谨以此论文献给我的家人，我的导师，以及所有帮助和支持我的朋友们。

------------------侯海东

论文作者身份识别及认领服务的研究

学位论文答辩日期：

指导教师签字：

答辩委员会成员签字：

**独 创 声 明**

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含未获得（注：如没有其他需要特别声明的，本栏可空）或其他教育机构的学位或证书使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名： 签字日期： 年 月 日

**学位论文版权使用授权书**

本学位论文作者完全了解国家有关保留、使用学位论文的法律、法规和学校有关规定，并同意以下事项：

1、学校有权保留并向国家有关部门或机构送交本学位论文的复印件和磁盘，允许论文被查阅和借阅；

2、学校可以将本学位论文的全部或部分内容编入学校学位论文数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编本学位论文；

3、学校可以基于教学及科研需要合理使用本学位论文。

需保密的学位论文在解密后适用本授权书。

学位论文作者签名： 导师签字：

签字日期： 年 月 日 签字日期： 年 月 日

论文作者身份识别及认领服务的研究

摘 要

关键词：

Research of author's identification and paper claim service

Abstract

Keywords:

目 录

[1 绪论 1](#_Toc503015826)

[1.1 研究背景和意义 1](#_Toc503015827)

[1.1.1 作者识别问题定义 2](#_Toc503015828)

[1.1.2 课题研究的意义 3](#_Toc503015829)

[1.2 研究现状及分析 4](#_Toc503015830)

[1.2.1 作者身份识别研究现状 4](#_Toc503015831)

[1.2.2 论文认领服务研究现状 4](#_Toc503015832)

[1.2.3 作者识别问题难点 4](#_Toc503015833)

[1.3 研究内容和主要工作 4](#_Toc503015834)

[1.4 论文内容安排 4](#_Toc503015835)

[1.5 本章小结 4](#_Toc503015836)

[2 知识库与测试数据集构建及准备工作 5](#_Toc503015837)

[2.1 实体知识库的构建 5](#_Toc503015838)

[2.2 测试数据集的构建 6](#_Toc503015839)

[2.2.1 文献记录的获取与整理 6](#_Toc503015840)

[2.2.2 测试数据集的选取 7](#_Toc503015841)

[2.3 评价指标 8](#_Toc503015842)

[2.4 作者识别的特征选择 10](#_Toc503015843)

[2.5 本章小结 11](#_Toc503015844)

[3 基于知识库的作者身份识别方法 12](#_Toc503015845)

[3.1 问题分析与方法流程 12](#_Toc503015846)

[3.2 预处理阶段 14](#_Toc503015847)

[3.2.1 作者姓名规范化 15](#_Toc503015848)

[3.2.1.1 姓名拼写形式分析 15](#_Toc503015849)

[3.2.1.2 姓名规范化算法描述 16](#_Toc503015850)

[3.2.2 作者所属单位抽取 16](#_Toc503015851)

[3.2.3 通讯作者标注 17](#_Toc503015852)

[3.3 候选作者生成阶段 18](#_Toc503015853)

[3.3.1 候选作者查找流程 18](#_Toc503015854)

[3.3.2 实验结果与分析 18](#_Toc503015855)

[3.4 本章小结 18](#_Toc503015856)

[4 基于合作网络的重名作者消歧 18](#_Toc503015857)

[4.1 相关概念介绍 18](#_Toc503015858)

[4.1.1 合作网络简介 18](#_Toc503015859)

[4.1.2 节点相似度 18](#_Toc503015860)

[4.2 作者合作网络的挖掘与构建 18](#_Toc503015861)

[4.2.1 合作关系图模型建模 18](#_Toc503015862)

[4.2.2 作者合作网络构建 18](#_Toc503015863)

[4.3 改进的作者身份识别方法 18](#_Toc503015864)

[4.3.1 候选作者选择阶段 18](#_Toc503015865)

[4.3.2 节点相似度计算 18](#_Toc503015866)

[4.3.3 实验结果与分析 18](#_Toc503015867)

[4.4 本章小结 18](#_Toc503015868)

[5 论文推送认领服务分析设计 19](#_Toc503015869)

[5.1 后台技术框架简介 19](#_Toc503015870)

[5.1.1 ASP.NET Web API 19](#_Toc503015871)

[5.1.2 Entity Framework 19](#_Toc503015872)

[5.1.3 Autofac 19](#_Toc503015873)

[5.1.4 AutoMapper 19](#_Toc503015874)

[5.2 推送认领服务分析 19](#_Toc503015875)

[5.3 系统设计与实现 19](#_Toc503015876)

[5.3.1 认领平台 19](#_Toc503015877)

[5.3.2 推送认领 19](#_Toc503015878)

[5.3.3 补充认领 19](#_Toc503015879)

[5.4 本章小结 19](#_Toc503015880)

[6 总结与展望 19](#_Toc503015881)

[6.1 论文总结 19](#_Toc503015882)

[6.2 研究展望 19](#_Toc503015883)

[参考文献 20](#_Toc503015884)

[附录 20](#_Toc503015885)

[致谢 21](#_Toc503015886)

[个人简历 22](#_Toc503015887)

[在校期间研究成果 22](#_Toc503015888)

# 绪论

## 研究背景和意义

随着高等教育改革工作的深入推进和稳步发展，我国高校科研水平迅速提升，科研论文产出数量也逐年增加。与此同时，科研论文的收录引用作为科研评价体系的重要指标之一，对于评价个人学术能力和科研水平也有至关重要的作用。我国现行科研体制下对于千人计划、杰出青年、长江学者、院士、创新群体等人才评定、职称评定以及科研项目、奖项的申报通常需要申请者出具自己科研论文是否收录在重要检索刊物以及论文被引用情况，以证明个人、团队或者科研成果的创新影响力。但申请者在收集整理自己论文的过程中，很多科研人员都有找不到论文的经历，他们为了节省时间和精力往往提交格式不够标准的论文数据，因此会希望科研管理部门辅助其统计论文信息，而他们只需从中复核挑选自己的论文。

同时论文作为科研成果的一种重要呈现方式，也在一定程度上反映了高校的科研实力和学术水平。近年来，我国高校越来越重视SCI学术论文的统计管理和分析利用[1-3]，并加快推进各自的知识库等学术资产管理系统的建设和实践[4-5]。因此，无论是高校还是个人，都需要准确的论文统计分析服务来为科研管理及人才发展提供参考依据。

但是，科研管理部门通过Web of Science等平台检索收集本单位作者发表的SCI论文信息后，在存缴入库的时候经常会面临一个棘手的问题：如何识别本单位的论文作者并将其归属到相应部门？目前，很多高校的科研管理部门都是采用人工识别判断的方式。他们根据现有的数据和经验，通过人工识别论文作者和归属部门，并按照不同部门对论文进行整理汇总报送至相应部门进行作者确认。相关部门人员在收到论文作者确认要求后，通过现有经验自行确认或者报送相关作者本人进行确认的方式完成论文作者的修正、确认后，再将论文数据整理汇总至科研管理部门。科研管理部门人员等待所有部门都完成论文作者确认后，将论文汇总结果存入数据库，本批次论文报送确认工作结束。整个工作流程如图 1‑1所示。



图 1‑1 作者识别确认流程时序图

目前，办公流程逐渐向着信息化、自动化、智能化发展。仔细分析上述论文作者识别、报送确认工作流程不难发现，整个流程需要跨越不同的职能部门和人员、全程人工参与线下操作，导致工作流程环节多、时间长，工作效率低下。同时论文作者采用人工判别的方式，不仅工作量大而且工作枯燥容易出错。尤其中文作者拼音写法具有多样性，易出现重名和引发混淆，导致科研管理部门无法快速准确的完成论文作者识别工作。因此，如何自动、准确的识别论文中的作者身份，建立论文作者与其实际作者的关联关系，并提供一个线上的论文认领确认系统作为支撑，成为科研管理部门亟待解决的问题。而这也是本文着重讨论的问题。

### 作者识别问题定义

作者身份识别与使用生物识别技术鉴别用户身份相类似，生物识别技术是利用人体固有的生物特征（包括指纹、人脸、虹膜、声音等）与特征信息库进行匹配，从而识别生物特征拥有者的真实身份。而作者身份识别是利用文献记录的一些特征属性，如文献标题、作者、作者单位、合作者、出版刊物等信息确定作者真实身份，并将文献中的该作者与专家库中的真实作者链接起来。下面以一篇论文记录为例来描述作者身份识别过程，在图 1‑2中，将论文文献信息导入系统后，系统首先对论文文献数据进行预处理和再加工，根据抽取出的特征信息查找候选作者集合，然后调用重名作者消歧程序执行消歧工作，最后将得到的候选作者与专家库中的专家关联起来。



图 1‑2 作者身份识别过程示意

作为实体链接[6]在文献检索领域的特化任务，论文作者身份识别旨在将文献记录中出现的作者名称正确链接到专家库中相应的专家，以解决作者的歧义性。其研究内容主要包括两部分，第一部分是作者身份识别问题，面临的主要挑战是作者重名现象，可以采用作者间共著关系来解决，很多研究者都对基于学术论文的合著网络进行了研究，有很多值得借鉴的地方；第二部分是建立作者识别结果反馈机制问题，以确保论文作者与其真实作者的准确关联，该问题相对来说较为简单，可以通过建立一种论文认领确认服务来实现。

### 课题研究的意义

本课题的研究目的在于通过对论文文献数据的充分挖掘，对作者身份识别问题以及面临的重名作者消歧问题进行研究，并探索一种可靠地服务机制确保作者身份的准确识别。通过本文的研究工作，我们希望可以通过作者身份识别系统，给科研管理人员提供一种更加方面高效的识别论文作者的方法，使工作人员摆脱繁琐沉重的人工识别工作，提高高校科研管理效率。

随着论文等学术资源的日益增长，给各高校的学术信息资源管理带来了新的挑战。目前，中国科学院、北京大学等高校和科研单位都已建成了自己的机构知识库，以实现对学术资产的数字化管理。然而知识库中获取的大量文献难以归属到个人，导致资源利用效率不高。本文所述作者识别方法同样也有利于高校推进机构知识库、成果管理系统等学术资产管理系统的建设和使用，方便科研管理人员对个人、机构的论文发表、收录情况进行统计分析，为高校决策、产出论文统计服务提供有力支撑。

## 研究现状及分析

### 作者身份识别研究现状

### 论文认领服务研究现状

### 作者识别问题难点

1.一词多义（一个名字对应多个人）。2.同义问题（一个作者多种署名方式）。

## 研究内容和主要工作

## 论文内容安排

## 本章小结

# 知识库与测试数据集构建及准备工作

## 实体知识库的构建

在实体链接任务中，主要借助维基百科等知识库中的信息将实体指称准确链接到相应实体。同样在作者身份识别系统中，自动识别论文作者也需依赖实体知识库的支撑。因此在作者身份识别系统设计和探究阶段，必须准备和构建专家、部门等实体知识库作为基础数据，这是对论文作者识别方法分析的关键基础。

实体知识库要为作者身份识别任务提供丰富的实体特征信息，主要包括以下实体数据：

1. 部门实体

部门实体指本单位及其下属部门，包括部门中文名称、编码、父部门、邮编等属性。但是作者在论文中对同一个部门的英文名称可能有几种不同的标注形式，比如信息科学与工程学院，作者在论文中标注的格式可能为Coll Informat Sci & Engn、Coll Informat Sci & Technol和College Informat Sci & Engn等。因此，对于每一个部门实体还应该构建其常见的英文名称作为别名信息。

1. 专家实体

专家实体指校内在编教师，包括工号、姓名、姓名全拼、姓名简拼（姓全拼+逗号+名所有单字首字母）、出生日期、身份证号、手机号、邮箱、所属学院、所属系（实验室）、双聘单位等信息。

因部门组织架构变化、正常人事调动或科研需求等原因，可能会存在部分专家的所属部门发生变动。所以在知识库的构建中也要考虑这部分因素，在专家与部门的所属关系中添加到职时间、离职时间以及历史任职部门，以记录这些变化。

1. 学生实体

学生实体指曾就读于本单位的硕士研究生和博士研究生，包括学号、姓名、出生日期、所属学院、所学专业、导师姓名、入校时间、离校时间等信息。

研究生在就读期间，通常会在导师的指导下，对某一课题进行深入研究，当积累了一定的学术成果后，往往会将成果以论文的形式发表，而导师通常是其中的一个作者。因此导师对学生的这种指导关系，能在一定程度上辅助进行作者身份识别。

1. 论文实体

论文实体指本单位师生所发表的SCI论文，包括论文标题、作者、发表期刊、发表时间、关键词、参考文献、被引频次等信息。

知识库中的论文记录均是人工标注作者并经过相关作者确认的真实数据。在人工标注作者过程中，对于校外作者标注为外单位人员，对于校内作者标注分为两种情况：（1）作者为校内专家，标注为专家的名字；（2）作者为校内学生，则直接标注为学生不作区分。这些标注数据可用于验证作者识别结果的正确性和专家间合作关系等数据的深层次分析和挖掘，是构建机构知识库的数据基础。

通过搜集、整理现有的数据和资料，构建出的实体知识库如表 2‑1所示。其中专家与学生实体在构建知识库的过程均放入了人员表中，以人员类型字段进行区分。实体基础数据准备完毕后，还需准备测试数据。下一节将详细介绍测试数据集的构建。

表 2‑1 知识库中的实体及说明

|  |  |  |
| --- | --- | --- |
| 实体类型 | 实体数目 | 说明 |
| 部门 | 225 | 包括学院、系、实验室等 |
| 专家 | 4857 | 截至2017年7月入职的专家 |
| 学生 | 27810 | 2006.09-2015.09期间就读于本单位的研究生 |
| 论文 | 13769 | 截至2017年7月人工标注作者的论文 |

## 测试数据集的构建

### 文献记录的获取与整理

在对作者身份识别方法进行研究的过程中，除了要准备实体知识库中的基础数据，构建测试数据也是对论文作者识别系统进行验证和改进的重要环节。本文用于测试系统的文献记录数据均是从在线文献检索系统检索下载得到的。在浩如烟海的文献资料中检索符合自己需求的数据，需要依赖文献检索系统的支持。常用的文献检索系统有Web of Science和Engineering Village。Web of Science（WOS）是世界范围内最权威、覆盖学科最多的科技文献检索平台，覆盖了SCIE、SSCI、CPCI-S等多个权威学术文献数据库。而Engineering Village是最权威的工程、应用科学领域文献检索平台，拥有著名的工程索引（EI）数据库。利用这些检索工具，可以按照作者、标题等多种筛选条件检索论文，并将检索结果保存为Excel等文件格式进行下载。

目前，各大已有的文件检索数据库并没有一个统一的数据格式，从不同的检索库中获取到的文献记录格式差异较大。因此获取到的文献记录需要经过数据融合构造成统一格式的数据，才能存入文献数据库。在进行文献数据融合前，首先要对记录进行查重处理。由于SCI和EI作为索引数据库收录了部分相同出版刊物，在获取到的文献记录中存在重复的文献数据，可以利用文献标题、作者等字段查找剔除重复数据。通过对SCI与EI字段格式进行详细分析（详见文后附录 A），对于查重处理后的文献记录，以SCI字段格式为蓝本，采用基于SCI和EI字段对应关系的文献数据融合算法[]将文献数据融合成统一格式的数据集，存入数据库。文献记录查重融合流程如图 2‑1所示。



图 2‑1 文献记录查重融合数据流图

### 测试数据集的选取

测试数据集是论证作者身份识别方法有效性的重要工具，选择与构建一个适当的测试数据集至关重要。在构建测试数据集时，应该满足以下条件：

1. 文献记录要包含足够丰富的文献特征信息，标题、作者等关键字段值不能缺失；
2. 选择的文献记录中作者姓名、作者地址要有代表性，要充分包含作者署名时的各种变化；
3. 文献记录应包含不同的作者类型，如校内专家、学生、外单位人员；
4. 文献记录要有合理的数据分布，如有些专家会单独发表文章，有些专家会和不同研究方向的专家合作发表文章等；
5. 文献记录的数目要有一定的样本量。

在文献数据库获取的众多论文中，有很多论文人工根本无法直接判断出作者属于专家库中的哪个专家。因此在选择测试数据时还要考虑已有的论文标注集的影响，应该选择已在标注集中进行人工标注作者的论文记录作为测试数据。

同时由于无效值、缺失值等脏数据的存在，获取到的文献记录并不能直接用于作者身份识别系统的测试，需要对数据进行清洗，剔除掉无用数据。数据清洗（Data Cleaning）是指为了提高数据质量发现并纠正数据中可识别错误的过程，包括检查数据一致性，处理无效值和缺失值等。利用数据清洗相关方法解决文献数据中存在的少量重复记录、发表日期字段格式不统一、发表日期中年份与发表年字段值不一致等问题后，即可得到干净、标准的文献测试数据。

基于清洗后的文献记录，本文在选择测试数据集的时候，充分考虑了上述条件中的情况，做到有针对性的选择。最终本文从中国海洋大学师生在2015年至2016年期间发表的SCI论文中筛选出符合条件的论文共3576篇作为测试数据，这些论文作者总人次为19483，其中校外作者为7223人次，校内专家为6289人次，校内学生5971人次。后面进行的作者身份识别实验均基于此数据集。

## 评价指标

作者身份识别实验的结果受到诸多因素的影响，主要有：部门及其别名的数量、专家与学生的数量、测试论文的数量、测试论文中作者样本的分布是否平衡等。为了评价作者身份识别系统的有效性，本文借鉴信息检索与分类问题中的评测方法，采用查全率（Recall）、查准率（Precision）和F测量值（F-Measure）作为评价指标。

在作者身份识别任务中，系统的目标是要为论文找出其所有的真实作者。定义表示系统中的论文集合，对于其中一篇论文，如果把识别论文的作者身份这个子任务当成分类器，那么论文的所有真实作者为正类，系统中除论文的真实作者外的所有其他人员为负类。则查全率可以定义为识别结果中正确找出的论文的真实作者样本数占所有论文的真实作者样本数的百分比。通过查全率可以考察对各个论文样本作者识别的完备性。

如表 2‑2中的二分类混淆矩阵所示，如果用表示正确识别出的论文的真实作者的样本数量，用表示实际是论文的真实作者但没有正确识别出的样本数量，则查全率的计算方法如式(2-1)所示。

(2-1)

表 2‑2 二分类混淆矩阵

|  |  |  |  |
| --- | --- | --- | --- |
|  | 实际类别 | | |
|  | | 正 类 | 负 类 |
| 识别类别 | 正类 | *TP*(True Positive) | *FP*(False Positive) |
| 负类 | *FN*(False Negative) | *TN*(True Negative) |

查准率是指识别结果中正确识别出的论文的真实作者样本数占所有系统找出的论文的作者样本数的百分比。通过查准率可以考察各个论文样本的作者识别的正确性。如果用表示正确识别出的论文的真实作者的样本数量，用表示实际不属于论文的真实作者，但被错误的判断为论文的作者的样本数量，查准率的计算方法如式(2-2)所示。

(2-2)

F测量值可以平衡考察查全率和查准率，来衡量作者识别方法的整体效果，是二者的调和平均值，通常采用F1测量值。其计算公式如式(2-3)所示。

(2-3)

微平均（Micro-averaging）可以综合各个论文作者识别子任务的识别结果综合评价系统的性能。微平均基于各个论文作者识别子任务中正确分类的样本数和错误分类的样本数，来计算系统整体论文识别的查全率、查准率和F测量值。查全率、查准率和F测量值的微平均计算公式如式(2-4)、式(2-5)和式(2-6)所示。

(2-4)

(2-5)

(2-6)

## 作者识别的特征选择

文献信息通常都会包含标题、作者、作者地址、摘要、关键词、期刊、期卷号、页码、参考文献、分类类别等特征元素。对于作者身份识别问题，这些特征元素不都具有识别能力，同时具有识别能力的特征之间对于解决同名作者消歧问题的能力大小也有差异。如作者全名、作者地址、出版日期等具有很强的作者识别能力，利用这些信息可以很容易识别作者并区分重名作者，像出版物信息虽然有一定的作者消歧能力，但仅依靠它们很难直接区分出重名作者。虽然文献数据库中论文记录有几十个特征元素，但实际上可以用于作者身份识别的特征元素屈指可数，而且不同的文献数据库提供的记录字段不同，在系统的设计之初应考虑尽可能的兼容更多的文献数据库。并且一般在文献数据库中很多字段都存在数据缺失的情况，从而使作者身份识别问题变得更加复杂化。为了有效解决作者身份识别问题，充分利用各个特征信息，需要对SCI论文字段进行深入、细致的分析。

1. 作者全名、作者简称字段

作者全名字段存储的是论文所有作者的英文名称，对于国内作者通常指姓名汉语拼音的全称，而作者简称字段存储的则是所有作者姓名的简称形式（姓氏全拼+逗号+名字所有单字首字母）。姓名作为作者在论文中的标识，具有很高的身份辨识能力，通过它可以很容易的找到该作者的身份。而且在该字段中还包含着在本篇论文中作者与作者之间的共著关系，共著关系是重名作者消歧过程中最重要的特征。因此该字段是作者识别过程中最关键的特征之一。

1. 作者地址、通讯作者地址字段

作者地址字段以一定的格式存储了每个作者在发表论文时的署名单位地址，而通讯作者地址则标注了通讯作者及其单位地址。透过这些字段，可以清楚的知道每个作者的所属单位，有助于辨别作者身份。因此这些字段也是本文在设计作者识别方法时重点考虑的特征。

1. 出版日期、出版年字段

出版日期（出版年）字段表明了论文在相应出版刊物发表的大致时间。通常论文是在作者署名单位在职期间发表的，因此结合人员的在职、离职情况与论文的出版时间，可以用来判断某个人员是否是论文的真实作者。本文在解决作者身份识别问题的过程中，这些字段也是考虑因素之一。

1. 作者关键词、附加关键词字段

关键词字段通常能够反映一个作者的研究方向，而且在论文发表的过程中具有一定的连续性，因此具有一定的作者区分能力。但在文献数据库中该字段存在着较多的数据缺失情况，而且相比于合作关系较强的作者辨别能力，关键词的作者辨别能力则相对弱一些，因此本文在解决作者身份识别问题的过程中暂不考虑该特征因素。

文献数据库中存储的SCI论文记录字段较为齐全，并且如作者地址、出版日期等重要字段数据缺失情况较少，携带着丰富的文献特征信息。本文从微观角度对SCI论文文献特征信息进行深入挖掘、分析，并在此基础上选择了作者全名（作者简称）、作者地址（通讯作者地址）、出版日期（出版年）字段作为文献记录特征，并将其运用到作者身份识别方法的设计中。

## 本章小结

本章主要介绍了研究和解决论文作者身份识别问题中的相关准备工作。君欲善其事，必先利其器。本章首先介绍了知识库中需要准备的实体数据，这些数据是后续研究作者识别问题的基础。然后讲述了测试文献数据的获取与整理以及测试数据集的选取与构建，并介绍了作者识别方法的评价指标。最后通过对SCI论文文献记录字段进行细致分析，选取了作者识别方法所用的文献字段作为识别特征。

# 基于知识库的作者身份识别方法

在作者身份识别任务中，论文记录字段数据缺失多，可用信息少，作者署名和单位存在大量不同表述形式等，给作者身份识别带来了不小的困难。本章对论文作者身份识别中存在的问题进行了分析，初步设计了作者识别方法及其流程。因此，如何最大限度的识别和链接论文中的作者是本章的主要工作。

## 问题分析与方法流程

论文作者身份识别任务的目标是，对于给定的一篇论文，找到一种有效的识别方法能够准确识别论文中出现的作者并将其链接到知识库中相应的实体。具体到本文所讨论的作者识别问题，指的是对于给定论文中的作者，若作者是外单位人员则将作者标记为外单位人员，若作者是单位内的学生则将作者链接到知识库中的学生实体，若作者是单位内的专家则将作者链接到知识库中的专家实体。下面给出论文作者识别问题的形式化定义。

令表示一篇已发表的论文，为论文的作者组成的集合，为知识库中专家、学生和外单位人员组成的集合，其中任意一个实体元素（元素除外）表示知识库中的某个专家或学生，特别地，元素代表所有的外单位人员。设关系为从集合到的一个映射，则作者身份识别问题可以定义为通过一定的技术手段，将论文中的每一个作者映射到实体集合中，并建立相应的链接关系，示意图如图 3‑1所示。因此本文的主要研究内容就是如何找出并建立映射关系。



图 3‑1 作者识别问题示意图

为了建立论文作者到实体集合的映射关系，通过对论文文献记录和知识库数据进行分析，发现其中存在不少问题，主要如下：

1. 同一个人在其不同背景下的论文中作者署名可能有多种拼写形式。如对于作者王建伟，他在论文中的署名形式可能为Wang Jianwei、Jianwei Wang、Wang J.W.等不同形式。
2. 多个人在其论文中可能因使用相同的署名标识而引起识别的冲突。如作者王伟和王威可能在论文中均署名Wang Wei。
3. 同一个部门的人在其论文中标注所属单位时，可能使用不同的标注形式。比如对于信息科学与工程学院，不同的人可能标注为Coll Informat Sci & Engn、Coll Informat Sci & Technol和College Informat Sci & Engn等。
4. 一个人可能因所属部门改变在不同时期的论文中使用不同的署名单位。同样部门名称的更换也可能会导致该部门的人在不同时期的论文中使用不同的署名单位。
5. 作者在论文中标注其所属单位时，可能存在少量的拼写错误。

针对以上问题，本文给出了如下解决方案，如下所示：

1. 对论文文献记录进行预处理，抽取作者识别方法所需的元数据，如作者全名、作者地址、出版日期等。
2. 构建了一个常用姓氏字典，根据汉语拼音人名拼写规律区分作者署名拼写形式，并将其规范化。
3. 构建本单位下属部门的英文名称别名字典，以识别部门的不同标注形式以及改名问题。同时使用基于编辑距离的相似度计算方法解决部门名称拼写错误问题。
4. 构建人员与部门的关系字典，以记录人员同时在多个部门任职和组织关系调动等情况。
5. 对于相同作者署名标识导致的识别冲突，利用作者在论文中标注的单位以及邮箱等个人信息进行区分。

根据上述解决方案，本文初步设计了基于知识库的作者身份识别方法，整体框架如图 3‑2所示。



图 3‑2 作者识别方法整体框架

根据图 3‑2可知，作者识别方法整个流程主要分为两个阶段：预处理阶段和候选作者生成阶段。预处理阶段主要是从论文文献记录信息中提取待识别作者的元数据，包括作者姓名、所属部门、论文发表时间等信息。候选作者生成阶段主要是根据提取出的论文作者信息作为识别特征，采用一定的匹配规则算法查找候选作者。下面将对这两个阶段的工作内容进行详细介绍。

## 预处理阶段

论文作者身份识别的关键问题是通过提取论文文献记录信息中的作者姓名和所属部门名称等信息作为识别特征，与已有的人员、部门信息进行匹配来识别作者身份。预处理阶段就是对于论文的每一个待识别作者，从论文文献记录中抽取识别该作者身份所需的特征信息，并将其记录下来用于后续查找、匹配作者的过程。对于论文的出版日期和出版年信息，可直接从文献记录数据库中的相应字段获取到。而对于作者姓名、作者署名单位信息的抽取以及通讯作者的标注，由于数据库中文献记录格式复杂、不同的作者有不同的标注形式，因此无法从相应字段直接获取，需要利用一些相应算法进行数据的抽取。

### 作者姓名规范化

#### 姓名拼写形式分析

通常我国学者在投稿英文刊物的时候，对于自己名字的拼写采用汉语拼音的形式。汉语姓名的拼写通常采用姓在前、名在后的形式，而在大多数西方国家英文姓名的拼写则一般采用教名+中间名+姓的方式。为此，很多作者在发表英文文章时，对于自己名字的拼写借鉴了英文姓名的拼写方法，造成在英文论文中汉语姓名的拼写形式比较混乱。在2011年颁布的《中国人名汉语拼音字母拼写规则》标准中，对于汉语人名的拼写规则做出了明确规定（可参考附录 B）。从该标准颁布实施以来，对文化教育、编辑出版、中文信息处理等方面的汉语人名拼写起到了很好的规范作用，姓名拼写形式混乱的现象大大减少。但是由于个人拼写习惯等原因，目前作者在英文论文署名时还存在着不同的拼写形式。

通过对文献数据库中的论文记录进行分析后发现，大部分作者的汉语姓名拼写形式基本符合标准规定，且同一篇论文中的汉语作者姓名拼写形式会保持一致的拼写规则。进一步对作者汉语姓名拼写形式进行整理汇总后，归纳出四种类型的拼写形式，如表 3‑1所示，其中大部分作者采用的是和这两种类型的拼写形式，只有少部分作者采用的是和。通过观察发现，作者姓名不仅拼写形式多种多样，而且在姓名的拼写中还存在着诸如连接号（-）、隔音符号（＇）等多余字符的情况。同时在知识库每个专家和学生实体中，姓名的拼写都采用了姓在前名在后的全拼（如Sun Youbin）和姓全拼+逗号+名所有单字首字母的简拼（如Luo, CP）这两种类型，因此需要识别论文中作者姓名的拼写形式、剔除作者姓名中的多余字符，并进行相应的规范化处理，以方便后续根据不同的姓名拼写形式匹配查找候选作者。具体来说就是将论文中姓名拼写采用形式和的作者姓名处理成符合知识库中全拼形式的姓名，将采用和的姓名处理成符合知识库中简拼形式的姓名。

表 3‑1 作者汉语姓名拼写形式

|  |  |  |
| --- | --- | --- |
| 形式编号 | 类型 | 示例 |
|  | 姓在前名在后的全写形式 | Sun, Youbin 孙幼斌  Yang, Wei-min 杨为民  Li, Xiao Wen 李晓文 |
|  | 名在前姓在后的全写形式 | Jianjun, Zhou 周建军 |
|  | 姓在前名在后的缩写形式一（姓全写，名取每个汉字拼音首字母） | Luo, C. P. 罗常培  Wang, J. -G. 王建国 |
|  | 姓在前名在后的缩写形式二（姓全写，名取第一个汉字拼音首字母） | Chen, X. 陈学海 |

#### 姓名规范化算法描述

基于上述分析，本文设计的作者姓名规范化算法模型由五元组构成，即。其中各元素说明如下：

* 1. 表示输入作者姓名集合，集合有一个属性表示该集合采用的姓名拼写形式类型，每个作者姓名都有一个属性表示该姓名是否是一个英文姓名，表示第个作者姓名是一个英文姓名；
  2. 是一个统计姓和名前后顺序的二元组，其中表示拼写形式中采用姓在前名在后的作者姓名数量，表示拼写形式采用名在前姓在后的姓名数量；
  3. 是一个统计姓名拼写采用全写或缩写形式的二元组，其中表示拼写形式采用全写的作者姓名数量，表示拼写形式采用缩写的姓名数量；
  4. 表示姓名拼写形式的集合，至分别对应着上述的四种拼写类型，表示名前姓后的缩写形式一，表示名前姓后的缩写形式二；
  5. 是作者姓名拼写形式映射函数，表示根据输入的作者姓名集合，确定该篇论文采用的是哪种类型的拼写形式。

映射函数是整个算法模型的核心，其基本设计思路是根据对集合中的每一个作者姓名的分析结果，总结出该篇论文作者采用的姓名拼写形式。对于集合中的其中一个作者姓名，其分析处理过程如下：

1. 作者姓名清洗。根据汉语拼音音节由声母和韵母构成的规律，构建了包含400个左右音节的汉语拼音音节字典，利用这个字典判断是否是由汉语拼音组成的作者姓名，如果是设置并进入下一步骤；否则设置并返回进行下一个姓名的处理操作。
2. 作者姓名拆分。根据分隔符（逗号或空格）将拆分成姓和名两部分，记为。
3. 判断姓名全写、缩写形式。如果或包含小圆点（.）且全部是大写字母，则是缩写形式，将加1；否则是全写形式，将加1。
4. 判断作者姓、名前后顺序。根据《中国四百大姓》整理常见的单姓、复姓的姓氏拼音字典，并根据以下规则判断姓、名前后顺序。
5. 为缩写形式。若为缩写部分，则为姓在前名在后的拼写形式，将加1；否则为名在前姓在后的拼写形式，将加1。
6. 为全写形式。根据姓氏拼音字典判断是否是姓氏，若为姓氏而不是，则为姓在前名在后的拼写形式，将加1；若为姓氏而不是，则为名在前姓在后的拼写形式，将加1；若同时为姓氏或同时不为姓氏，则的姓名前后顺序与集合其他作者姓名的拼写顺序保持一致。

对集合中每一个作者进行上述分析后，得出该篇论文作者采用的姓名拼写形式，映射函数整个算法描述如所示。



图 3‑3 姓名拼写形式判定算法

分析得出每篇论文作者采用的姓名拼写形式后，需要对作者姓名进行处理、规范化为知识库中相应的姓名拼写形式。对于英文作者姓名保持拼写原样不做处理，对于汉语作者姓名根据表 3‑2中的处理规则进行处理。

表 3‑2 作者姓名拼写形式处理规则

|  |  |  |
| --- | --- | --- |
| 拼写形式 | 目标形式 | 处理规则 |
|  | 知识库中全拼形式 | 剔除多余字符 |
|  | 知识库中全拼形式 | 姓、名颠倒顺序，剔除多余字符 |
|  | 知识库中简拼形式 | 剔除多余字符 |
|  | 知识库中简拼形式 | 姓、名颠倒顺序，剔除多余字符 |

### 作者所属单位抽取

#### 问题分析

作者单位信息是论文作者署名的必要组成部分，在文献检索、信息传递等方面起到了重要作用[]。论文中作者署名单位信息通常包括单位名称、所在地名称、邮政编码、省份名称、国别。而在单位名称中，作者一般会标注自己的所属院系、实验室等部门名称。因此在抽取作者所属单位时，应尽量精确到细粒度的部门名称，以减少后续作者定位、查找的工作量。

在论文中标注单位地址时，不同的作者对同一个单位可能有不同的写法，为此需要根据单位、部门的常见名称写法构建一个部门别名字典。同时作者在标注自己的单位时可能会有漏写字母、前后字母顺序颠倒等拼写错误，因此在抽取作者的所属部门时适合采用粗粒度的基于编辑距离相似度的字符串模糊匹配方法。编辑距离指的是从源字符串转换到目标字符串所需的最小编辑操作成本，可以用来衡量两个字符串之间的相似度。字符串与之间的相似度计算公式定义如下：

(3-1)

其中，分别表示字符串与的长度，表示字符串与之间的编辑距离，其计算公式如式(3-2)所示。

(3-2)

#### 抽取算法描述

在文献数据库存储的记录中，作者单位标注共有两种形式，如表 3‑3所示。当作者地址中仅有作者单位时，表明论文所有作者均从属于同一个单位，在处理时可以统一转换为作者+作者单位的形式，便于处理。

表 3‑3 作者单位标注形式示例

|  |  |  |
| --- | --- | --- |
| 标注形式 | 作者全名 | 作者地址 |
| 作者+作者单位 | Su, Yiming; Dong, Ping | [Su,Yiming; Dong,Ping] Ocean Univ China, Coll Marine Geosci, Qingdao 266100, Peoples R China |
| 作者单位 | Zhang, Yue | Ocean Univ China, Inst Mat Sci & Engn, Qingdao 266100, Shandong, Peoples R China |

在抽取每个作者的署名单位前，首先要从作者地址中提取出每个单位地址及与标注作者的对应关系。由于作者地址中的单位标注有一定规律，可以利用正则表达式进行匹配提取。作者地址中可能有多组作者及对应的标注单位，利用正则表达式(?<=\[)[^]]\*(?=\])提取地址中的作者放入数组中，同样使用正则表达式(?<=\[.+\])[^[]\*(?=\[|$)提取作者标注单位放入数组中，中下标为的作者其对应的标注单位为。

在抽取一篇论文中每个作者的所属单位时，采用如下步骤：

1. 根据作者姓名在数组中进行匹配，并在相应的中找到其标注单位。
2. 根据本单位名称标识判断该作者是否本单位人员。如果是本单位人员，继续后续处理流程；否则标记为外单位人员，返回处理下一个作者。
3. 将作者标注单位根据分隔符（，）拆分成单位名称、部门名称等几部分，并与部门别名字典中的每一个部门名称分别计算相似度，选取满足相似度阈值（本文设为0.6）且值最大的部门名称作为该作者的所属部门。

### 通讯作者标注

通讯作者担负着保证论文可靠性的责任，其重要性不言而喻。虽然通讯作者标注与否不会影响作者查找、匹配的最终结果，但考虑论文作者身份识别的完备性，对通讯作者进行标注是十分有必要的。

通过分析论文记录中的通讯作者地址字段发现，通讯作者标注具有明显的规律特征，一般是作者+(reprint author)+作者单位的形式，而且作者的拼写形式与作者简称或者作者全名字段中的拼写形式一致，如表 3‑4所示。因此可以设计正则表达式(?<=^)[^(]\*(?=\()提取通讯作者姓名，在处理每一个待识别作者生成记录时，根据当前作者的作者简称或者作者全名字段与提取出的每一个通讯作者进行匹配，如果匹配成功则标注当前作者为通讯作者。

表 3‑4 通讯作者标注形式示例

|  |  |  |
| --- | --- | --- |
| 作者全名 | 作者简称 | 通讯作者地址 |
| Zou, Xuemei; He, Qi | Zou, XM; He, Q | Zou, XM; He, Q (reprint author), Ocean Univ China, Coll Marine Geosci, Qingdao 266100, Peoples R China. |
| Li, Wei Jun | Li, WJ | Li, Wei Jun (reprint author), College of Marine Geosciences, Ocean University of China, Qingdao, China |

## 候选作者生成阶段

候选作者生成阶段的主要目标是在候选作者集合尽可能小的情况下包含目标作者。如果候选作者集合中没有包含目标作者，那么就不可能正确的将待识别作者与目标作者进行链接。而为了让候选作者集合中包含目标作者，则可能会把更多的专家或学生作为候选作者加入到集合中。本节主要解决的问题是如何在保证候选作者集合召回率的情况下尽可能减小候选作者集合中的候选作者数量。

### 候选作者查找算法

在预处理阶段，已经为论文的每一个待识别作者提取到了姓名、所属单位、论文出版日期等关键特征信息，接下来需要根据这些特征信息在实体库中查找、匹配作者。对于一个给定的待识别作者，在候选作者生成阶段会为查找产生候选作者集合。由于实体库中没有校外人员信息，因此若待识别作者为校外作者，则生成一条候选作者记录并标记为外单位人员即可。而对于校内作者，按图 3‑4所示流程进行查找。对于经过查找但未在实体库中找到候选作者的校内作者，即，称作者为未登录作者，将其标记为不存在（）。



图 3‑4 候选作者查找流程

在对校内作者进行查找、生成候选作者集合的过程中，采用如下规则进行人员匹配：

1. 论文中作者署名信息与人员库中姓名信息进行匹配。

将在预处理阶段提取到的论文作者姓名与人员库中的人员姓名进行匹配，只有这个匹配规则成立才能确保目标作者有可能是论文的实际作者。由于论文作者姓名存在不同的拼写形式，具体采用如下方法进行匹配：

1. 拼写形式为全写形式（）。在人员姓名全拼字段中利用作者姓名进行精确匹配。
2. 拼写形式为缩写形式一（）。在人员姓名简拼字段中利用作者姓名进行精确匹配。
3. 拼写形式为缩写形式二（）。在人员姓名简拼字段中以作者姓名为前缀字符串进行前缀匹配。
4. 论文中作者署名单位信息与实体库中人员-部门信息进行匹配。

根据在预处理阶段抽取的作者所属单位信息，与人员的所在部门信息进行匹配。同时在单位信息匹配的过程中，还要考虑人员在各部门任职的时间区间与论文的发表时间的匹配。通常署名单位为C的论文是作者在C单位任职期间发表的，但学术论文的发表具有一定的滞后性，可能作者在调任D单位后论文才被发表。此时论文中作者的署名单位还为C，但论文发表时间并不在作者在C单位任职时间区间内。因此本文在考虑人员在单位任职时间区间与论文发表时间的匹配时，将任职时间区间延长3年，只要论文发表时间落在这个区间内的即认为匹配成功。

经过上述匹配过程后，已经为待识别校内作者生成候选作者集合。但集合中可能包含多个候选作者，有必要对候选作者集合进行进一步的过滤，在保留目标作者的前提下尽可能减少集合中候选作者数量。针对不同情况本文采用的过滤方法如下：

1. 多个候选作者其实是同一个人。

由于数据迁移中的少量错误或者学生毕业后留校任职等特殊原因，可能存在同一个人在人员库中有多条记录的少量情况，因此集合中查找到的多个候选作者可能是同一人。针对这种情况，本文根据人员的身份证号、手机号、邮箱信息进行过滤，对于身份证号、手机号、邮箱任意一种方式值相同的候选作者认为是同一人，保留最新记录对应的候选作者。

1. 通讯作者匹配出的候选作者有多个人。

通常在论文投稿时，通讯作者会在论文中写明自己的邮箱等联系方式。因此针对这种情况，可以利用通讯作者标注的邮箱与候选作者的邮箱进行匹配来完成候选作者筛选。

1. 导师或学生有一方匹配出的候选作者有多个人。

一般情况下研究生在发表论文时，由于导师和学生这种特殊的指导关系，导师和学生会同时出现在论文作者列表中。因此针对这种情况，本文利用这种特殊指导关系，在导师或学生有一方只有一个候选作者的情况下，即可间接印证另一方作者身份，实现候选作者的筛选。

### 实验结果与分析

## 本章小结

# 基于合作网络的重名作者消歧

## 相关概念介绍

### 合作网络简介

### 节点相似度

## 作者合作网络的挖掘与构建

### 合作关系图模型建模

### 作者合作网络构建

## 改进的作者身份识别方法

### 候选作者选择阶段

### 节点相似度计算

### 实验结果与分析

## 本章小结

# 论文推送认领服务分析设计

## 后台技术框架简介

### ASP.NET Web API

### Entity Framework

### Autofac

### AutoMapper

## 推送认领服务分析

## 系统设计与实现

### 认领平台

### 推送认领

### 补充认领

## 本章小结

# 总结与展望

## 论文总结

## 研究展望

# 参考文献

1. 魏晓,刘洋.东北大学2006—2015年SCI论文统计分析[J].图书情报导刊,2016,1(12):138-142.
2. 杜春芳.SCIE收录科技论文统计分析——以河南科技大学为例[J].内蒙古科技与经济,2016(10):60-62.
3. 周环,苏莉娜.基于Web of Science的河北大学科研论文统计分析[J].科技情报开发与经济,2015,25(01):130-133.
4. 朱立禄,宋世俊,王琳.国内外机构知识库建设现状及建议[J].现代情报,2017,37(03):109-115.
5. 刘文云,毕煜.基于科研管理需求驱动的机构知识库管理政策——以山东理工大学SDUT IR为例[J].情报理论与实践,2017,40(01):31-35+40.
6. 王睿. 实体链接的研究与实现[D].北京邮电大学,2015.

# 附录

附录 A SCI/EI文献记录字段对应关系及说明

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 字段  名称 | 出版物类型 | 作者全名 | 文献标题 | 出版物名称 | 丛书标题 | 语种 | 会议标题 | 会议日期 |
| SCI字段标识 |  |  |  |  |  |  |  |  |
| EI字段标识 |  |  |  |  |  |  |  |  |

附录 B 本文涉及的部分汉语人名拼写规则

|  |  |  |
| --- | --- | --- |
| 条目 | 规则 | 示例 |
| 5.1.1 | 正式的汉语人名由姓和名两个部分组成。姓和名分写，姓在前，名在后，姓名之间用空格分开。复姓连写。姓和名的开头字母大写。 | Wáng Fang 王芳  Yáng Wèimín 杨为民 |
| 5.1.2 | 由双姓组合（并列姓氏）作为姓氏部分，双姓中间加连接号，每个姓氏开头字母大写。 | Liú-Yáng Fān 刘扬帆 |
| 5.1.4 | 国际体育比赛等场合，人名可以缩写。汉语人名的缩写，姓全写，首字母大写或每个字母大写，名取每个汉字拼音的首字母，大写，后面加小圆点，声调符号可以省略。 | Li X.L. 李小龙 |
| 5.1.5 | 中文信息处理中的人名索引，可以把姓的字母都大写，声调符号可以省略。 | ZHANG Ying 张颖 |

# 致谢

# 个人简历

1992年3月9日出生于山东省聊城市。

2011年9月考入中国海洋大学信息科学与工程学院计算机科学与技术专业，2015年7月本科毕业并获得工学学士学位。

2015年9月保送中国海洋大学信息科学与工程学院计算机应用技术专业攻读硕士学位至今。

# 在校期间研究成果

1. 负责开发《烟草所成果信息管理系统》