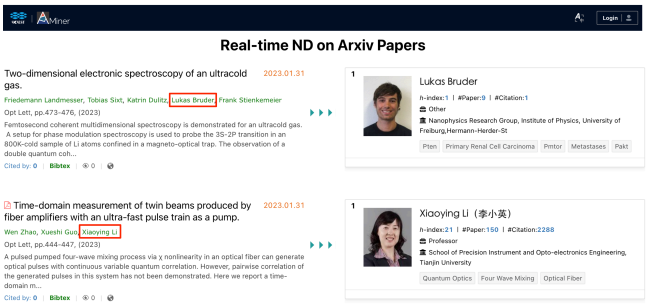


图8：从 arXiv.org 中消歧每日论文的演示。

结果。 表3和表4展示了各种名称消歧方法在两项任务上的性能。所提出的 SND-all、RND-all 以及竞赛获胜者的表现显著优于其他基线方法，成对F1分别高出 25.74~28.10%，加权 F1 高出 2.35~11.80%。我们提出的方法与近期研究提出的基线方法之间存在显著的性能差距，这表明当前常用的姓名消歧方法仍远未达到令人满意的水平，也反映了 WhoIsWho 基准的重要性。此外，我们提出的简单而有效的方法略优于竞赛获胜方法，这说明我们的实证因素分析成功捕捉到了提升名称消歧方法有效性的关键要素。

4.5 真实场景中的性能

WhoIsWho 基准中的论文通常包含丰富的信息，因为标注人员更倾向于处理具有大量属性的论文，这些属性能提供有用的证据来支持他们的判断。不幸的是，在线数字图书馆中总是有许多属性稀疏的论文，也就是说，缺少具有多个属性的论文。以 AMiner 为例，近一半新收录的论文缺少机构属性。为了了解这些论文中的在线姓名消歧场景，我们在这些属性稀疏的案例上执行 SND-all 和 RND-all。

结果。结果如图7所示。其中，没有作者姓名的论文表现最差，成对F1下降了 36.72%，加权 F1 下降了 15.58%。缺少机构或关键词属性也会显著降低姓名消歧算法在两项任务中的在线性能，成对F1下降 8.97~10.51%，加权 F1 下降 4.28~6.40%。结果表明，在线姓名消歧场景比我们在 WhoIsWho 上展示的更加复杂。未来，我们将更新包含稀疏属性的数据集，以推动更多贴近真实世界的在线姓名消歧场景研究。

5 WHOISWHO 工具包

通过自动化数据加载、特征创建、模型构建和评估流程，WhoIsWho 工具包易于研究人员使用，并帮助他们开发新的姓名消歧方法。工具包流程概览如图 9 所示。该工具包与 PyTorch 及其相关的深度学习库（如 Hugging Face [39]）完全兼容。此外，该工具包提供与库无关的数据集对象，可供

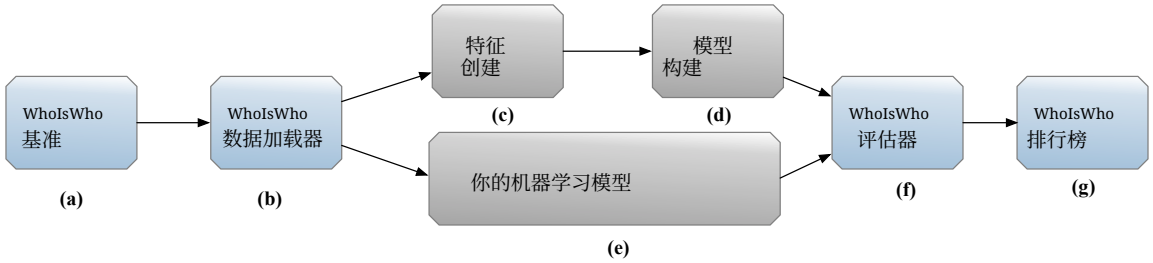


图 9：WhoIsWho 工具包流程概览。（**a**）WhoIsWho 提供了一个具有高度歧义性和海量数据的大规模基准测试。（**b**）WhoIsWho 工具包可自动完成数据集处理和划分。也就是说，数据加载器会自动加载任意版本的数据集，并以标准化方式进一步划分数据集。（**c**）WhoIsWho 工具包提供了灵活的模块，用于特征创建，包括语义特征表征和关系图构建。在此基础上，（**d**）研究人员可以采用工具包库中预定义的模型进行训练和预测。此外，（**e**）研究人员还可以自行构建特征处理流程并开发机器学习模型。（**f**）WhoIsWho 会根据具体任务评估模型，并输出模型在验证集上的性能。最后，（**g**）WhoIsWho 提供公开排行榜，以跟踪最新进展。

任何其他 Python 深度学习框架使用，例如 TensorFlow [1]。为简化说明，我们着重介绍如何使用 PyTorch 构建一个基础的 RND方法，如代码清单 1 所示。更多详情请参阅 https://github.com/THUDM/WhoIsWho。

Arxiv论文消歧。 我们在 AMiner 上部署了由工具包实现的 RND-all方法，以实时消歧来自 arXiv.org 的每日论文。图8展示了一个演示页面。详情请参阅第 A.8 节。我们人工检查了最新的 100 条消歧结果，发现其中 90% 的归属准确。

6 相关工作

这里，我们回顾当前主流的姓名消歧数据集和最先进的姓名消歧算法。

姓名消歧数据集。 数据集的规模会显著影响姓名消歧算法的性能。为解决这一问题，学界近期创建了大量姓名消歧数据集。其中，一些工作直接从现有数字图书馆中收集数据集，包括 PubMed [40, 45], DBLP [14], 等 [17, 37, 46]。然而，如图2所示，归属错误阻碍了有效算法的发展 [4, 44]。另一些工作则尝试基于现有数据库中的噪声数据，人工标注少量数据，以降低数据噪声 [11, 13, 20, 24, 26, 30, 32, 34, 38, 47]。然而，如图1所示，这些数据集大多实例数量不足。详细的数据统计请参阅表6。其中一些数据集的适用范围有限，例如，SCAD-zbMATH [24] 专为数学领域定制。脆弱的归纳偏置会影响在这些数据集上训练的姓名消歧方法的性能和泛化能力。Subramanian 等人 [31] 通过汇总若干小规模数据集构建了一个统一数据集。然而，他们并未检查各组成数据集的质量。

实际任务与算法。 大多数工作都聚焦于 SND 任务。通常，这些工作分为三个步骤：阻塞、论文相似度匹配和聚类。Backes [2] 讨论了姓名阻塞步骤。若干研究着重探讨论文相似度匹配和聚类步骤。早期尝试通过设计人工制定的相似度度量 [5] 来衡量论文之间的相似度。随后，研究人员发现，构建论文相似度图更擅长学习高阶相似度 [6, 8, 13, 27, 47]。至于聚类步骤，采用层次凝聚聚类和 DBSCAN 等聚类方法。其中，实践者更青睐 DBSCAN，因为它无需预先指定簇的数量。

RND任务旨在将新到的论文分配给现有作者，这对于在线学术系统而言是一个更切实可行的场景。除采用已有基线方法外，Qian 等人 [26] 通过共同作者和关键词属性，预测某篇论文由特定作者撰写的可能性。Pooja 等人 [25] 利用动态图嵌入对不断演化的图进行建模。还有一些研究 [18, 42] 进一步采用概率模型进行在线论文分配。

不可避免的累积误差会严重影响姓名消歧算法的有效性。因此，IND 任务对于确保学术系统的可靠性至关重要。遗憾的是，这一问题尚未受到足够关注 [4]。

以往的方法通常在各种小规模数据集上进行评估，这阻碍了该领域的发展。因此，应当建立一个大规模基准测试，设置包含综合任务的常设排行榜，并提供一个易于使用的工具包，用于网络规模学术姓名消歧。

7 结论

本文推出了 WhoIsWho，其中包括用于网络规模学术姓名消歧的基准、排行榜和工具包。具体而言，该大规模且具有高度歧义性的基准有助于设计鲁棒算法。设有两条赛道的赞助竞赛推动了姓名消歧领域的发展。常设排行榜面向公众开放，用于跟踪最新进展。我们还设计了一个易于使用的工具包，使最终用户能够快速构建自己的算法，并将结果发布到记录最新进展的常设排行榜上。总之，WhoIsWho 是一个持续开展、由社区推动的开源项目。我们也鼓励社区贡献力量。

致谢。 本研究得到了中国科学技术部科技创新重大项目（项目编号：2020AAA0108400和2020AAA0108402）、国家自然科学基金杰出青年科学基金（项目编号：61825602）、国家自然科学基金（项目编号：62076245、62276148）、CCF-Zhipu202306，以及中国人民大学公共计算云的支持。

参考文献

[1] Martín Abadi、Paul Barham、Jianmin Chen、Zhifeng Chen、Andy Davis、Jeffrey Dean、Matthieu Devin、Sanjay Ghemawat、Geoffrey Irving、Michael Isard 等。2016。Tensorflow：一种面向大规模机器学习的系统。。载于 OSDI，第 16 卷。美国佐治亚州萨凡纳，265–283。[2] Tobias Backes。2018。姓名匹配与阻塞对作者消歧的影响。载于第27届 ACM信息与知识管理国际会议论文集。803–812。[3] Bo Chen、Jing Zhang、Jie Tang、Lingfan Cai、Zhaoyu Wang、Shu Zhao、Hong Chen 和 Cuiping Li。2020。CONNA：即时解决姓名消歧问题。IEEE Transactions on Knowledge and Data Engineering（2020）。[4] Bo Chen、Jing Zhang、Xiaokang Zhang、Yuxiao Dong、Jian Song、Peng Zhang、Kaibo Xu、Evgeny Kharlamov 和 Jie Tang。2022。GCCAD：用于异常检测的图对比学习。TKDE'22（2022）。[5] Ricardo G Cota、Anderson A Ferreira、Cristiano Nascimento、Marcos André Gonçalves 和 Alberto HF Laender。2010。一种用于书目引文姓名消歧的无监督、基于启发式的层次方法。美国信息科学与技术学会杂志 61，9（2010），1853–1870。[6] Yuxiao Dong、Nitesh V Chawla 和 Ananthram Swami。2017。metapath2vec：面向异构网络的可扩展表示学习。载于 KDD'17。135–144。[7] Martin Ester、Hans-Peter Kriegel、Jörg Sander、Xiaowei Xu 等。1996。一种用于在含噪大型空间数据库中发现簇的基于密度的算法。。载于 KDD，第96卷。226–231。[8] Xiaoming Fan、Jianyong Wang、Xu Pu、Lizhu Zhou 和 Bing Lv。2011。论基于图的姓名消歧。数据与信息质量杂志 2，2（2011），1–23。[9] Janaína Gomide、Hugo Kling 和 Daniel Figueiredo。2017。书目数据同义词歧义问题中的姓名使用模式。Scientometrics 112，2（2017），747–766。[10] Michael Gusenbauer。2019。Google Scholar 会令其他搜索引擎相形见绌吗？比较12种学术搜索引擎与书目数据库的规模。Scientometrics 118，1（2019），177–214。[11] Hui Han、Hongyuan Zha 和 C Lee Giles。2005。使用 k 路谱聚类方法对作者引文中的姓名进行消歧。载于第5届 ACM/IEEE-CS数字图书馆联合会议论文集。334–343。[12] Jorge E Hirsch。2005。一种量化个人科学研究产出的索引。PNAS'05 102，46（2005），16569–16572。[13] In-Su Kang、Pyung Kim、Seungwoo Lee、Hanmin Jung 和 Beom-Jong You。2011。构建用于作者消歧的大规模测试集。《信息处理与管理》47，3（2011），452–465。[14] Jinseok Kim。2018。评估数字图书馆中的作者姓名消歧：以 DBLP 为例。Scientometrics 116，3（2018），1867–1886。[15] Jinseok Kim 和 Jana Diesner。2016。基于姓名首字母的消歧对大规模合著网络测量的扭曲效应。美国信息科学与技术学会杂志 67，6（2016），1446–1461。[16] Jinseok Kim、Jinmo Kim 和 Jason Owen-Smith。2019。为作者姓名消歧自动生成标注数据：一种迭代聚类方法。Scientometrics 118，1（2019），253–280。[17] Jinseok Kim 和 Jason Owen-Smith。2021。用于大规模评估作者姓名消歧的 ORCID关联标注数据。Scientometrics 126，3（2021），2057–2083。[18] Na Li、Renyu Zhu、Xiaoxu Zhou、Xiangnan He、Wenyuan Cai、Ming Gao 和 Aoying Zhou。2021。论作者消歧：自底向上的协作网络重建。载于2021年 IEEE 第37届数据工程国际会议（ICDE）。IEEE，888–899。[19] Xiao Liu、Da Yin、Jingnan Zheng、Xingjian Zhang、Peng Zhang、Hongxia Yang、Yuxiao Dong 和 Jie Tang。2022。Oag-bert：面向学术知识服务的统一骨干语言模型。载于第28届 ACM SIGKDD知识发现与数据挖掘会议论文集。3418–3428。[20] Gilles Louppe、Hussein T Al-Natsheh、Mateusz Suszhanik 和 Eamonn James Maguire。2016。使用半监督学习进行对族裔敏感的作者消歧。载于 KESW'16。272–287。[21] Dongsheng Luo、Shuai Ma、Yaowei Yan、Chunming Hu、Xiang Zhang 和 Jin-peng Huai。2020。一种学者姓名消歧的集体方法。IEEE Transactions on Knowledge and Data Engineering 34，5（2020）。[22] Xiaoxiao Ma、Jia Wu、Shan Xue、Jian Yang、Chuan Zhou、Quan Z Sheng、Hui Xiong 和 Leman Akoglu。2021。基于深度学习的图异常检测综合综述。IEEE Transactions on Knowledge and Data Engineering（2021）。[23] Tomas Mikolov、Kai Chen、Greg Corrado 和 Jeffrey Dean。2013。向量空间中词表示的高效估计。arXiv预印本 arXiv:1301.3781（2013）。[24] Mark-Christoph Müller、Florian Reitz 和 Nicolas Roy。2017。作者姓名消歧数据集：实证分析与新资源。Scientometrics 111，3（2017），1467–1500。[25] KM Pooja、Samrat Mondal 和 Joydeep Chandra。2022。演化数字图书馆中的在线作者姓名消歧。《神经计算》493（2022），1–14。[26] Yanan Qian、Qinghua Zheng、Tetsuya Sakai、Junting Ye 和 Jun Liu。2015。面向不断扩展的数字图书馆的动态作者姓名消歧。Information Retrieval Journal 18，5（2015），379–412。[27] Ziyue Qiao、Yi Du、Yanjie Fu、Pengfei Wang 和 Yuanchun Zhou。2019。利用异构图卷积网络嵌入进行无监督作者消歧。载于 IEEE Big Data'19。IEEE，910–919。[28] Cristian Santini、Genet Asefa Gesese、Silvio Peroni、Aldo Gangemi、Harald Sack 和 Mehwish Alam。2022。一种基于知识图谱嵌入、利用字面量进行作者姓名消歧的方法。Scientometrics 127，8（2022），4887–4912。[29] Qiaomu Shen、Tongshuang Wu、Haiyan Yang、Yanhong Wu、Huamin Qu 和 Weiwei Cui。2016。Nameclarifier：用于作者姓名消歧的可视分析系统。IEEE Transactions on Visualization and Computer Graphics 23，1（2016），141–150。[30] Min Song、Erin Hea-Jin Kim 和 Ha Jin Kim。2015。探索 PubMed 规模下的作者姓名消歧。Journal of Informetrics 9，4（2015），924–941。[31] Shivashankar Subramanian、Daniel King、Doug Downey 和 Sergey Feldman。2021。S2and：用于作者姓名消歧的基准与评估系统。载于 JCDL'21。IEEE，170–179。[32] Jie Tang、Alvis CM Fong、Bo Wang 和 Jing Zhang。2012。数字图书馆姓名消歧的统一概率框架。IEEE Transactions on Knowledge and Data Engineering 24，6（2012），975–987。[33] Jie Tang、Jing Zhang、Limin Yao、李娟子、Li Zhang 和 Zhong Su。2008。Arnetminer：学术社交网络的提取与挖掘。载于 SIGKDD'08。990–998。[34] Dina Vishnyakova、Raul Rodriguez-Esteban 和 Fabio Rinaldi。2019。一种面向 MEDLINE 作者消歧的新方法及金标准。Journal of the American Medical Informatics Association 26，10（2019），1037–1045。[35] Haiwen Wang、Ruijie Wan、Chuan Wen、Shuhao Li、Yuting Jia、Weinan Zhang 和 Xinbing Wang。2020。利用对抗式表示学习，在异构信息网络上进行作者姓名消歧。载于 AAAI人工智能会议论文集，第34卷。238–245。[36] Kuansan Wang、Zhihong Shen、Chiyuan Huang、Chieh-Han Wu、Yuxiao Dong 和 Anshul Kanakia。2020。Microsoft Academic Graph：专家力量也有不足之时。Quantitative Science Studies 1，1（2020），396–413。[37] Ruijie Wang、Yuchen Yan、Jialu Wang、Yuting Jia、Ye Zhang、Weinan Zhang 和 Xinbing Wang。2018。AceKG：面向学术数据挖掘的大规模知识图谱。载于第27届 ACM信息与知识管理国际会议论文集。1487–1490。[38] Xuezhi Wang、Jie Tang、Hong Cheng 和 S Yu Philip。2011。Adana：主动姓名消歧。载于 ICDM'11。794–803。[39] Thomas Wolf、Lysandre Debut、Victor Sanh、Julien Chaumond、Clement Delangue、Anthony Moi、Pierric Cistac、Tim Rault、Rémi Louf、Morgan Funtowicz、Joe Davison、Sam Shleifer、Patrick von Platen、Clara Ma、Yacine Jernite、Julien Plu、Canwen Xu、Teven Le Scao、Sylvain Gugger、Mariama Drame、Quentin Lhoest 和 Alexander M. Rush。2020。Transformers：最先进的自然语言处理。载于 EMNLP'20。计算语言学协会，线上，38 45 2020 6 – 。[40] 2020 https://www.aclweb.org/anthology/ 。emnlp-demos。Tong Zeng 和 Daniel E Acuna。。利用近似网络结构进行大规模作者姓名消歧。载于计算社会科学国际会议。[41] Baichuan Zhang 和 Mohammad Al Hasan。2017。利用网络嵌入对匿名图进行姓名消歧。载于2017年 ACM信息与知识管理会议论文集。1239–1248。[42] Baichuan Zhang、Murat Dundar、Vachik Dave 和 Mohammad Hasan。2019。用于粒子滤波主动在线姓名消歧的狄利克雷过程高斯混合。载于 JCDL'19。IEEE，269–278。[43] Chuxu Zhang、Chao Huang、Lu Yu、Xiangliang Zhang 和 Nitesh V Chawla。2018。Camel：面向作者标识的内容感知与元路径增强度量学习。载于2018年万维网会议论文集。709–718。[44] Jing Zhang 和 Jie Tang。2021。AMiner 中的姓名消歧。Science China-Information Sciences 64，4（2021），10–1007。[45] Li Zhang、Yong Huang、Qikai Cheng 和 Wei Lu。2020。通过关联开放书目数据库挖掘 PubMed 作者标识符。载于2020年 IEEE 第20届软件质量、可靠性与安全国际会议伴随会议（QRS-C）。IEEE，209–212。[46] Li Zhang、Wei Lu 和 Jinqing Yang。2021。arXiv 预印本 arXiv:2104.01821（2021）。[47] Yutao Zhang、Fanjin Zhang、Peiran Yao 和 Jie Tang。2018。AMiner 中的姓名消歧：聚类、维护与人在回路。载于第24届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集。1002–1011。[48] Zhenyu Zhang、Bowen Yu、Tingwen Liu 和 Dong Wang。2020。有无神经网络条件下作者姓名消歧的强基线方法。载于亚太知识发现与数据挖掘会议。Springer，369–381。[49] Xin Zheng、Pengyu Zhang、Yanjie Cui、Rong Du 和 Yong Zhang。2021。用于作者姓名消歧的双通道异构图网络。Information 12，9（2021），383。

```
# 模块 -1:      Data 正在加载
来自 whoiswho.dataset 导入 LoadData,              随机拆分数据
# 加载数据集的特定版本。
训练集 = LoadData(姓名="v3",     类型="训练集", 分区 =None)
# 将数据拆分为未分配论文和候选作者
未分配论文, 候选作者 = SplitDataRND(训练集, 拆分="时间",
        比率 =0.2)

# 模块 -2: 特征创建
来自 whoiswho.featureGenerator 导入 AdHocFeatures

# 提取默认的 n 维临时                                特征。
pos_特征, neg_特征 = AdHocFeatures(unassigns ,
        候选项,          特征_模式="默认", 负样本 =3)

# 模块 -3:      模型构建
来自whoiswho.loadmodel 导入 ClassficationModels

# 构建一个基础分类模型。
预测器=ClassficationModels（类型="MLP",              集成 =False)
# 自动训练
来自whoiswho.training 导入 AutoTrainRND
预测器 = AutoTrainRND(inputs = (pos_feats , neg_feats),
        预测器 , 训练轮次 =1, bs =1, early_stop =None)

# 模块 -4: 在验证数据上进行评估
# 加载验证数据
未分配项、候选项、真实值 = LoadData("v3",          类型="验证集".
        任务="RND")
# 分配未分配的          论文
分配_结果 = 预测器.predict(unassigns ,              候选项)
# 评估 RND 结果
来自whoiswho.evaluation 导入 RNDeval
加权_精确率, 加权_召回率,                    加权_F1 =
        RNDeval(assign_res , gt)
```

列表 1：基础 RND 算法

A 附录

A.1 WhoIsWho 工具包流程。

图9展示了 WhoIsWho 工具包流程的概览。清单1展示了构建基础 RND 算法的一个简单示例。

A.2 WhoIsWho 基准中的数据组织结构。

截至目前，WhoIsWho 已发布三个版本的数据集，即 WhoIsWho-v1、-v2 和 -v3，并为 IND 任务专门提供了一个数据集 WhoIsWho-v3.1。其中，v1、v2、v3 数据集包含相同的机构，用于 SND 和 RND 任务。下面简要介绍数据组织结构。

A.2.1 WhoIsWho-v1/v2/v3。数据集采用两级字典的格式组织，即姓名-作者-论文，如列表 2 所示。第一级字典的键是作者姓名，值是具有“同名”的作者档案。“同名”一词指的是使用姓名阻塞技术统一姓名的方式 [2, 14], ，例如将姓氏移至名字前，或保留除姓氏以外的所有姓名首字母。例如，“Jing Zhang”的变体包括“Zhang Jing”、“J. Zhang”和“Z. Jing”。作者档案也以字典形式组织，其中键是作者 ID，值是该作者的论文ID。对于每篇论文，我们收集其标题、所有作者的姓名、机构、关键词、摘要、出版年份以及发表场所（会议或期刊）等属性。列表 4 展示了论文ID为“9PgiwDo7”的一个示例。

A.2.2 用于 IND 任务的 WhoIsWho-v3.1。WhoIsWho-v3.1 采用单层字典结构，因为 IND 任务旨在检测并移除每位作者名下的错误论文。字典的键为作者 ID，值为属于该作者的论文，即正常数据，以及人工检测出的错误论文，即离群点。示例见列表3。

A.3 WhoIsWho竞赛的数据划分

我们介绍了如何将 WhoIsWho数据集划分为训练集、验证集和测试集，分别用于竞赛的三项任务，即 SND、RND 和 IND。

从头开始的姓名消歧。SND 任务旨在将同一姓名下的论文划分为不同的组。每组包含同一作者的论文，而不同组中的论文则属于不同作者。因此，我们首先根据特定比率，按作者姓名层级将数据集划分为训练集、验证集和测试集。随后，对于验证集和测试集，我们删除每个姓名下作者与论文之间的作者关系，如清单5所示。研究人员应将属于同一作者的论文正确归入同一组。

实时姓名消歧。 RND任务旨在将新到达的论文分配给现有作者。因此，首先，我们也按照特定比率，以作者姓名为级别将数据集划分为训练集、验证集和测试集。随后，对于验证集和测试集，我们按照发表年份升序对每位作者的论文进行排序。为模拟真实的RND场景，我们将每位作者最新的论文视为新到达的未分配论文，将其余论文视为现有作者档案，如列表6所示。我们还向未分配论文中添加了若干NIL论文，即无法分配给任何现有作者档案的论文。研究人员不仅需要将论文正确分配给相应作者，还需要识别NIL论文。

错误分配检测。 IND 任务旨在检测并移除每位作者名下的错误论文。具体而言，我们按如下方式构建 IND 任务的数据集：1）如表 1 所示，整体数据标注流程包含一个“清理”步骤，标注人员在此步骤中移除或拆分明显不属于相关作者的论文。随后的“验证”步骤允许标注人员对那些更难识别的错误分配论文执行同样的“清理”操作。这两个阶段为 IND 任务提供了足够数量的待检测错误分配论文。2）一些作者会手动维护自己的个人资料，例如添加新论文或移除不属于自己的论文。我们还收集了这些被移除的论文，作为 IND 任务的检测目标。因此，我们按作者组划分训练集、验证集和测试集。

```
{
    "官华_都": {        # 作者姓名，
        "zsOOUZxZ": [# 作者 ID        ，
            "QDMcmF8V" ，# 论文ID      ，
            "9PgiwDo7"
        ]
    },
    "bin_yu":    {
        "HoH18DsE": [
            "VMYs96sn" ，
            "YT4XzThC" ，
            "S3RARClD" ，
            "wM8dXlKT"
        ],
        "WYZVZfO0":   [
            "OTvYjfnt",
            "EzzruFin"
        ],
        "9BMlVP0u":   [
            "brlzUqnH" ，
            "HWAfXDPx" ，
            "RZPhHOMm" ，
            "vebukM2n"
        ]
    }
}
```

列表 2：WhoIsWho-v1/v2/v3 的数据组织结构

```
{
    "HoH18DsE": { # 作者      标识符，
        "姓名":   "xxx"，# 姓名      作者的，
        "正常_数据": [# 论文          属于     to  作者，
        "VMYs96sn" ，
        "YT4XzThC" ，
        "S3RARClD"，
        "wM8dXlKT"
        ],
        "离群点": [# 错误分配的论文          分配      to  the
    作者
        "OTvYjfnt",
        "EzzruFin"
        ]
    },
}
```

列表3：WhoIsWho-v3.1 的数据组织结构

```
{
    "9PgiwDo7": {
        "id": "9PgiwDo7  ",
        "标题":"预应变 TiNi 纤维嵌入金属基体中的受限相变
    预应变 TiNi 纤维嵌入金属基体中
    智能复合材料",
        "摘要": "TiNi 纤维的逆马氏体
    相变嵌入金属
     基智能复合材料中          has  的相变已得到研究。结果
     表明，在                 温度和
     回复      应力，      逆马氏体
    TiNi 纤维的相变可以                    be  分为
     two  具有不同                 动力学特征的部分
    :  自适应          的逆向转变
    马氏体              与定向马氏体的
    马氏体。        两者之间的关系        马氏体分数
    与温度之间的关系已计算。”                    ，
        “关键词”:      [
            “null”
        ],
        "作者":       [
            {
                "姓名":   "Yanjun    郑".
                "org":  "大连      大学          of
    技术（大连              大学          大连理工大学，大连
     大学     技术），大连，中国"
            },
            {
                "姓名":   "Lishan   Cui",
                "机构":  "中国     大学          石油大学
    - 北京（中国          大学          石油大学，
    大学          of  石油），北京，中国"
            },
            {
                "姓名":   "Dan  Zhu",
                “机构”:  “中国     大学          石油大学
    - 北京（中国          大学          石油，
    大学          of  石油），北京，中国”
            },
            {
```

```
{
    "bin_yu": {
        [
            "VMYs96sn" ，
            "YT4XzThC",
            "S3RARClD" ，
            "wM8dXlKT"
            "OTvYjfnt",
            "EzzruFin" ，
            "brlzUqnH" ，
            "HWAfXDPx" ，
            "RZPhHOMm" ，
            "vebukM2n"
        ]
    }
}
```

清单5：RND任务中验证集/测试集的数据组织结构

RND任务

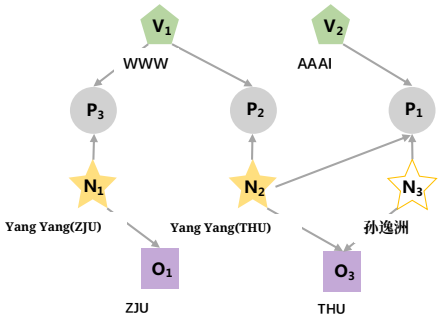
```
# 现有        作者      档案。
{
    "HoH18DsE": [
        "VMYs96sn",
        "YT4XzThC" ，
        "S3RARClD" ，
        "wM8dXlKT"
    ],
    "9BMlVP0u":     [
        "brlzUqnH   ",
        "HWAfXDPx"  ，
        "RZPhHOMm"  ，
    ]
}

# 新到的              未分配          论文。
{
    [
    "vebukM2n -1",          # "-1"  表示    the  2nd   （作者     索引
     从 0 开始）作者应当                    消歧。
    "RZPhHOMm -3  "
    ]
}
```

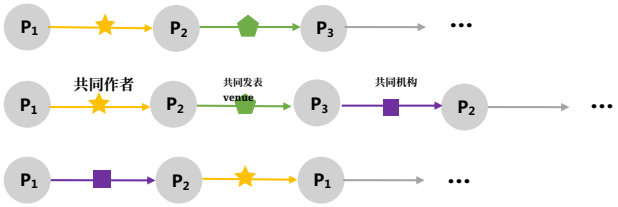
图 6：WhoIsWho-v1/v2/v3 的数据组织结构

A.4 运行环境

我们使用 PyTorch 实现所有实验模型，并在一台企业级 Linux 服务器上运行代码。该服务器配备 40 核 Intel(R) Xeon(R) CPU（E5-2640 v4 @ 2.40GHz，内存 252G）和 1 个 NVIDIA Tesla V100 GPU 核心（内存 32G）。



(a). 关系图



(b). 元路径

图10：SND-all 方法采用的异构关系图和元路径。

A.5 SND-all 的实现细节

这里介绍 SND-all 方法的技术细节，包括软语义特征、异构关系图特征以及组合模式。

A.5.1 软语义特征。我们从 WhoIsWho 基准中提取所有论文属性，包括标题、摘要、发表场所、关键词、年份和作者所属机构，并使用通过 genism 实现的 word2vec 训练词嵌入模型，嵌入维度设为 100。超参数定义如下：min_count=2、window=5、negative=5。然后，对于每篇论文，我们将标题、关键词和待消歧作者所属机构的词嵌入取平均，作为语义特征，得到最终的论文嵌入。具体而言，标题和关键词字段直接取自论文属性。对于机构，我们提取待消歧目标作者所属的机构。

A.5.2 关系语义特征。对于每个目标姓名，我们在论文之间构建一个异构网络，其中包含三种关系类型：共同作者、共同机构和共同发表 venue。采用的元路径如图10所示。异构边的权重由共现次数决定，即共同作者关系中的共同作者人数、共同机构关系中机构里的共现词数量，以及共同发表 venue 关系中会议/期刊里的共现词数量。我们将论文通过共同发表 venue 边游走至相邻论文的随机概率设为0.1，以减小发表场所中常见词噪声的影响。游走长度设为20，每个节点的随机游走次数设为5。随后，我们保存每篇论文对应的路径集合，并使用 word2vec进行训练。对于缺少邻居的孤立论文，我们将其保留以供后续匹配处理。

A.5.3 特征组合。最后，我们分别根据论文的语义与关系嵌入计算论文之间的成对距离。随后，我们将两个相似度矩阵直接相加，得到组合相似度矩阵。为获得聚类结果，我们采用 DBSCAN 生成簇，其中 db_eps 设为 0.2，db_min 设为 4。对于无法归入任何组的离群点，我们采用基于规则的方法进行后续匹配。具体而言，我们对共同作者的姓名进行字符匹配，作为 CoA；对两篇论文的标题和关键词进行字符匹配，作为 CoW；并采用 Tanimoto 距离计算机构和发表场所的相似度，分别作为 CoO 和 CoV。CoA、CoW、CoO 和 CoV 的权重分别设为 {1.5, 0.33, 1.0, 1.0}。随后，如果分数高于预设阈值（在我们的方法中为 1.5），我们就将论文分配到相应的组中。

A.6 RND-all 的实现细节

这里介绍 RND-all方法的技术细节，该方法由软语义特征、特定语义特征、自我图和组合模式组成。

A.6.1 软语义特征。同样，我们像 SND-all 一样，利用 OAGBERT [19], 获取论文的软语义嵌入，并据此计算目标论文与候选作者的每篇论文之间的软语义相似度。

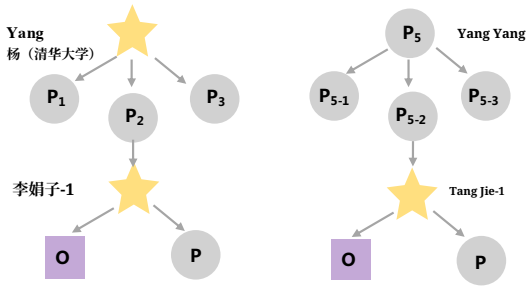
A.6.2 临时语义特征。表5详细列出了每篇未署名论文与候选作者配对之间的36维人工设计特征。

A.6.3 关系自我图特征。我们收集历史书目信息，为目标论文和候选作者构建异构自我图，其中作者、论文和机构作为节点。如图11所示，对于作者这一中心节点，我们首先使用 OAGBERT 获取论文嵌入，然后对属于该作者的所有论文嵌入取平均，得到作者嵌入。之后，通过训练图注意力网络（GAT）获得更新后的论文或作者嵌入，并据此计算中心论文与中心作者之间的关系相似度。

A.6.4 特征组合。 对于软语义特征和关系自我图特征，我们采用 [3] 中使用的 RBF核函数，基于目标论文与候选作者之间的相似度分数提取 41 维聚合特征。对于特定语义特征，我们直接采用 36 维手工构造特征。最后，我们简单地将三个特征模态的聚合特征拼接起来，得到目标论文与候选作者之间的 118 维特征，并采用集成 GBDT 模型进行预测。

A.7 基线方法的实现细节

下面详细介绍论文中使用的姓名消歧基线方法。



（a）. 作者的自我中心图　　（b）. 论文的自我中心图

图11：RND-all方法中目标论文与候选作者构建的自我中心图。

表5

36维人工设计特征的详细定义。 p：目标论文， a： p中的目标作者， c：候选人。

编号.特征描述
1 a的合著者在 c中的TF-IDF分数
2 a的合著者在 c 中的TF-IDF分数乘以 c中的共现次数
3 a的合著者在 p的作者姓名中的比率
4 a的合著者在 c的作者姓名中的比率
5 a的标题共同部分在 c中的TF-IDF分数
6 a的标题共同部分在 c 中的TF-IDF分数乘以 c中的共现次数
7 a的标题共同部分在 a的标题中的比率
8 a的标题共同部分在 c的标题中的比率
9 a的标题与 c的标题之间的最大杰卡德相似度
10 a的标题与 c的标题之间的平均杰卡德相似度
11 a的标题与 c的标题之间的最大Jaro–Winkler相似度
12 a的标题与 c的标题之间的平均Jaro–Winkler相似度
13 a的发表 venue 共同部分在 c中的TF-IDF分数
14 a的发表 venue 共同部分在 c 中的TF-IDF分数乘以 c中的共现次数
15 a的发表 venue 共同部分在 a的发表场所中的比率
16 a的发表 venue 共同部分在 c的发表场所中的比率
17 a的发表场所与 c的发表场所之间的最大杰卡德相似度
18 a的发表场所与 c的发表场所之间的平均杰卡德相似度
19 a的发表场所与 c的发表场所之间的最大Jaro–Winkler相似度
20 a的发表场所与 c的发表场所之间的平均Jaro–Winkler相似度
21 a的机构共同部分在 c中的TF-IDF分数
22 a的机构共同部分在 c 中的TF-IDF分数乘以 c中的共现次数
23 a的机构共同部分在 a的机构中的比率
24 a的机构共同部分在 c的机构中的比率
25 a的机构与 c的机构之间的最大杰卡德相似度
26 a的机构与 c的机构之间的平均杰卡德相似度
27 a的机构与 c的机构之间的最大Jaro–Winkler相似度
28 a的机构与 c的机构之间的平均Jaro–Winkler相似度
29 a的关键词共同部分在 c中的TF-IDF分数
30 a的关键词共同部分在 c 中的TF-IDF分数乘以 c中的共现次数
31 a的关键词共同部分在 a的关键词中的比率
32 a的关键词共同部分在 c的关键词中的比率
33 a的关键词与 c的关键词之间的最大杰卡德相似度
34 a的关键词与 c的关键词之间的平均杰卡德相似度
35 a的关键词与 c的关键词之间的最大Jaro–Winkler相似度
36 a的关键词与 c的关键词之间的平均Jaro–Winkler相似度

A.7.1 SND 任务的基线方法。**G/L-Emb**。这是一种利用合著连接构建论文-论文图的方法。该流程首先对论文中的所有词元对应的 Word2Vec 嵌入进行加权平均，从而生成初始论文嵌入。接下来，该方法先在全局论文-论文网络上进行学习，再通过图自动编码将这些嵌入调整到针对每位作者的局部论文-论文网络。最后，采用层次凝聚聚类算法（HAC）将这些论文划分为不同的组。**LAND**。这是一种用于构建异构学术知识图谱（KG）的方法，其中涵盖论文、作者、发表场所、所属机构等多种实体。这些知识图谱还包含各种关系，例如合著关系和发表场所关系。实体输入嵌入使用 BERT 模型初始化，随后应用知识图谱嵌入技术，生成论文和作者嵌入。最后，利用层次凝聚聚类（HAC）方法，根据嵌入对这些实体进行分组。**IUAD**。这是一种通过将论文视为节点，并在两篇论文拥有同一作者时在它们之间创建边来构建协作网络的方法。它采用概率生成方法，判断协作网络中的两篇论文是否属于同一位作者，目标是准确重建完整的协作网络。概率模型训练完成后，将其用于执行 SND（监督式姓名消歧）任务，以提升整体消歧效果。**SND-all**。这是一种基于竞赛获胜方法的实证研究提出的基线方法。它首先使用标题、关键词和机构作为软语义特征，估计待消歧论文之间的语义相关性。对于每篇论文，SND-all 通过 Word2Vec 将这些特征映射到相应的词嵌入，并对其取平均以生成论文嵌入。随后，根据这些语义嵌入计算论文之间的余弦相似度。此外，SND-all 构建了一个异构网络，其中包含论文之间的三种关系边：共同作者、共同机构和共同举办地。该方法使用 metapath2vec 生成论文的关系嵌入，并根据这些嵌入计算关系相似度分数。最后，SND-all 将多模态相似度结合起来，计算论文之间的总体相似度。DBSCAN 根据这些相似度得出最终聚类结果。**竞赛获胜方法**：采用与 SND-all 类似的流程。不过，它使用所有论文属性来估计论文之间的语义相关性；事实证明，与 SND-all 方法采用标题、关键词和机构等少数信息丰富的属性相比，这种做法效果较差。此外，在估计论文之间的关系相似度时，竞赛

表6：主流人工标注姓名消歧数据集的统计信息。姓名数较少、作者和论文数较多，表明数据集的歧义度更高。

数据集	姓名数	作者数	论文数	来源
Song-PubMed	36	385	2,875	PubMed（生物医学）
GS-MEDLINE	-	-	3,756	PubMed（生物医学）
Han-DBLP	14	479	8,453	DBLP（计算机科学）
Qian-DBLP	680	1,201	6,783	DBLP（多个领域）
Tang-AMiner	110	1782	8386	AMiner (通用领域)
SCAD-zbMATH	2919	2946	33,810	zbMATH (数学)
Zhang-AMiner	100	12,798	70,258	AMiner（通用领域）
INSPIRE	**12,458**	36,340	360,066	INSPIRE（物理学）
WhoIsWho	2,495	**72,609**	**1,102,249**	AMiner（通用领域）

获胜方法仅考虑共同作者和共同机构关系边，因此与SND-all 相比，所采用的视角更为有限。

*A.7.2 RND*任务的基线方法。**IUAD**也用于RND任务，通过重建新到论文与现有作者之间的协作网络来完成。**CONNA**是一种自底向上、基于交互关系的模型。为了确定目标论文与候选作者之间的相似度，它首先为目标论文和候选作者的每篇论文对应的属性词元嵌入构建基本交互关系。随后，将不同的属性矩阵聚合起来，形成论文级交互关系。具体而言，CONNA使用所有论文属性，将作者姓名视为一个字段，将其他所有属性视为另一个字段。最后，使用学习排序技术，根据聚合后的论文级矩阵，对目标论文与所有候选作者之间的一致程度进行评分。**CONNA+临时**方法是一种组合方法，将手工构造特征整合到CONNA框架中。为确保公平比较，该方法采用了RND-all中使用的相同临时特征。**RND-all**是根据消融研究结果提出的另一种方法。它使用第4.1节所述的软语义特征和临时语义特征。此外，为了纳入关系特征，RND-all为目标论文和所有候选作者构建异构自我图，如附录图11所示。该方法使用图注意力网络（GAT）获取目标论文和候选作者的融合中心节点嵌入。通过计算中心节点嵌入之间的余弦相似度，确定目标论文与各候选作者之间的关系相关性。最后，RND-all结合软语义特征、临时语义特征和所提出的关系特征进行预测。**竞赛获胜方法**仅使用软语义特征来估计目标论文与各候选作者之间的一致程度和临时语义特征。该方法在构建语义特征时也使用了所有论文属性；事实证明，与RND-all方法采用少数信息丰富的属性相比，这种做法效果较差。

A.8 arXiv 每日论文消歧的在线部署

我们已在 AMiner 上部署了所提出的 RND-all 方法，用于对 arXiv.org 的每日论文进行消歧。具体来说，对于论文中每个需要消歧的姓名，我们没有采用姓名阻塞策略（即将姓氏移至首位，或保留除姓氏外的所有姓名首字母），而是采用 Elastic-Search [11] 进行在线模糊搜索。最后，我们使用 RND-all 估算每位候选作者与目标论文之间的相似度。为处理不存在正确作者的 NIL 案例，我们预先设定了一个阈值，并将得分超过该阈值的最高分候选作者作为 AMiner 上的正确作者返回。

A.9 数据集统计。

详细的数据统计信息如图6所示。

[11]https://www.elastic.co

A.10 交互式标注工具

图12展示了所设计的交互式标注工具的框架，该工具由两个主要部分组成，即标注面板和信息面板。

标注面板。 前3个区域构成了由三层堆叠结构组成的圆环。具体而言，外层即区域“1”展示了收集到的、尚未分配且目标作者姓名为“Andrea Rossi”的论文。中间层中的每个区块，即区域“2”，代表一位名为“Andrea Rossi”的作者。为帮助标注人员掌握论文之间的全局关系，我们采用聚类方法，将每位作者的论文划分为若干组，如区域“3”所示。每组包含具有相似属性的论文，例如发表在同一发表场所或由相同作者共同撰写的论文等。点击区域“2”中的作者后，我们可以看到论文之间的交叉关联（虚线）和内部关联（实线）。标注人员可以通过区域“6”自由选择不同的属性。随后，标注人员可以通过区域“5”执行操作。

Information Panel. 区域“7”“8”和“9”提供了所选论文和作者的全面信息，帮助标注人员准确执行操作。其中，区域“7”展示所选作者之间的档案比较。区域“8”则展示所选作者之间更详细的比较，例如合著者、所属机构和关键词。最后，区域“9”提供所选论文的完整信息。

总的来说，交互式标注工具不仅提供便捷的原子操作来提高标注人员的效率，还准备了全面的信息来支持他们做出决策。借助这一有效的可视化工具，我们计划在未来为 WhoIsWho 标注并更新更多数据集。

收稿日期：2007年2月20日；修订日期：2009年3月12日；接受日期：2009年6月5日

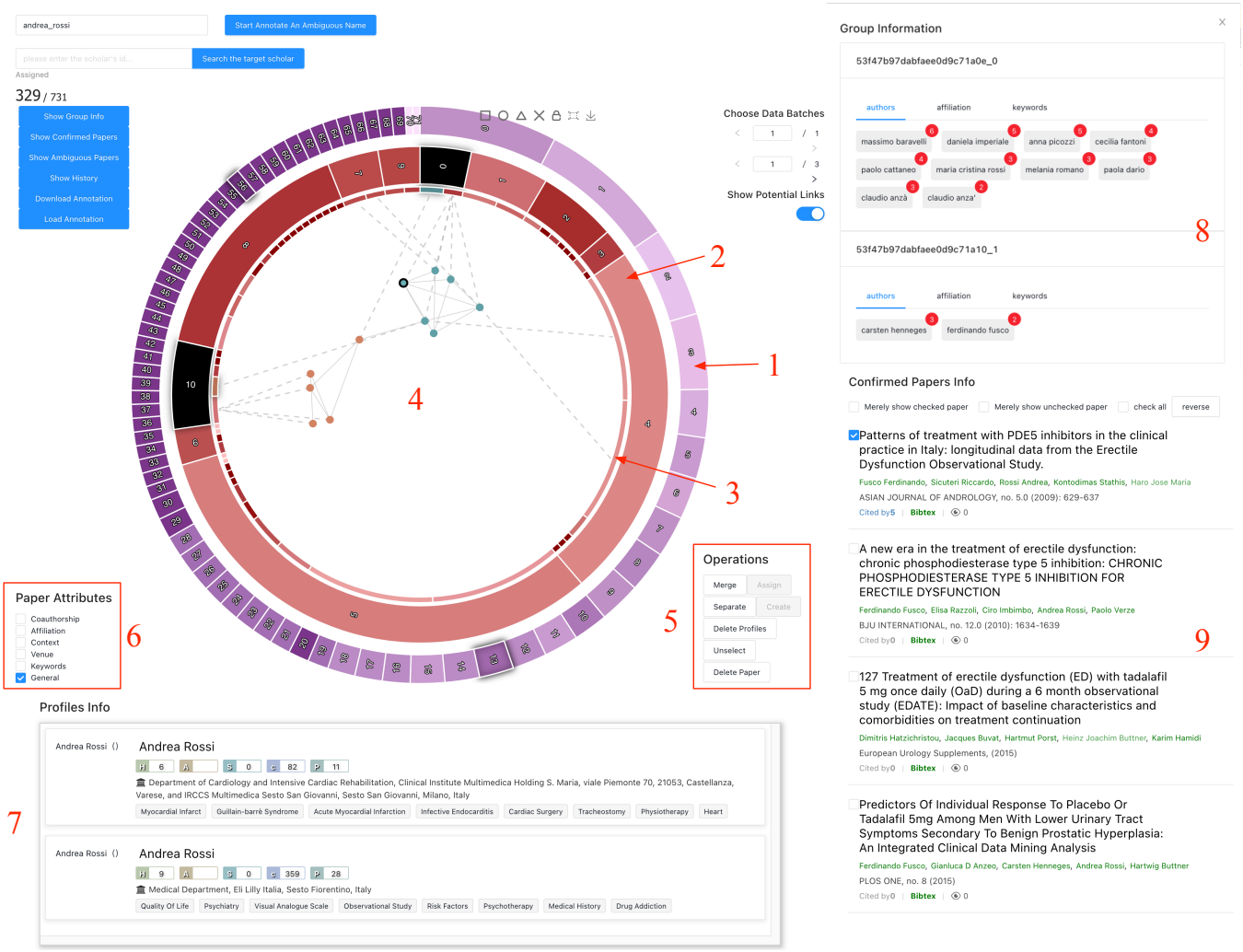


图12：使用作者姓名“Andrea Rossi”进行作者标注的简化标注示例。