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](#_Toc501354883)

[1.1 研究背景和意义 1](#_Toc501354884)

[1.2 研究现状 1](#_Toc501354885)

[1.2.1 作者身份识别研究现状 1](#_Toc501354886)

[1.2.2 作者识别问题难点 1](#_Toc501354887)

[1.3 研究内容和主要工作 1](#_Toc501354888)

[1.4 论文内容安排 1](#_Toc501354889)

[1.5 本章小结 1](#_Toc501354890)

[2 测试数据集构建与特征选择 2](#_Toc501354891)

[2.1 测试数据集的构建 2](#_Toc501354892)

[2.1.1 文献记录的获取与整理 2](#_Toc501354893)

[2.1.2 测试数据集的选取 2](#_Toc501354894)

[2.2 评价指标 2](#_Toc501354895)

[2.3 作者识别的特征选择 2](#_Toc501354896)

[2.4 本章小结 2](#_Toc501354897)

[3 论文作者身份识别初步设计 2](#_Toc501354898)

[3.1 作者识别基本流程 2](#_Toc501354899)

[3.2 作者姓名提取与规范化 2](#_Toc501354900)

[3.2.1 作者姓名拆分 2](#_Toc501354901)

[3.2.2 作者姓名清洗 2](#_Toc501354902)

[3.2.3 作者姓名前后顺序分析 2](#_Toc501354903)

[3.2.4 作者姓名规范化 2](#_Toc501354904)

[3.3 作者所属单位提取 2](#_Toc501354905)

[3.3.1 编辑距离 2](#_Toc501354906)

[3.3.2 相似度 3](#_Toc501354907)

[3.3.3 单位名称对应字典 3](#_Toc501354908)

[3.3.4 作者单位提取方法描述 3](#_Toc501354909)

[3.4 实验结果与分析 3](#_Toc501354910)

[3.5 本章小结 3](#_Toc501354911)

[4 基于合作网络的重名作者消歧 3](#_Toc501354912)

[4.1 合作网络简介 3](#_Toc501354913)

[4.2 基于合作网络的作者消歧方法 3](#_Toc501354914)

[4.2.1 基于合作网络的作者表示模型 3](#_Toc501354915)

[4.2.2 论文作者合作网络构建 3](#_Toc501354916)

[4.2.3 合作者集与论文作者集相似性计算 3](#_Toc501354917)

[4.2.4 重名作者消歧 3](#_Toc501354918)

[4.3 实验结果与分析 3](#_Toc501354919)

[4.4 本章小结 3](#_Toc501354920)

[5 论文推送认领服务设计 3](#_Toc501354921)

[5.1 后台技术框架简介 3](#_Toc501354922)

[5.1.1 ASP.NET Web API 4](#_Toc501354923)

[5.1.2 Entity Framework 4](#_Toc501354924)

[5.1.3 Autofac 4](#_Toc501354925)

[5.1.4 AutoMapper 4](#_Toc501354926)

[5.2 推送认领服务分析 4](#_Toc501354927)

[5.3 系统设计与实现 4](#_Toc501354928)

[5.3.1 认领平台 4](#_Toc501354929)

[5.3.2 推送认领 4](#_Toc501354930)

[5.3.3 补充认领 4](#_Toc501354931)

[5.4 本章小结 4](#_Toc501354932)

[6 总结与展望 4](#_Toc501354933)

[6.1 论文总结 4](#_Toc501354934)

[6.2 研究展望 4](#_Toc501354935)

[参考文献 4](#_Toc501354936)

# 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，科研成果产出数量也逐年增加。而科研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。因此近年来，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]。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。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。

目前，高校科研管理部门在对论文信息进行收集统计、存缴入库的过程中，通常采用如下两种方式：由作者自行提交论文信息，科研管理部门审核后汇总；由科研管理部门收集整理论文信息，交由作者确认后汇总。在论文收集整理的过程中，很多科研人员都有找不到科研成果的经历，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。

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论文信息后，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？如果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，从而导致即便科研管理部门获取了大批SCI论文数据却无法快速准确得到某作者或部门的产出情况。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，成为科研管理部门亟待解决的问题，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1-1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 1‑1 作者身份识别过程示意

命名实体链接（Named Entity Linking, NEL）是自然语言处理中一项非常重要的任务，旨在将文本中出现的实体指称正确链接到知识库中相应的实体，以解决实体的歧义性[6]。而论文作者识别问题作为实体链接在文献检索领域的特化任务，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（实体链接），面临的主要挑战是作者重名现象，可以采用自然语言处理的相关技术来解决，很多研究者都对NLP领域中的实体链接问题进行了研究，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，即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，可以通过建立论文推送认领服务来实现。

### 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

## 研究内容和主要工作

## 论文内容安排

## 本章小结

# 测试数据集构建与准备工作

## 知识库的构建

## 测试数据集的构建

准确构建数据集是论文作者识别分析的关键基础。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，获取到的文献记录格式差异较大。因此构建一个统一格式的文献数据库是首先需要解决的问题。其次，随着大数据时代的来临，文献数据以一种惊人的速度快速增加。在数量如此庞大的文献信息中，由于无效值、缺失值等脏数据的存在，无法直接将其用于作者身份识别系统的测试。本节将详细阐述如何解决上述问题。

### 文献记录的获取与整理

准确构建测试数据集是评价论文作者识别系统的关键基础。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。对于查重处理后的文献记录，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。



图 2‑1 文献记录查重融合数据流图

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。

### 测试数据集的选取

测试数据集是检验作者身份识别方法有效性的重要工具，选择与构建一个合适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. 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；
2. 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；
3. 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；
4. 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

## 评价指标

## 作者识别的特征选择

## 本章小结

# 论文作者身份识别初步设计

## 作者识别基本流程

## 作者姓名提取与规范化

### 作者姓名拆分

### 作者姓名清洗

### 作者姓名前后顺序分析

### 作者姓名规范化

## 作者所属单位提取

### 编辑距离

### 相似度

### 单位名称对应字典

### 作者单位提取方法描述

## 实验结果与分析

## 本章小结

# 基于合作网络的重名作者消歧

## 合作网络简介

## 基于合作网络的作者消歧方法

### 基于合作网络的作者表示模型

### 论文作者合作网络构建

### 合作者集与论文作者集相似性计算

### 重名作者消歧

## 实验结果与分析

## 本章小结

# 论文推送认领服务设计

## 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

# 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》