项目编号：2021048

CALIS 全国农学文献信息中心研究项目
结题报告

项目名称：基于消歧算法的机构命名识别研究——以石河子大学为例

项目关键词：消歧算法；机构；命名识别；石河子大学

项目单位（盖章）：石河子大学图书馆

通信地址：新疆省石河子市北四路石河子大学图书馆

邮编 832000

项目主持人：罗晓玲

联系电话：13999730185

电子邮件：23797552@qq.com

提交日期：2022 年 5 月 11 日

项目结题验收单

专家验收表

<table>
<tr><td>项目名称</td><td colspan="4">基于消歧算法的机构命名识别研究——以石河子大学为例</td></tr>
<tr><td>主持人</td><td>罗晓玲</td><td>职务/职称</td><td colspan="2">副研究馆员</td></tr>
<tr><td>所在单位</td><td colspan="4">石河子大学图书馆</td></tr>
<tr><td>专家意见</td><td colspan="4">2022年5月10日，石河子大学图书馆组织专家，对中国高等教育文献保障系统CALIS全国农学文献信息系中心研究项目“基于消歧算法的机构命名识别研究——以石河子大学为例”(编号2021048)进行了结项验收，专家通过查看资料、听取汇报、提问质询、核实资料，经认真讨论形成验收意见如下：
1.该项目研究方法正确，研究思路明确，立意新颖，提供的反应研究过程实例鲜明、创新实践成果在提高科学研究活动、提高并保存知识储存资源中起到了重要作用。
2.本课题通过总结机构实体名称识别的理论基础和方法，分析作者隶属机构名称因其基本职能、组织结构的发展演变而产生机构名称中英文全称、简称的错写、漏写或书写不规范等问题现状。以石河子大学为例，对石河子大学机构名称的规范数据和命名识别特点进行探讨和分析，构建适应于石河子大学的机构向量，提出了基于石河子大学在短期内相对稳定特征的名称演化识别的消歧算法，实现了石河子大学机构实体间的精准关联。该项目的创新之处在于，基于人工审核构建机构特征词表；设计了一种关于特征词的特殊识别算法——消歧算法，完成对石河子大学整个机构名称形式的有续集中。基于算法和特征词标的搭建，在GO FRONT系统中对石河子大学机构名称中的不同形式进行了验证。验证结果基本满足了本校机构教学、科研等不同成果的存储和利用需求，进一步为高校文献检索、机构评价、统计挖掘和学科影响力评价等活动提供应用基础性数据保障。
3.该项目基于GO FRONT系统，开发了学科分析和机构成果模块，发表相关论文一篇。
验收专家一致认为，该项目完成了任务书规定的各项任务，达到了预期目标，同意通过验收。

（如需要可增加页数）</td></tr>
<tr><td>专家签字</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>职务/职称</td><td></td><td></td><td></td><td></td><td></td></tr>
</table>

基于消歧算法的机构命名识别研究——以石河子大学为例

摘要：本课题通过对石河子大学机构-作者向量的相似度以及机构名称中的更名、合并、拆分与重组关系等多元问题进行分析，构建适应于石河子大学机构名称的机构向量，提出了基于石河子大学机构在短期内相对稳定特征的名称演化识别的消歧算法，从而实现石河子大学各机构实体间的精准关联，进一步为高校文献检索、机构评价、统计挖掘和学科影响力评价等活动提供应用基础性数据保障。

关键词:石河子大学、机构-作者、命名识别、关联、消歧算法

1 研究背景、目的及意义

1.1 研究背景

新兴机构的泛起，传统机构的淘汰、更名、拆分、重组与归并，使同一机构存在一个乃至多个曾用名、相似名称，加之机构全称、机构简称以及不规范的机构名称书写形式等交替使用，导致现有机构名称识别度降低、从属机构和相关机构的组织结构模糊、与其他属性的关联融合难度提高，使管理系统与服务系统之间不能准确记录机构的信息，最终导致每个独立的系统不能集成数据，从而影响数据的有效传输，增加大数据时代的数据挖掘效率和成本。针对机构重名、别名、名称繁简变化以及名称变更等现象，开展机构名称的规范化识别研究是当前亟待研究的课题。在大数据日益发展的今天，不同高校利用或共享大数据资源，采取了适用与本身机构名称识别的严格规定，精确识别与定位科研实体，消除机构名称中的不同表现形式，建立统一和规范化的机构名称，实现对高校机构名称的规范控制。

1.2 研究目的

本课题以基于地区高校机构的名称识别与消岐为研究对象，以面向历史沿革变迁，消除简称/代称尤其是新疆兵团地区——石河子大学机构特有称谓间的差异为目标，通过分析大量的数据实例，深入研究机构名称的构成特点，对石河子大学的机构名称从词性统计与构词方式两方面进行分析总结,进而为精准的文献计量分析、机构评价、文献统计等提供数据和决策支撑。

1.3 研究意义：

1.3.1 理论价值

对机构命名识别的构建研究，可以提升以机构名称为联接点的信息检索、统计分析、计量评价等活动的可信度，有效解决信息检索、计量评价等科研活动中机构名称著录混乱、层级结构模糊的瓶劲问题。从而提高高校科研机构人员对复杂和非对称的高端技术科研情报感知能力，对高校机构的信息情报服务提供参考依据和决策支撑。

1.3.2.应用价值

本课题针对机构名称识别与机构名称中的歧义现象和消歧算法进行研究，试图对机构实体名称进行正确的识别，进而消解机构名称的异质性，提高信息检索的查全率以及科学计量的信度。为此，构建石河子大学名称识别的消歧算法，实现石河子大学各机构实体间的精准关联，满足高校机构教学、科研等不同成果的存储和利用需求，带动其他文献要素的整合与组织、聚类分析以及机构之间的导航与统计评价,为更准确的反应评估对象的真实水平,科学开展学科影响力评价，更好地发挥科学文献中作者机构的作用。便于科技成果交流与共享等活动的顺利进行。

2 研究内容及方法（思路、方法、具体内容）

2.1 研究思路

本研究从大数据时代情境下文献作者实体机构信息检索、计量评价等科研活动中机构名称著录混乱、层级结构模糊的瓶劲问题出发，提出研究目的和研究意义；总结机构实体名称识别的理论基础和方法，分析作者隶属机构名称因其基本职能、组织结构的发展演变而产生机构名称中英文全称、简称的错写、漏写或书写不规范等问题现状，构建基于规则与算法相结合的机构命名识别方法。以石河子大学为例，在分析石河子大学图书馆命名规则的基础上，对石河子大学机构名称的规范数据和命名识别特点进行探讨和分析，以其能够有效实现石河子大学机构数据的集中汇聚和资源整合的完全利用。

2.2 研究方法

2.2.1 文献调研法：通过人工检索在万方、中国知网、Elsevier、Web of Science等网络上查找相关文献，以及其他社会信息服务机构网站、图书馆（国家图书馆、大学图书馆等）书目检索系统的调查，了解国内外各类机构名称的不同表现形式，为本论文的研究提供基础材料。

2.2.2 案例分析法：案例分析法主要体现在综合考察国内外机构名称的规范研究项目实践，分析其在实践过程中构成的规范文档、机构数据库的数据结构、机构名称处理及命名规范方法以及研究项目的发展、研究成果应用情况等。

2.2.3 文献计量方法：文献计量学是基于文献体系和文献相关媒介为研究对象，利用数学，统计学，计算机科学和系统科学等方法来科学分析数量分布的研究，以及科学技术动态特征进行研究的一门学科。通过利用这种基于数学及统计学的定量分析的方法，对机构名称进行频次统计分析，对机构-作者实体间的相似度进行计算，以实现机构名称的归一。

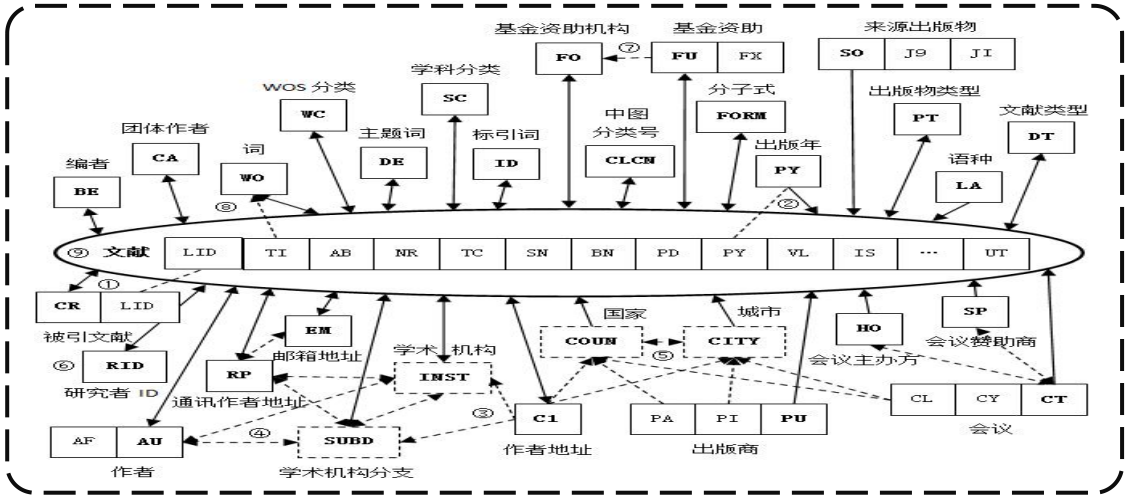
2.3 研究内容

科研机构是海量数字资源中的主要科研实体之一，对机构命名进行识别是开展高校机构评价的前提。通过证实消歧义算法对高校机构的命名特征以及识别关系的有效性，为本课题的工作提供了有益的思路。课题内容主要包括：

2.3.1 调研国内机构命名识别的研究状况：调研国内高校机构的名称规范，对各大高校机构名称/命名的收词特点，共性异性做以比较，归纳总结我国高校机构目前需要解决和研究的问题。

2.3.2 构建基于规则的特征词表（石河子大学为例）。

即通过人工构建筛选规则，逐步缩小同一机构名称的集合，进而确定指向同一机构实体的名称集合。其中规则大多数依赖于地区、邮政编码、机构类型、基金等附加属性以及机构名称关键词的语言学特征等，并基于人工审核构建石河子大学机构特征词表（见图）。



2.3.3 设计石河子大学机构名称的消歧算法：

消歧框架和算法：以 WOS、EI、CSSCI 、CSCD 数据库中的论文为数据源，依次构建机构-作者向量与机构-年度向量（机构-年度向量示意图）；依据更名、合并与拆分等特定的机构形式所特有的名称映射关系与时间分布规律筛选潜在的“匹配机构对”；通过归并邻近几年的机构-作者向量以减少人员流动对实验结果的影响，并进行相似度计算。机构-年度向量 T = (t1999, t2000, …, t2014, t2015, count, flag) 构建算法如下。

（1） 机 构 - 年 度 向 量 中 包 含 17 个 年 份 属 性

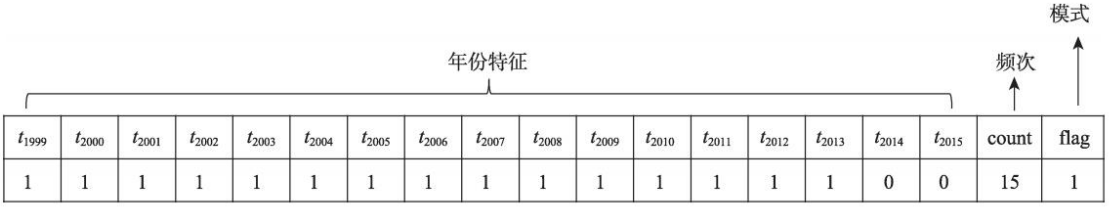
（t2010~t2020），将所有年份属性初始化为 0。

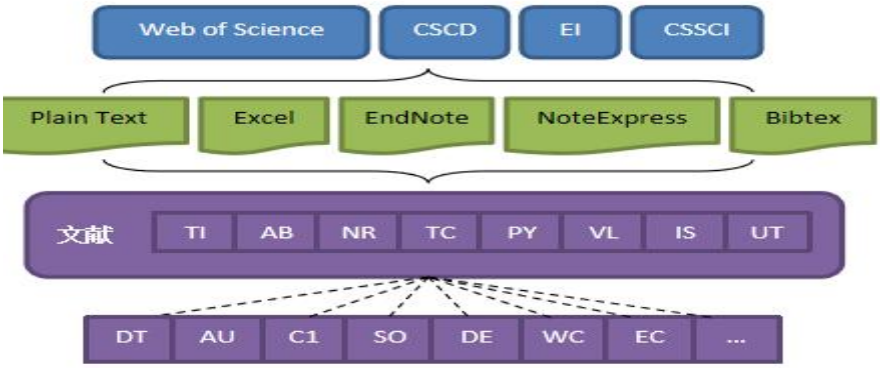
2） 遍历机构-作者向量，若某机构第 i 年作者向量的“规模”不为零，则将该机构的机构-年度

向量中对应 ti 值置为 1。

3） 统计各机构的机构-年度向量中 ti 为“1” 的频次，记为 count。

4） 根据年度向量中 ti 的分布特征，生成“模 式”（flag） 属性，值域为{0,1,2,3}。





2.4 创新点

2.4.1 构建基于机器学习的石河子大学特征词表，同时对多源异构数据下高校机构的名称进行有效清洗及提取。

2.4.2 基于文献计量学及统计学的定量分析方法，对机构名称进行聚类和识别分析，对石河子大学机构名称实体间的相似度进行计算。

2.4.3 将同一名称的不同演化名称进行归一，建立石河子大学机构规范名-别名映射表，采用消歧算法完成对石河子大学图书馆整个机构名称形式的有续集中。

3 结论

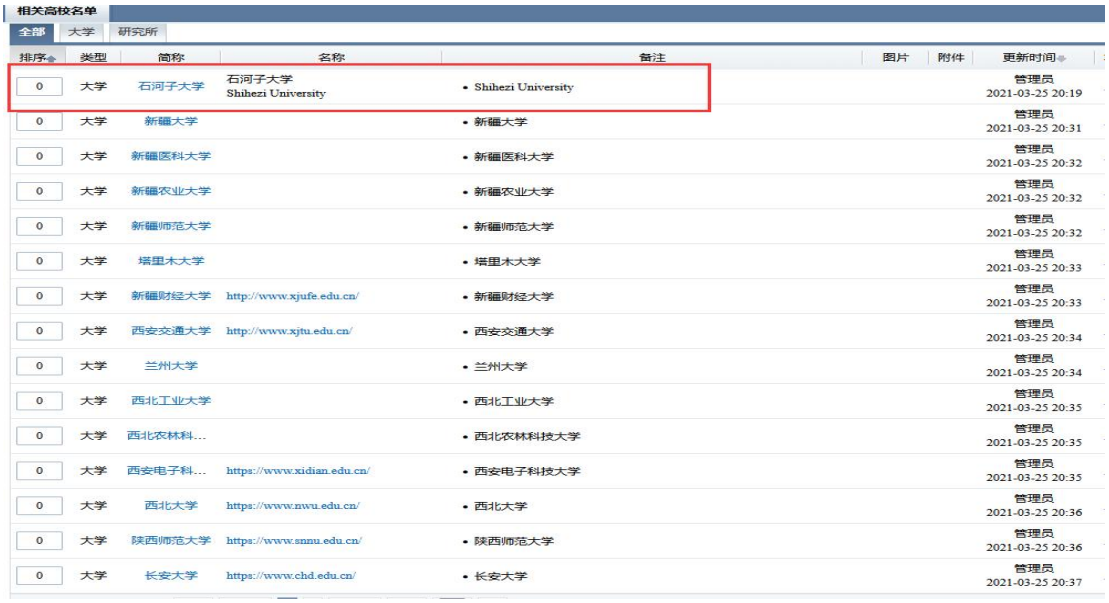
本文主要讨论机构名称命名识别的方法，并基于此方法辅助石河子大学及二、三级机构名称的扩展，为后续机构数据库的构建和作者成果认领奠定了基础。通过语料训练的词向量，构建适应于石河子大学机构名称的科研机构-作者向量，结合石河子大学机构-作者-年度向量的相似度与映射规则识别机构间的更名、合并与拆分等关系；同时，考虑到机构规模是判断重名风险效应量的关键因素，辅之以作者绝对共现量在系统中（Go Front）进一步精准识别结果集。为消除各种原因引起的机构名称变体间的歧义，并将其与机构实体建立精准映射，从而在系统中（Go Front）应用中实现该实体所有信息的快速定位。

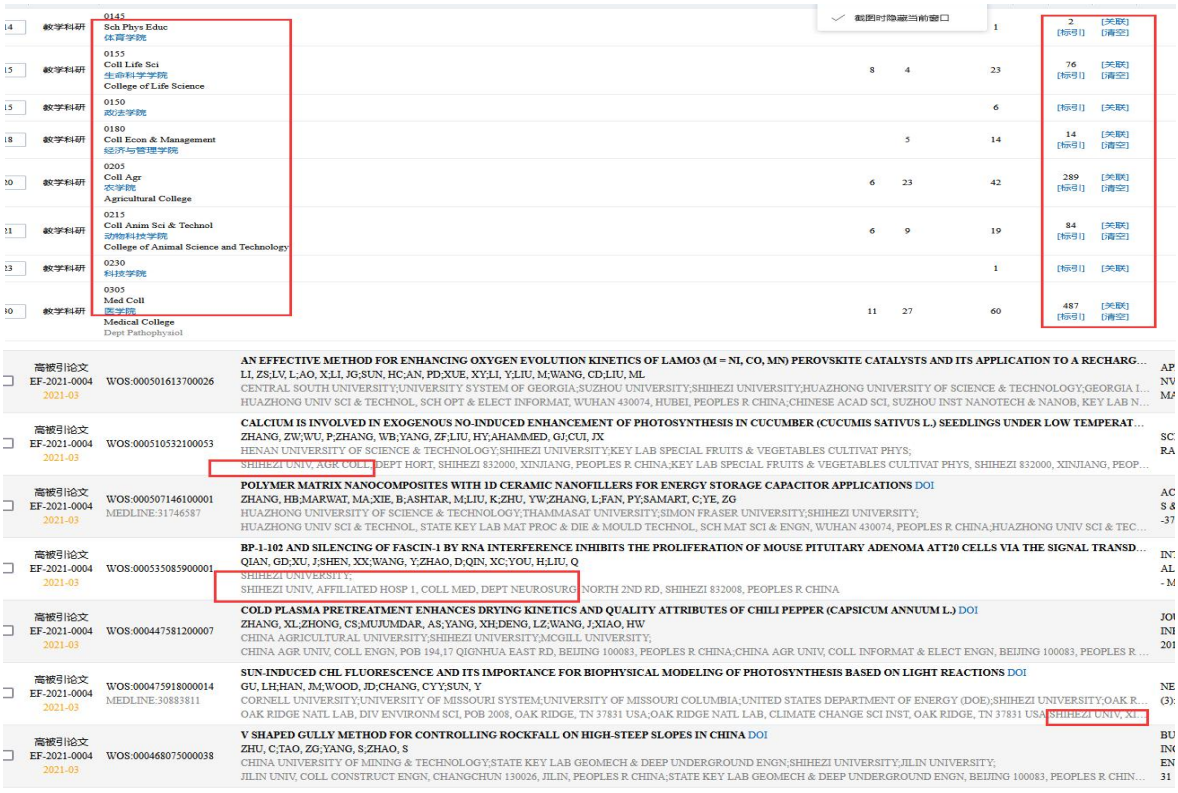
4 项目成果（发表的文章、开发的软件、取得的实践效果等）

1.开发了以石河子大学为核心的科研产出仓储管理系统（Go Front 系统）。该系统在我校各学院及图书馆各部门之间进行了有效的管理与应用，摸清了本校学术成果基本情况，保存保护知识资产，建立了石河子大学科研学术的历史载体。并对石河子大学机构的所有科研产出做集中展示，实现一站式检索，完整地收集、组织、

保存本校的各类学术成果，并对成果进行多维度的展示，提高学校学术成果的可视度。

通过 Go front 系统梳理出表中石河子大学机构名存的多种变体形式。结果表明，该系统对于作者发表文献中署名单位（以不同机构变体名称不同形式出现）的识别率可达 96%。在面对机构知识库中的海量题录实体数量时，通过 Go front 基本上展现文献题录数据的全貌以及作者-机构名称的精准关联。





2.发表文章 1 篇（外审中）：

罗晓玲,陈嘉勇,蓝文钦.信息组织和信息检索之无力论在信息服务中的应用[J].图书情报工作外审中。

5 参考文献

[1]张冬梅. 基于构成模式的中文机构名识别[D]. 北京:北京师范大学,2010.

[2]付晓梅. 高校机构名称归一化研究[D]. 山西:山西大学,2017.

[3]柴宝杰. 中文自动分词若干技术的研究[D]. 河北:燕山大学,2007.

[4]贤信. 机构规范文档结构及构建方式研究[D]. 北京:中国科学技术信息研究所,2015.

[5]胡万亭,杨燕,尹红风,等. 一种基于词频统计的组织机构名识别方法[J].计算机应用研究,2013,30(7):2014-2016.

[6]冯冲,陈肇雄,黄河燕. 采用主动学习策略的组织机构名识别[J]. 小型微型计算机系统,2006,27(4):710-714.

[7]麦合甫热提,米日姑·肉孜,麦热哈巴·艾力,等. 基于语法语义知识的维吾尔文机构名识别[J]. 计算机工程与设计,2014(8):2944-2948.

[8]滕青青,吉久明,郑荣廷,等. 基于文献的中文命名实体识别算法适用性分析研究[J]. 情报杂志,2010,29(9):157-161,169.