情报工程 ISSN: 2095-915X
第1卷 第3期 31-37 2015年6月
TECHNOLOGY INTELLIGENCE ENGINEERING
ISSN: 2095-915X Vol.1 No.3 31-37 Jun 2015

doi:10.3772/j.issn.2095-915x.2015.03.004

研究者唯一识别及其在专家档案系统中的实施

陈田田，吴广印
（中国科学技术信息研究所 北京 10038）

摘要：当前的情报检索系统中还存在研究者唯一标识问题，及由此带来的基于作者姓名检索的查全/查准等问题。本文在梳理了国内外对研究者唯一识别问题的研究及分析了当前国内研究者身份识别应用的局限性及不足后，阐述了自助式情报管理与服务系统通过建设专家库系统，并由外部系统利用专家库系统中的专家工作经历字段构建带有唯一标识的学者词表，为检索系统提供接口服务，最终实现了研究者身份识别并解决了文献查全/查准的问题。最后对专家库系统的技术和应用特点进行了总结。

关键词：研究者唯一标识，专家档案系统

中图分类号：G250

文献标识码：A

Study of Researchers' Identification and its Implementation on the Expert Archiving System

CHEN tiantian, WU Guangyin
(Institute of Scientific and Technical Information of China, Beijing, 100038,China)

Abstract: Currently, there still exists the problem of researchers' identification and the resulting problems of recall and precision based on the retrieval of author's name in the Information Retrieval System. After combing the study of researchers' identification at home and abroad and analyzing limitations and shortcomings of domestic application of researchers' identification ,this article elaborates that through the build of expert

基金项目：本项目得到国家科技支撑计划项目科技文献动态数字出版技术研发与应用示范（项目编号 2012BAH90F03）支持。
作者简介：陈田田（1991-）女，中国科学技术信息研究所硕士研究生；吴广印（1965-）男，中国科学技术信息研究所研究员，北京万方软件股份有限公司董事长。研究方向：云计算、知识组织、大数据挖掘与分析。

database system in self-information management and service system, then external system can make use of the expert work experience field in it ,providing interface service for the IR system, to build scholars thesaurus which with identification.Finally, this method realized the researchers' identification and solve the problem of recall and precision.At last we conclude the technology and application features of expert database system.

Keywords: Researcher identification, expert filing system.

1. 引言

人名识别问题在科学研究中的阻碍作用越发凸显并受到广泛关注。解决网络环境下的研究者唯一识别问题，是当今社会人们适应大数据时代环境深刻变化的需要。然而以大量文献资源的堆砌为特征的情报服务系统，仅能满足用户的基本检索需求，不能有效的解决人名识别问题及由此带来的基于作者姓名的文献查全 / 查准、学术评价等问题。

目前，关于研究者唯一识别问题研究的较多，包括为每位注册研究者提供一个国际唯一识别号、构建机构词表等解决方法。但是这些方法仅对研究者身份识别起到一定作用。在当前情报检索系统中，通过对人名或是分配给作者的国际唯一识别号进行检索，仍然不能实现已有文献系统的基于姓名检索的学者与其科研成果的准确关联问题。

一种可行的解决方案是通过专家库系统的建设构建学者词表，能在很大程度上解决作者重名问题。基于此，本文在梳理了国内外对研究者身份唯一识别问题的研究及分析了当前国内研究者身份识别应用的局限性及不足后，阐述了自助式情报管理与服务系统通过建设专家库系统，并由外部系统利用专家库系统中的专家工作经历字段构建带有唯一标识的学者词表，为检索系统提供接口服务，实现了研究者身份识别，并解决了文献查全 / 查准的问题。最后对专家库系统的技术和应用特点进行了总结。

2. 研究者唯一识别问题研究

研究者唯一识别问题是一个伴随着文献数字化的产生而产生的问题，已经成为检索系统、自然语言处理和信息抽取等应用中亟待解决的问题。检索系统中同一作者姓名对应着现实世界中不同的实体人物不仅不利于对机构成员科研产出的统计分析，也难以将科研成果与作者准确的对应。在搜索引擎中体现为查全率和查准率的低效，降低了使用搜索引擎的满意度；在学术传播中表现为加剧识别科学界研究同行、发现相关学术圈及广泛地合作的难度，阻碍学术交流和知识共享。

2.1 国外相关研究

2.1.1 ReacherID

2009 年汤森路透推出的 ReacherID 注册服务是对科学研究领域中研究者唯一识别问题提供的一种新的解决方法。通过注册机制 [1] 它给每个成员赋予全球唯一的身份标识符，研究者能对他们的全部出版文献进行管理及通过该平台提供的研究者信息，确定可能的合作者。同时，该平台将研究者的 Researcher ID 信息与 Web of Science 进行整合。每个作者被分配了唯一的标识码并对其发表的文章、书籍等相关学术成果进行匹配，其他浏览者可以对已经高度匹配了的作者及其著作关系进行检索，加速了作者的研究成果的广泛传播。不仅如此，汤森路透 [2] 还积极参与基于 ReacherID 的延伸服务研究——学术唯一标识实

验室。通过建立研究者与机构、基金、期刊会议等之间的相互联系，向科学界展示研究者的研究方向、研究成果及在学术界的影响大小。

2.1.2 ORCID

开放研究人员和贡献者 ID(ORCID) 是美国特拉华州的一个非盈利机构国际组织，成立于 2008 年。ORCID 的设计原理 [3] 和数字目标识别系统 (Digital Object Identifier) 类似，它为每个注册的研究者提供可以进行身份唯一识别的字符编号。该平台可以有效规避因姓名排序、姓名缩写及人事关系变动等因素导致的科研成果与研究者不能准确匹配的问题。

随着科研成果的开放获取不断成熟，ORCID 提供的学者信息 [1] 能很好的记录了参与科学研究相关科技工作者的研究思路、动态、走向，给在相关领域同时进行研究的人员提供相互交流借鉴的平台。可以说，一些在向科学进军的大道上相关必要的知识储备、经验的积累等隐性知识就能通过该系统向其他科研人员展示出来，缩短了学术信息传递的周期，极大的激发了研发人员进行学术创造的热情。

2.2 国内相关研究

2.2.1 构建人名 – 机构层级词表

中国科学技术信息研究所的张建聪等 [4] 从知识导航的角度对海量的信息资源进行分析后，总结出来用以描述信息资源的五个代表性知识节点：人物、机构、主题、关键词和基金，并指出机构要素是知识导航要素中比较典型的一个，即通过将机构与人名建立关联的方式识别出同名人中真实存在的不同实体。在机构长期的发展演变中 [5]，多数机构经历了重组、改名、合并等变革，机构名称也随之改变导致对应的机构成员不能得到唯一性识别。万方软件公司建立的“机构多层级词表”模型总结了机构名称多样性的多种情况：合并名称、曾用名称、机构简称、机构独立名称及非规范名称，并采用叙词表编制模型中的“用”、“代”、“属”、“分”、“族”的体系对上述各种名称映射到“机构多层级词表”中，通过机构与人名的准确匹配，解决了机构—人名识别问题。

2.2.2 统一出版机构和数据库的元数据

中国科学工作者广泛活跃在世界舞台，取得了卓越的成果。2013 年，我国的 SCI 论文数量已达 23.14 万篇，位居世界第二 [6]。越来越多的中国论文被国际数据库收录，作者姓名伴随着论文的收录也被记录在数据库中。但是不同国家形成的不同语言风格，在表达姓名信息时产生了巨大的中西方差异。由于国际上缺乏对作者姓名的元数据规范，论文中中英文作者姓名信息不一致现象时有发生 [7]。

国内作者文章被国外期刊收录后，作者姓名信息的标注依期刊的不同而不同；国外大型数据库对姓名标注的方法也“各自为政”。鉴于国内数字出版机构和数据库存储了大量国内外科研成果，中国地质大学 (武汉) 姚戈等 [8] 提出在国内通过规范出版机构和数据库收录方的出版和提交信息格式，以此对著者姓名歧义现象进行控制。

2.2.3 构建中国研究者标识系统

目前，随着研究成果及科研信息的公开化趋势不断加强，构建公共性的研究者标识系统成为学术界共识。成立于 2009 年的 ORCID 项目与其他研究者标识系统相比更加中立化、国际化。为了更好地解决中国研究者姓名标识问题，中科院文献情报中心与 ORCID 合作 [9] 于 2014 年 10 月推出 iAuthor, 即中国科学家在线 (http://iauthor.cn)，努力推动在中国科学院乃至全国赋予中国研究人员一个国际化的唯一标示符。iAuthor 简单地通过三步为研究者提供识别服务。首先，iAuthor 帮助研究者注册一个 ORCID 号或关联到已有的 ORCID 号。其次，帮助研究者创建一个永久的个人主页。最后，支持研究者在 iAuthor 和

ORCID 之间同步个人信息和著作信息，轻松获取研究者使用 ORCID ID 发表文章和相关的数据集的信息。目前，iAuthor 已经在中国科学院 100 多个研究所得到应用，注册用户数量超过 15000 人 [10]。

2.3 国内研究者身份识别应用的局限性及存在的不足

尽管目前国内对研究者唯一标识问题的研究中，已经产生具有一定影响力的成果（如 iAuthor 等），但是从情报检索的查全 / 查准率角度来看，还存在着一定的局限性。

以 iAuthor 中国科学家在线 (http: //iauthor.cn) 为例，系统平台为注册后的学者发放一个国际唯一的身份标识号解决作者识别问题，并通过与基金申请等科研工作流进行集成，促进研究工作的识别。学者在登录该平台后即可对自己的信息进行维护。但是国际唯一身份标识号不能实现对已有文献的关联。具体来说，尽管学者已经获得一个唯一身份标识号，从研究者自身的角度来看，实现了身份的唯一识别，但从信息检索的角度来看，已有的文献信息元数据中不包括作者唯一标识号，因此在当前的情报检索系统中仍然不能直接通过检索唯一身份标识号解决文献与作者的精确匹配问题。

同时，在国内学术数据资源库中，对人名歧义的处理还不尽人意，存在着不足。

一方面，对于机构名称不规范问题没有很好的处理。例如，在万方学术搜索平台检索姓名为“李小军”的学者，得到返回结果中有“同名”的 229 人。检索返回结果中，机构单位有“北京海淀区中国地震局地球物理研究所”、“中国地震局地球物理研究所”等。实际情况是，这是学者隶属机构用名不规范导致的人名歧义，而当前的学术资源库中不能很好的处理这类问题。

另一方面，对于同一学者隶属多个工作单位不能实现信息的整合。以学者“李小军”为例，在知网中检索姓名为“李小军”的学者，返回记录为 2 条，分别为隶属单位“中国地震局工程力学研究所”和“中国地震局地球物理研究所”。这两个机构是该学者在人事关系变动中产生的，学者信息却不能在同一人名环境下进行同步维护。

因此在这样的检索环境下，检索结果的查全 / 查准率就很难得到保证。

3. 自助式情报管理与服务系统中专家库系统的建设

自助式情报管理与服务系统 [11] 是一种企业级平台系统，本身集合了传统的情报检索服务、情报评价、用户沟通等功能，为用户提供了传统的文献资源及实时更新的网络资源等服务。

通常来说，查全率和查准率是评价情报检索系统的两个重要指标 [12]，而科研产出与学者之间的准确关联是解决基于作者姓名检索查全 / 查准问题和学术评价问题的基础。因此，系统通过专家库系统构建的学者词表为识别专家提供保证。同时系统还为注册用户提供专家库系统中的学者信息终身维护机制和系统内唯一的学者 ID 号，这些都为完善专家库系统建设及利用专家词表提供基础。

此外，用户在系统中的互动性也会促进情报服务发挥作用。 用户互动层面包含情报评价和用户沟通。用户可对信息源做出评价、转载、上传发布，不仅使自身信息需求得到满足，用户参与的评价信息也为其他用户提供参考。因此系统提供了多种类多渠道的用户互动机制，如邮件沟通、私信、公共会话、FAQ、博客等。

自助式情报管理与服务系统涉及众多功能，由于篇幅限制其他功能不再赘述。下文仅从该系统中的专家库系统的建设阐述其在身份识别中的

应用。

3.1 专家库系统概况

在科研活动中，人是参与知识活动中的主体。理清人物活动及其社会关系才能准确的评价出其在学术界中真实影响力及对社会的价值。专家库旨在建立学者科研档案库，利用专家库系统构建的学者词表可实现对专家的身份识别，具体实现参见下文说明。传统的情报服务系统向用户展示的专家信息大多来自于计算机的自动生成，专家信息较为单一且准确性差强人意。专家库中学者信息由学者本人维护，且涵盖了专家个人基本信息、教育经历、学术经历等，包含了专家基本信息、简历及专家评议、教育及工作经历、获奖情况及荣誉称号、项目承担及社会兼职、科研产出及其他（包括获奖情况、专利发明、论文、论著等）、指导研究生情况、会议论文情况及记录辅助信息等信息，为用户呈现了立体的学者形象。

3.2 基于专家库系统学者唯一 ID 的应用

专家库系统是自助式情报服务系统实现研究者唯一标识的基础。外部系统利用专家库系统中的专家工作经历字段集，构建学者词表并且对外提供接口服务。系统为每位专家分配一个系统内的唯一学者 ID 号，此 ID 号可用于系统内检索的入口词。学者 ID 号在数据库中对应为作者姓名与其相关单位的组合查询，实现了对同一学者在不同工作单位时期文献的全部获取。这对应的解决了现实中一个学者对应多个工作单位及属于不同单位的研究者唯一标识问题。

以检索学者“李小军”的相关信息为例，系统实现身份唯一标识过程的流程图见图 1。检索系统在接收到检索请求后，通过调用接口服务中提供的学者词表，将返回具有全部重名作者的学者 ID 号和单位信息列表。用户选择其中一个学者 ID 号，如选择“李小军 01”，再将 ID 号传递给学者词表，此 ID 号对应在数据库中的数据库中

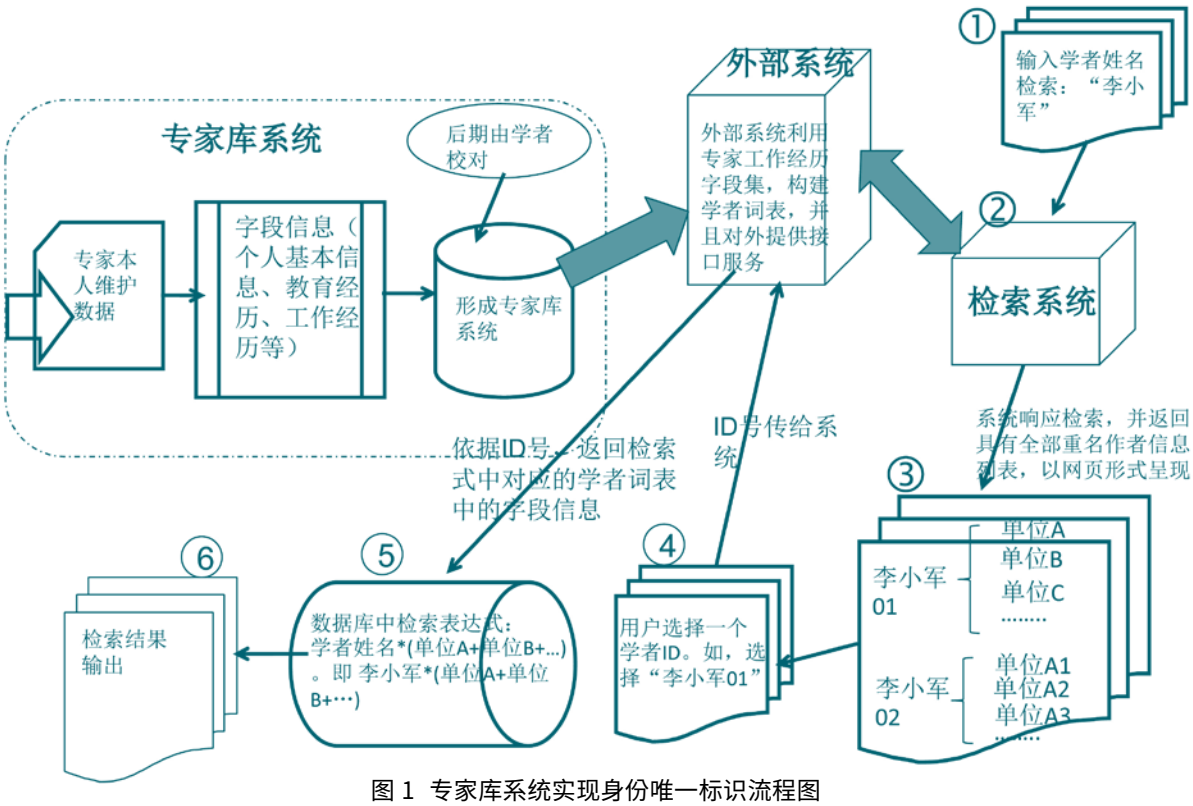

图 1 专家库系统实现身份唯一标识流程图

检索表达式：学者姓名 *(单位 A+ 单位 B+…)，对应在本次检索中为：李小军 *(单位 A+ 单位 B+…)。最后，检索系统依据此检索式对应的学者词表中的字段信息，形成检索结果集并将结果输出。

专家库系统后期由学者对其校对，解决了系统中机构名称不规范问题，规避了检索中因机构名称不规范问题导致的检索结果的缺失 ；同时，由于学者本人对自己的信息进行维护，及时更新专家本人隶属单位信息，实现了对同一学者隶属多个工作单位的信息整合。解决了上文中提到的现有情报检索系统中存在的不足。因此，这样的一个专家库系统在与外部系统集成后，对于改进情报检索系统的查全 \ 查准问题及基于作者的学术评价问题有着重要意义。

4. 专家库系统的技术和应用特点

4.1 专家库系统的技术特点

该系统涵盖了对专家完整科研信息的管理，其中涉及字段集、多值等非结构化数据库管理特性，数据库存储系统采用 MySQL 的最新版本 (MySQL Cluster V7.3) 支持云存储 . 其中每一字段和字段集为一个独立表格，支持海量数据存储。搜索引擎采用 Apache Solr 支持分布式多核索引；另外该系统也可和万方软件自主开发的云搜索引擎 RMSCloud 进行集成 (RMSCloud 是“863”中国云专项的科研成果，已通过国家正式验收并获得专家好评，多项技术和应用填补我们国家的“科研空白”)，支持基于云计算架构的云接口服务。

4.2 专家库系统的应用特点

现有科技文献服务系统大都是通过文献服务机构对各类文献数据的搜集和加工建设而成的，其完整性是可想而知的。开发本专家库系统的一个目的也是为了解决文献系统建设所面临的一些问题。本系统建设的出发点是让学者自行维护自己的科研档案，“谁的孩子，谁抱养”。该系统支持数据中心建设模式，构建全国或区域性专家数据中心，各机构自行认证，学者自建，具有分级管理的功能；系统分超级管理员、机构审查员、学者用户三类角色，机构审查员负责机构内学者档案的审查，学者只能按分配的唯一标识维护和添加自己数据，终身维护。另外该系统还可借助万方软件现有的文献仓储和机构词表为学者预填一些数据，然后由学者进行校对和补充维护，大大提高机构、学者的维护效率。同时一些授权用户还可以通过系统提供的接口服务抽取不同的数据项构建自己的文献服务系统，这一服务模式对传统文献建设模式是更高层次的创新。

该系统包含科研人员基础信息、同行评价、专家评价、教育经历、工作经历、学术兼职、获奖情况、项目承担情况、科研产出等科研档案信息。每一类信息都是一个数据集合，包含详细的子项，可以作为一个独立的文献数据库使用。或者说可以利用专家档案库来提取各类文献数据库，包括科技报告、成果、专利等科技文献数据流。通过专家工作经历、曾用名、英文名等信息构建研究者唯一识别用知识库，利用该知识库可有效识别具有同一姓名的不同专家，然后为专家分配一个系统内的唯一学者标识 (也可转化为对应的标准化标识)。这一唯一标识将成为学者检索、分析、评价的核心基础。

4.3 专家库系统的应用推广

我们在开发本专家系统的同时北京万方软件股份有限公司的评价中心已经完成了基于本系统数据结构的专家评价指标体系的建设工作，可用于行业专家的评价和推荐工作，对于机构用户也可用于本单位的科研绩效和支撑评价工作。

自助式情报服务与管理系统的建设涉及众多的新技术和新模式，本文不做详细介绍，读者可

参考作者在《情报工程》2015 年第一期的“自助式情报管理与服务系统的研究与实现”一文。系统的详细功能和服务也可访问 :http://sbis.sciinfo.cn/sbis 了解详情。

结束语

在数字信息环境中更好的满足用户的知识服务，是当前情报服务系统领域关注的重点。而一切与“人”有关的知识服务都必须建立在学者身份唯一识别的基础上。国外的 ORCID、ReacherID，国内 IAuthor 及自助式情报管理与服务系统的开发与研制是对数字信息化时代学者身份唯一标识的迫切性及重要性的验证。ORCID、ReacherID 及国内的 iAuthor 通过为每位研究者提供一个唯一标识号，实现对研究者身份的识别。然而通过作者的唯一标识符仍不能解决基于作者姓名检索的文献查准问题；自助式情报服务管理与系统通过专家库系统的建设，构建学者词表并向外提供接口服务，达到识别出重名的作者。同时为专家分配的系统内部的学者唯一标识号可以进行学者检索，进而实现了作者与文献信息的精确匹配及文献查全 / 查准等问题。

参考文献

[1] 李颖 , 徐硕 , 姚长青等 . 研究者标识系统的整合及其应用 [J]. 中国科技资源导刊 (中国信息导报), 2014(5):90-94.DOI:10.3772/j.issn.1674-1544.2014.05.015.

[2] 窦天芳 , 张成昱 , 张蓓 , 邹志华 .ResearcherID 现状分析及应用启发 [J]. 图书情报工作 , 2013(4):40-45.DoI:10.13266/j.issn.0252-3116.2014.04.007.

[3] 何巍 . 由著作者身份识别系统引发的思考 [J]. 科技导报 (北京),2010,28(9):120-120.

[4] 张建聪 , 吴广印 . 面向知识导航的机构要素元数据规范及互操作 [J]. 情报学报 ,2010,29(1):84-92.DOI:10.3772/j.issn.1000-0135.2010.01.013.

[5] 杨奕虹 , 李雅萍 , 张立丽 , 林霄剑 . 机构多层级词表的编制及在文献计量评价与科研绩效管理中的应用 [J]. 数字图书馆论坛 ,2013(6):57-63.DOI:10.3772/j.issn.1673-2286.2013.06.010.

[6]SCI 论文考核高校 [EB/OL].(2015-03-02).http://news.sciencenet.cn/htmlnews/2014/10/304822.shtm.

[7] 于双平 , 王文武 , 姜晓舜等 . 科技论文的中英文作者信息不对称现象 [J]. 中华医学图书情报杂志 ,2010,19(6):74-76.DOI:10.3969/j.issn.1671-3982.2010.06.024.

[8] 姚戈 , 王淑华 . 科技期刊著者姓名规范控制及身份识别分析和探讨 [J]. 中国科技期刊研究 ,2015,26(1):41-46.DOI:10.11946/cjstp.201409250932.

[9]iAuthor 与 ORCID: 中国科学院文献情报中心支持中国研究人员获得国际识别号 [EB/OL].(2015-03-15).http://orcid.org/blog/2014/12/03/iauthor 与 orcid- 中国科学院文献情报中心支持 中国研究人员获得国际识别号 ?lang=orc.

[10] 孙坦 , 黄金霞 , 张建勇等 . 科学家国际化识别研究 [J]. 图书情报工作 ,2015,59(1):17-22,44.DOI:10.13266/j.issn.0252-3116.2015.01.002.

[11] 吴广印 , 徐飞 , 冯丹等 . 自助式情报管理与服务系统的研究与实现 [J]. 情报工程 ,2015(1):8-15.DOI:10.3772/j.issn.2095-915x.2015.01.001.

[12] 吴广印 . 浅谈 " 查全 / 查准率 "[J]. 数字图书馆论 2010,(7):30-34.DOI:10.3772/j.issn.1673-2286.2010.07.007.