

# 多特征融合的英文科技文献增量式 人名消歧应用研究\*

阮光册<sup>1</sup> 涂世文<sup>1</sup> 田欣<sup>2</sup> 张莉<sup>2</sup>

(1. 华东师范大学经济与管理学部信息管理系 上海 200241; 2. 上海科技发展有限公司 上海 200235)

**摘要:**[目的/意义] 英文作者重名现象十分普遍,为解决科技文献增量式人名消歧问题,以提高学术检索平台作者检索的精度。[方法/过程] 提出一种融合文献外部基本特征和内部语义特征的人名消歧方法,解决新增英文学术文献作者归属的问题。首先,提取学术文献中人名消歧所需的元数据字段,采用 BERT 模型对元数据中包含语义信息的文本内容进行向量表示;随后,将融合多特征的数据输入 XGBoost,完成机器学习;最后,用学习好的模型实现新增文献的作者分配。[结果/结论] 通过实验对比,该方法表现出较好的效果, F1 取得了 95.6% 的分值。

**关键词:** 人名消歧; 科技文献; 多特征融合; BERT; XGBoost

中图分类号: G250

文献标识码: A

文章编号: 1002-1965(2021)09-0147-07

**引用格式:** 阮光册,涂世文,田欣,等. 多特征融合的英文科技文献增量式人名消歧应用研究[J]. 情报杂志, 2021, 40(9): 147-153.

## Application Research of Incremental Person Name Disambiguation in English Scientific and Technological Literature Based on Multi Feature Fusion

Ruan Guangce<sup>1</sup> Tu Shiwen<sup>1</sup> Tian Xin<sup>2</sup> Zhang Li<sup>2</sup>

(1. Department of Information Management in Faculty of Economics and Management,

East China Normal University, Shanghai 200241;

2. Shanghai Technology Development Co., Ltd, Shanghai 200235)

**Abstract:** [Purpose/Significance] Since the phenomenon of duplicate names of English authors is very common, in order to solve the problem of incremental name disambiguation in scientific and technological literature, and improve the accuracy of author retrieval in academic retrieval platform. [Method/Process] This paper proposes a method of name disambiguation, which combines the external basic features and internal semantic features of the literature, to solve the problem of the author's attribution of the newly added English academic literature. Firstly, this paper extracts the metadata fields needed for person name disambiguation in academic literature, and uses the Bert model to represent the text content containing semantic information in the metadata vector; then, the data fused with multiple features is input into XGBoost to complete machine learning; finally, the author assignment of new literature is realized by using the learned model. [Result/Conclusion] Through the experimental comparison, this method shows good results, F1 achieved 95.6% of the score.

**Key words:** person name disambiguation; technological literature; Multi feature fusion; BERT; XGBoost

## 0 引言

计量、科学评价等领域研究的基础,也是情报学研究的重要问题。随着世界科学研究的蓬勃发展,各类学术文献数量正以惊人的速度增长。STM(Scientific Tech-

准确获取指定作者的发文信息是文献计量、科学

收稿日期: 2020-12-27

修回日期: 2021-02-08

基金项目: 上海市经信委项目“上海人工智能公共研发资源图谱”(编码: XX-RGZN-01-19-5037)。

作者简介: 阮光册(ORCID: 0000-0001-8685-5234),男,1976年生,博士,副教授,硕士生导师,研究方向: 信息分析、文本挖掘;涂世文,男,1995年生,硕士研究生,研究方向: 数据挖掘;田欣,女,1990年生,硕士,研究方向: 生物信息学;张莉,女,1987年生,博士,研究方向: 控制理论与控制工程。

nical and Medical,简称 STM,国际科学、技术和医学出版商协会)报告显示<sup>[1]</sup>,2018年,全球范围内的研究人员数量达到710万,且每年以3%~4%的速度持续增长。在学术文献数据库中,作者的名称属性通常是识别和区分学术文献实体最常用的标识符,但相较于模糊匹配和逻辑检索功能在文献数据库中的广泛应用,对英文数据库的作者检索则受到作者同名、重名以及人名处理标准不一致等问题的困扰,影响了检索结果的精度。

人名消歧属于自然语言处理的研究范畴,科技文献的人名消歧问题存在于各种语言中。就科技文献的人名消歧来说,其目的是准确获取指定作者的发文信息,其核心问题是判断出现在不同文献中相同的作者名是否指向同一个人。解决该问题相对简单的方法是为每一个科研工作者提供一个唯一的标识符,如ORCID,但由于涉及隐私政策等原因,在网络化和数字图书馆快速发展的今天,采用唯一标识码的方法已无法有效解决海量科技文献数据增长的问题。利用自动化方式进行人名消歧是目前研究的方向,主要的做法是采用某种规则或者算法,将同名作者加以区分,将无歧义的信息呈现给用户<sup>[2]</sup>。然而,在实际研究中,由于作者英文署名存在多种形式使得构建的模型排歧精度不高、可扩展性欠佳等诸多问题有待进一步解决。

为此,本文以学术文献增量式人名消歧为研究对象,解决新增学术文献论文归属问题。首先提取学术文献中的多种特征数据,借助语言模型BERT进行语义特征抽取,将融合了文献外部特征和文本内部语义特征的数据输入XGBoost集成模型进行相似度匹配,通过计算,将新增文献分配给同名作者中相似度最高的作者,完成学术文献的增量消歧。最后,本文使用DBLP学术搜索平台的学术文献数据进行实验,获得了较好的实验结果。

## 1 研究现状

科技文献作者同名消歧问题属于命名实体消歧的范畴<sup>[3]</sup>。自从Bagga和Baldwin首次提出跨文本的同指消歧(Co-Reference)<sup>[4]</sup>之后,人名共指的研究逐渐引起了学界的关注。2001年,数字图书馆联合会议就作者消歧问题展开讨论,研究解决数字参考文献检索系统中作者同名问题。

作者同名消歧本质上是一个聚类或分类问题,一般包括特征抽取、相似度计算、消歧处理等步骤。

基于文献特征的人名消歧是最早被使用的研究方法,利用学术文献的元数据,将对同名作者有着较大区分度的特征提取出来,然后采用特征组合和构造的方式,选择并保留有效特征,借助模型实现对同名作者的

分类,进而实现消歧<sup>[2]</sup>。在特征提取时,作者的个人信息或论文的题录信息是常用的消歧特征。实践研究表明,选择有效的特征能准确辨识作者的真实情况。如使用文献的合作者信息进行人名消歧<sup>[5]</sup>,或将作者和论文题录的多个概念特征进行组合,作为消歧特征组,通过构建相似度矩阵,借助聚类算法进行消歧<sup>[6]</sup>。除了学术文献基本元数据外,有学者<sup>[7]</sup>将论文的主题作为消歧特征提取,融合合作者信息、姓名关联信息等多特征,实现人名消歧。

根据对所使用特征处理方式的不同,人名消歧的方法又可以进一步划分为无监督的消歧方法、有监督的消歧方法和半监督的消歧方法。无监督的人名消歧基本思路为:将所选特征转化为一组数值,通过计算,将相似度比对结果满足阈值要求的论文归属为同一作者。采用的算法包括:K均值算法<sup>[8]</sup>、基于密度的聚类算法<sup>[9]</sup>、凝聚层次聚类算法(HAC)<sup>[10]</sup>以及各种改进算法等等。由于无监督方法使用非标注的文献数据,借助选取的特征值计算各文献间的相似度,面对大量文献时,运算效率会有所下降,且聚类文献与现实中作者的对应关系也存在问题。虽有学者提出了多阶段的聚类策略<sup>[11]</sup>,一定程度上提高了无监督聚类的准确率,但实体对应关系问题依然极大地限制了其使用场景。有监督的方法利用标注好的训练数据集来学习分类模型,其做法为:总结已知作者发文、所属单位等特征,依据这些特征对新出现论文进行判断,决定归属。文献<sup>[12]</sup>采用随机森林和DBSCAN聚类的方法,在USPTO专利数据集上进行实验,获得了较好的人名消歧效果。有监督方法效率与精度较高,不足的地方在于需要大量标注好的样本,有时需要专业人员耗费大量的时间对数据进行标注,限制了其在大型数据库中的应用。结合非监督和监督算法的优点,研究人员开始尝试采用将少量标注数据与大量无标注数据相结合<sup>[13]</sup>,通过训练模型,进行人名消歧。然而,半监督的方法需要人工定义规则,以实现数据标注<sup>[14]</sup>,在处理大规模数据集的人名消歧任务时,仍存在不足。

考虑到学术文献数据中可利用的信息有限,研究者尝试整合外部的资源和知识来达到数据增强的效果。该类方法的一般思路是:结合外部公开的资源和知识库,通过创建新的规则和类别,将待消歧的姓名与现实世界中人物信息中区分度较强且准确的社会属性建立联系,从而获得更丰富的人物特征,并基于这些社会属性进行分类,从而实现消歧的目的。文献<sup>[15]</sup>通过获取包含作者文章的Web网页,判断两个待消歧作者的文献是否同时出现在一篇Web文档中,从而区分人名。然而,借助外部资源获取额外的信息,客观上会影响文献检索的效率,此外如何避免外部信息所带来的

噪音,也是该方法面临的难题。

目前,一些研究开始探讨对学术论文所包含的语义特征进行计算,对同一作者的研究成果进行辨识,从而实现有著者姓名的消歧。如文献[16]利用语义分析技术对机构知识库进行作者人名消歧,而文献[17]则使用生物神经网络层级时序记忆(Hierarchical Temporal Memory, HTM)对论文摘要进行信息表示,实现作者人名消歧。

综上所述,对于学术文献作者姓名消歧问题,现有的多种研究方法均存在各自的优势与不足。本文以增量式人名消歧场景作为研究对象,将BERT预训练语言模型引入作者人名消歧的研究,借助深度学习强大的语义特征提取和表示能力,解决英文文献作者消歧的问题。本文旨在探索深度学习方法在解决姓名歧义问题时的可用性并评价其消歧效果。

## 2 方法设计

人名增量消歧本质上是一个分类问题。传统的基于机器学习的分类方法主要是通过将文本表示为特征向量,利用特征对文本进行降维,选择算法模型(如:SVM,朴素贝叶斯等)实现分类。这种基于词袋模型的分​​类方法对特征工程的依赖度较高,在复杂任务及大数据量的情况下,从原始数据中自动学习抽象的、高层次的全局特征的能力不强,使得分类模型的泛化能力较差。相对于机器学习方法,深度学习方法能够从大规模无标注语料中学习词的语义和句法信息,通过组合多个非线性模型,将文本数据转化为更高层次的知识表示<sup>[18]</sup>,在无需大量特征工程的情况下,从大数据中自动学习文本特征并刻画出文本的内在信息,提高分类的效果。

在学术文献不断动态增长的情景下,无法依靠人工方式建立分类特征。基于此,本文以深度学习框架为核心,在较少的人工干预情况下,利用词向量方法对特定应用情境下的语义特征进行提取,快速高效的从历史数据中学习高层次知识表达,实现科技文献增量式人名消歧方法。

**2.1 科技文献的元数据特征分析** 科技文献的元数据因类型不同而异,文献[19]按照是否涉及文献内容将科技文献元数据分为文献的内部特征和外部特征。基于这一分类方式,本文选择进行人名消歧的文献元数据如表1所示。

依据文献[19],本文选择作者、合作者、作者单位、期刊名、发表时间等元数据作为科技文献外部特征进行消歧计算。这些特征不涉及(或较少涉及)科技文献的内容,属于文献的外部标识,在实践应用中,本文采用特征相似性匹配的分析方法,通过阈值,判断学

者与新增科技文献之间的归属问题。相对应的,关键词、标题和摘要等内部特征,其具有较强的文本语义属性,本文采用语义表征学习模型BERT对其进行计算,构建蕴含文献语义信息的特征向量,识别学者与待归属文献之间的语义关联。

表1 用于作者名消歧的文献元数据

文献元数据	对应文献特征	特征分类	特征使用说明
作者	科研人员	外部特征	基本特征的相似性 匹配分析
合作者	合著关系		
作者单位	科研机构		
期刊名	来源特征		
发表时间	时间		
关键词		内部语义特征	基于BERT进行 语义向量转换
标题	文本特征		
摘要			

**2.2 融合多特征增量式人名消歧方法设计** 学术检索平台的科技文献数量庞大,数据处于增量式更新状态,且新增文献的规模以及更新时间无法确定。在数据不断动态增长的情景下,如果仍旧采用全局人名消歧的方法,不仅聚类计算的时间复杂度非常高,同时聚类也将耗费大量的计算资源,为此,在性能和速度上往往让人难以接受。

在实践应用过程中,科技文献作者人名增量消歧的过程,是在已经拥有一批消歧文献数据基础上进行的,其核心思路是快速、准确的将新增学术文献分配给已有作者。传统上,图书馆或科研管理机构在进行作者人名规范时,采用将人名与机构名相结合的方式进​​行作者分配,但受到机构名变更、简称重名等问题的干扰,分配结果仍需要大量的人工进行识别,使得这种方法不仅费时,准确率也无法保障。

为实现增量式人名消歧,本文首先构建了两个文献集:现有作者档案文献集和新增学术文献集。实验步骤为:通过机器学习,从现有作者档案文献集中,计算出“作者-文献”的特征向量模型,然后利用该模型对新增学术文献进行匹配计算,将符合阈值的文献分配给已有作者。如果新增文献的作者不在现有作者档案文献集中,则新建其档案,加入现有作者档案文献集。

本文提出的融合文献内容外部特征的人名增量式消歧方法实现流程如图1所示。

由图1所示,本文的方法分成两大过程,即:特征学习和增量式消歧。其中,特征学习是本文方法的核心,借助机器学习和深度学习的方法,从现有作者档案文献集中学习一个特征向量模型,步骤为:构建已明确作者归属的“作者-文献”集,首先,生成正负样本集,其中,正样本集为正确归属文献的数据集合,负样本为



非正确文献归属的数据集合。对于负样本集,本文以随机方式,从现有已明确作者归属的文献集中选取同名作者,并将作者对应的归属文献进行随机的错乱重排;随后,抽取正负样本集中文献的外部特征(见表

1),并对文献内部特征采用 BERT 模型计算向量;最后,将获得的文献外部特征与 BERT 获得向量特征融合为一个特征序列,输入 XGBoost 模型进行训练。

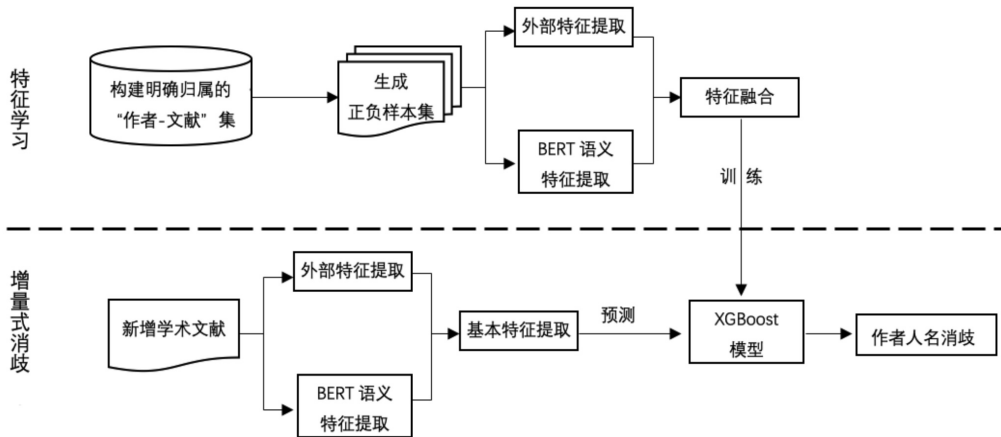


图 1 增量式人名消歧方法流程图

对于增量式消歧部分,首先提取新增文献的外部特征,并采用 BERT 模型对文献内部特征进行向量计算,生成一个新增文献特征序列,借助 XGBoost 训练好的模型对新增学术文献的特征序列进行分类计算,预测作者归属,将新增文献分配给各种特征维度上最为相似的作者,进而完成增量式人名消歧。

**2.3 学术文献内部语义特征提取方法** 人名消歧方法常用论文的标题、关键词以及论文合作者等作为文献作者的研究方向,进而实现人名消歧<sup>[17]</sup>。关键词和标题虽然可以反映论文的主题信息,但受限于关键词的规范性以及主题粒度大小不一的问题,仍然无法完整地表达论文的核心思想。目前,一些研究已经开始使用摘要作为消歧特征之一,具体做法是通过对摘要进行分词和去停用词等预处理后,抽取特征词描述文献的主题,其本质是将长文本信息转换为词的形式<sup>[8-9]</sup>,一定程度上损失了语义信息。为此,本文选择 BERT 语言模型,借助该模型对句子级别文本的语义表示能力,最大限度的保留摘要文本的上下文语义信息。

BERT(Bidirectional Encoder Representations from Transformers)模型<sup>[20]</sup>是 Google AI 团队在 2018 年开源的自然语言处理模型,该模型的主要特点是利用两个方向的上下文信息获得文本深层次的语义,是一种新的自然语言文本表征方法。相对于 Word2Vec 等其它广泛使用的词向量技术,BERT 模型通过遮蔽语言模型(Masked Language Model)和下一句预测两个预训练任务,在大规模的语料集上进行无监督的预训练,再以预训练模型为基础,通过模型微调(Fine-tuning)实现下游文本语义分析任务。BERT 预训练模型在无监督环境下,能够学习到语料库中的语言先验知识,可

以在标注语料稀缺的场景下完成训练任务。在预训练过程中,BERT 结合了语料库中的语言先验知识,使得模型在文本特征的提取和表示方面明显具备优势。在解决一词多义问题方面,BERT 模型通过词汇间上下文关系,记录了文献中词汇的语义信息,可以较好的解决不同语境下的词义问题。

**2.4 基于 XGBoost 的文献匹配** 增量消歧可以认为是一个新增文献与已有作者的匹配问题,其本质上是一种分类问题。为了实现新增文献与已有作者的匹配,首先需要提取已经明确了文献与作者匹配关系的数据集的特征,借助机器学习方法,构建模型;随后,提取新增文献的特征,并将其输入已构建好的模型中,通过计算,实现新增文献与已有作者的匹配。

本文选择极端梯度提升算法 XGBoost<sup>[21]</sup>进行特征计算,并构建模型。XGBoost 是基于 CART 回归树的一种 boosting 集成算法,其核心思想是通过建立多棵回归树,使样本预测尽可能接近样本的真实值,在实践中,算法具有一定的泛化能力。XGBoost 算法在训练大数量时,可以通过同层节点的并行化计算方式,提高计算效率。

科技文献元数据会有一些缺失值,如摘要和关键词等,这就使得数据具有稀疏性。XGBoost 模型在训练数据时,首先对没有缺失值的数据进行分裂,然后计算缺失值最佳的分裂方案,这使得该模型对缺失值并不敏感,具有较好地处理稀疏型数据的能力。

为获得更好的模型预测能力,本文在模型训练时采用交叉验证的形式提高消歧的效果。

**2.5 方法评估** 本文使用加权 F1 值(weighted f1-score)作为模型评估度量。

对于单一作者的情况,模型的准确率、召回率和

F1 值的计算规则如下:

$$\text{Precision} = \frac{\# \text{CorrectlyPredictedToTheAuthor}}{\# \text{TotalPredictedToTheAuthor}}$$
$$\text{Recall} = \frac{\# \text{CorrectlyPredictedToTheAuthor}}{\# \text{TotalPapaerBelongToTheAuthor}}$$
$$\text{Weight} = \frac{\# \text{UnassignedPagerOfTheAuthor}}{\# \text{TotalUnassignedPager}}$$

对于有多个作者的情况,其准确率为单一作者的准确率乘以单个作者的 F1 值。同理,多个作者的召回率和 F1 值均为单个作者的 F1 值加权后的结果。其计算方式如下:

$$\text{WeightedPrecision} = \sum_{i=1}^M \text{Precision}_i \times \text{weight}_i$$
$$\text{WeightedPrecision} = \sum_{i=1}^M \text{Recall}_i \times \text{weight}_i$$
$$\text{WeightedF}_1 = \frac{2 \times \text{WeightedPrecusuib} \times \text{WeightedRecall}}{\text{WeightedPrecision} + \text{WeightedRecall}}$$

3 实验过程

3.1 样本的构建 本文以 DBLP(<https://dblp.uni-trier.de/db/>)学术搜索平台中的学术文献数据集为实验数据,首先获取论文数据,并对其进行了预先的消歧和标注,实现“作者-文献”的关联,构建现有作者档案文献集,为确保数据集的准确性,本文选取带有 ORCID 的作者,构建“文献-作者”数据集,作为学术文献同名作者消歧的预训练和效果检验。实验数据如表 2 所示。

表 2 数据集概览

	数据量	作者名	实际作者数
原始数据	107 002	109	11 420
训练数据	101 797	109	11 420
测试数据	5 205	-	1 542

通过检索,文本共获取实际作者 11 420 个,共计 107 002 篇论文。由于存在同名情况,共获得 109 个作者名。实验中,本文选择了全部作者的 101 797 篇论文作为模型训练,并随机选择了 1 542 个实际作者的 5 205 篇论文作为测试集。

如图 1 所示,本文在模型训练时,为提高模型的分类效果,将训练集划分为正、负两个样本集合。其中负样本集的作用是提高模型的分类效果,负样本集的生成流程如下所示:

- (1)选取训练集中发表文献数量大于 6 篇的作者,构建“作者-文献”集合;
- (2)在“作者-文献”集合中,随机选取若干个同名作者的集合,并随机抽取集合中每位作者 20% 的文献,对这些文献的作者与文献随机错乱重排,形成负样

本。

3.2 数据预处理 本文选择科技文献的元数据包括作者名、合作者、作者单位、期刊(会议)名、论文标题、发表时间、论文关键词和摘要信息。由于科技文献的来源不同,这些元数据的格式也不同,为此,在特征提取前,需要对其进行专门的处理。具体处理思路如下所示:

- a. 机构名、期刊(会议)名。对于这些短文本数据,本文采用常规的处理方法,首先去除文本中的特殊字符,并将所有内容转换为小写格式。
- b. 作者名。由于不同文献对作者名的格式要求不同,使得文献集合中作者名存在多种格式,如作者“Wang Ping”,会存在“Wang P”“Ping Zhang”“WANG PING”等多种形式,为此,需要对其进行处理并统一格式。本文采用字典映射的方式对其进行处理,首先构建标准作者名格式,随后将文献中的作者名映射为标准格式。对于一些姓名前后颠倒的形式,本文采用统计字母个数的方式,对共现字符数相同的姓名进行规范格式的映射。通过人为检测,字典映射方法基本可以准确的实现作者名的关联。
- c. 论文发表时间。发表时间预处理需要解决的问题是字段缺失和少量的错误数据。由于人为的错误,少数论文发表年份为错误数据,如“2030 年”。对于缺失数据和错误数据,本文采用的策略是,如果该作者有多篇论文,则以该作者所有论文发表时间的中位数进行填充,如果该作者仅有一篇论文,则使用数据集中所有论文发表时间的中位数进行填充。

d. 论文标题和摘要信息。预处理的目的是降低文本的维度,主要采用词形归一化处理,使用 NLTK 工具进行了词形还原。

预处理完成后,每篇文献形成如下数据表示:  
Article: {year, author, coauthor, orgs, title, venue, keywords, abstract}

3.3 文献元数据的特征提取 本文实现人名消歧,需要构建文献内外部元数据的交叉特征集,具体的操作如下:

a. 时间元数据的特征。时间是科技文献的重要特征,可以反映作者的研究主题和特点。考虑到学者在一个连续时间内的研究成果具有一定相似性的特征,消歧模型构建时,本文采用了多个时间统计策略,具体为:

$$\text{year:} \{ \text{year}_{\text{erally}}, \text{year}_{\text{newest}}, \text{year}_{\text{mean}}, \text{year}_{\text{std}}, \text{year}_{\text{median}} \}$$
上述的统计策略分别表示某位作者发表第一篇文献的时间,最新论文发表的时间,发表论文时间的均值、标准差和中位数。对于新增文献,分别与同名作者的 5 个时间统计特征进行比对,计算时间序列上匹配

程度。

b. 作者、单位、期刊名等元数据的特征。本文发现这一类特征采用简单的匹配方式就可获得较好的效果,为此,本文在实验中采用相等匹配和集合匹配两种方法。

相等匹配即为字符串完全相等。对于作者、期刊名等元数据,通过规范化处理,对完全相同的元数据则认为匹配成功。

集合匹配主要针对作者所在单位。由于作者单位元数据存在格式不一致的问题,在实验中,本文发现有的作者单位是简略的信息,而有些是非常完整的信息。如:简略的形式为“\* \* \* University”,而完整的形式为“\* \* \* University \* \* \* Department, \* \* \* Street”。为此,实验中,首先对作者单位元数据按照空格分词,形成词汇集;在匹配时,将新增文献的作者单位与已有作者的单位进行比对,计算两个集合交集与最短集合长度之间的比值,如果达到一定的阈值,则认为匹配成功。

c. 文献内部元数据的特征。人名消歧选取的文献内部元数据包括:摘要、标题和关键字。为了更好地获取这些元数据的语义特征,本文借助 BERT 模型将文献的这些元数据转换成一个带有语义信息的向量。假设文献  $a = \{ abstract, title, keywords \}$ ,使用 BERT 模型将  $a$  转换为向量  $v_a$ ,则某位作者的文献向量集合即为  $\{v_{a1}, v_{a2}, \dots, v_{an}\}$ ,其中  $n$  为该作者的发文总量。对于新增文献,其向量表示为  $v_b$ ,论文归属问题及转换为  $v_b$  和向量集合文献的相似度计算。

在构建文献内部特征向量时,使用 PyTorch 深度学习框架实现的 BERT 模型 (<https://pypi.org/project/sentence-transformers/>) 来进行语义特征表示,实验中,将元数据转换为一个 768 维的数值型向量表示。在匹配计算时,关键字、标题和摘要分别转化为 5 个统计特征,即最小值、最大值、均值、标准差和中位数。以摘要为例,图 2 显示了文献 a 和文献 b 的摘要经过 BERT 语义向量表示后的结果。

	abstract_a	abstract_b
0	[-0.4023114, 0.5755189, -0.30681625, 0.0454310...	[[0.3303879, 0.6347718, 0.85132164, 0.23749858...
1	[-0.4023114, 0.5755189, -0.30681625, 0.0454310...	[[0.45477986, 0.43924397, -0.02296334, 0.4726...
2	[-0.4023114, 0.5755189, -0.30681625, 0.0454310...	[[

图 2 学术文献摘要 BERT 语义向量表示

图 3 显示了文献 a 和文献 b 的摘要相似度计算结果。

	abstract_sims	abstract_sims_min	abstract_sims_max	abstract_sims_mean	abstract_sims_std	abstract_sims_mmd
0	[0.49968525767326355]	0.499685	0.499685	0.499685	0.0	0.499685
1	[0.537568986415863]	0.537569	0.537569	0.537569	0.0	0.537569
2	[0]	0.000000	0.000000	0.000000	0.0	0.000000

图 3 学术文献摘要相似度特征

最后,将外部特征和内部特征融合后的结果输入 XGBoost 进行模型训练。

3.4 作者匹配 根据上文对文献内外部特征提取完成后,基于构建的正负样本,本文使用 XGBoost (<https://pypi.org/project/xgboost/>) 模型完成特征的融合和作者匹配。

具体匹配的步骤为,首先将样本数据输入 XGBoost,通过调参,完成模型的训练;随后,对于新增文献,先将新增文献作者与现有作者进行匹配,获得一些列同名作者;随后采用 XGBoost 对新增文献的特征与同名作者的文献特征进行相似度计算;最后输出匹配分数最高的作者 ID 作为新增文献的所属作者。

3.5 结果对比 为了评估模型的效果,本文对比了多组实验的结果,实验具体设置如下:

a. BERT 语义特征和学术文献基本特征的方法,本文方法。

b. Word2Vec(使用 gensim 包完成)特征的 SVM 分类方法,简化表示为 Word2Vec + SVM。

c. Word2Vec 特征的 XGBoost 分类方法,简化表示为 Word2Vec + XGBoosts。

d. 对比文献<sup>[17]</sup>提出的基于 SDR 的人名消歧方法。

具体的实验结果如表 3 所示。

表 3 增量消歧实验结果

增量消歧方法\评测指标	Weighted Precision	Weighted Recall	Weighted F <sub>1</sub> -score
本文方法(BERT + GBoost)	0.96299	0.94277	0.95674
Word2vec+SVM	0.917999	0.880788	0.899009
Word2vec +XGBoost	0.89723	0.85672	0.88921
基于 SDR <sup>[17]</sup> 的人名消歧	0.9821	0.7675	0.8617

从表 3 中几种方法在测试集上的指标数值可见,本文提出的方法表现出较好的效果,F1 取得了 95.6% 的分值,高于其他模型的性能表现。相比于基于 Word2vec + XGBoost 的方法,本文方法提高了约 7%,相比于 Word2vec+SVM 的方案,本文方法提高了约 6%。

此外,对比文献[17]的实验结果,虽然基于 SDR 方法的准确值更高,但是召回率和 F1 值方面来看,本文的方法更优。且文献[17]采用的是数据量小(实验数据为 19 个作者的 88 篇论文),因此从总体上来看,本文方法具有一定的合理性。

4 结 语

增量消歧,需要快速且准确地将文献分配给系统中已有作者,这是学术文献数据库在进行更新时最亟待解决的问题。本文提出了一种融合多特征的相似度



匹配方法实现增量式人名消歧研究,借助 BERT 自然语言处理模型和 XGBoost 分类模型进行相似度匹配,将新增文献分配给相似度得分最高的作者。通过实验对比,本文方法获得了较好的增量式人名消歧效果。

然而,该文的研究也存在一些改进的地方,如:

a. 本文选取了科技文献的多个元数据,形成了多特征融合的消歧模型,但在实际应用中,如何设定每个特征的权重值,合理分配外部特征匹配和内部语义特征所占的比重,进一步提高模型的准确率,是实践应用中需要解决的问题。

b. 本研究仅限于对英文语言的学术文献作者同名消歧,对跨语言的同名作者消歧并未涉略。因为不同语言之间的名字形式不一致,跨语言学术文献同名作者消歧更具有挑战,同时也是进行多来源学术文献组织和管理的难题,今后会对这方面逐步开展研究。

#### 参 考 文 献

- [1] Johnson R, Watkinson A, Mabe M. The STM report an overview of scientific and scholarly publishing [ED/OL]. [2021-02-01]. [https://www.stm-assoc.org/2018\\_10\\_04\\_STM\\_Report\\_2018.pdf](https://www.stm-assoc.org/2018_10_04_STM_Report_2018.pdf).
- [2] 付 媛,朱礼军,韩红旗. 姓名消歧方法研究进展[J]. 情报工程, 2016, 2(1):53-58.
- [3] Li S, Gao C, Miao C. Author name disambiguation using a graph model with node splitting and merging based on bibliographic information[J]. *Scientometrics*, 2014,100:15-50.
- [4] Bagga A, Baldwin B. Entity-based cross-document conferencing using the vector space model [C]//Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, 1998:79-85.
- [5] Fan X M, Wang J Y, Pu X, et al. On graph-based name disambiguation [J]. *Journal of Data and Information Quality*, 2011,2(2):23-56.
- [6] 线岩团,余正涛,洪旭东,等. 基于特征加权重叠度的中文实体协同消歧方法[J]. *中文信息学报*,2017,31(2):36-41.
- [7] 张 雄,陈福才,黄瑞阳. 基于融合特征相似度的实体消歧方法研究[J]. *计算机应用研究*,2017,34(2):347-350, 396.
- [8] 朱亮亮. 利用改进的 K-means 算法实现文献著者人名消歧[J]. *软件导刊*,2013(5):63-66.
- [9] 任景华. 利用优化的 DBSCAN 算法进行文献著者人名消歧[J]. *图书馆理论与实践*, 2014(12):61-65.
- [10] Delgado A D, Martinez R, Fresno V, et al. A data driven approach for person name disambiguation in web search results [C]//Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics, Dublin, 2014:301-310.
- [11] Han W, Xu B, Zhao T. Study on Chinese person name disambiguation based on multi-stage strategy [C]//Proceedings of 2011 the 8th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). Shanghai, China: IEEE,2011.
- [12] Kim K, Khabsa M, Giles C L. Random forest DBSCAN for USP-TO inventor name disambiguation [EB/OL]. [2020-04-27]. <https://arxiv.org/abs/1602.01792>.
- [13] Li G, Lai R, D'Amour A, et al. Disambiguation and co-authorship networks of the U. S. patent inventor database (1975-2010) [J]. *Research Policy*, 2014, 43(6):941-955.
- [14] Levin M, Krawczyk S, Bethard S, et al. Citation based bootstrapping for large-scale author disambiguation[J]. *J Am Soc Inf Sci Tec*, 2012,63(5):1030-1047.
- [15] Pereira D A, Ribeiro-Neto B, Ziviani N, et al. Using web information for author name disambiguation [C]//Proceedings of the 9th ACM/IEEE-CS Joint International Conference on Digital Libraries (JCDL'09). New York: ACM, 2009:49-58.
- [16] José O, José L S, Xavier S, et al. Authors semantic disambiguation on heterogeneous bibliographic sources [C]//2017 XLIII Latin American Computer Conference (CLEI), 2017:1-9.
- [17] 翟晓瑞,韩红旗,张运良,等. 基于稀疏分布式表征的英文著者姓名消歧研究[J]. *计算机应用研究*,2019,36(12):3534-3538.
- [18] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. *Nature*, 2015,521(7553):436-444.
- [19] 张 晗,徐 硕,乔晓东. 融合科技文献内外部特征的主题模型发展综述[J]. *情报学报*, 2014, 33(10):1108-1120.
- [20] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: NAACL, 2019: 4171-4186.
- [21] Chen T, Guestrin C. XGBoost: A scalable tree boosting system [C]// the 22nd ACM SIGKDD International Conference, 2016: 785-794.

(责编:王育英;校对:王平军)