学者智能目录体系构建研究*

郑杨[1] 石进[1]

（1.南京大学，信息管理学院，江苏南京，210023）

摘要：[目的/意义]知识经济时代，学者的同领域及跨领域合作已成为普遍现象。但知识需求者在搜集相关领域学者信息时往往会碰壁，因此，急需构建完善的学者智能目录体系，促进知识主体之间的合作。[方法/过程]分析现有学者目录构建研究和学者信息检索工具的利弊及其对构建学者目录的启示；为学者智能目录体系的构建提出建议。[结果/结论] 学者智能目录体系相对于传统的学者信息检索工具，具有以用户为中心、优化知识管理模式等优势，促进了知识主体之间的合作、实现双赢。

关键词：人工智能 学者目录 信息组织 知识管理

Research on the Construction of Intelligent Bibliography System for Scholars

Zheng Yang[1], SHI Jin[1]

(1. School of Information Management, Nanjing University, Nanjing 210023, China)

Abstract: [Purpose/Significance] In the era of knowledge economy, cross-domain cooperation among scholars has become a common phenomenon. However, knowledge demanders often run into a wall when collecting information about scholars in related fields. Therefore, it is urgent to build a perfect intellectual catalog system for scholars to promote cooperation among knowledge subjects. [Method/Process] Analyze the advantages and disadvantages of existing scholars' information retrieval tools and its enlightenment to the construction of scholars' catalogue; This paper puts forward some suggestions for the

construction of intelligent catalog system for scholars. [Results/Conclusion] It is necessary and feasible to construct an intelligent directory system for scholars, which is conducive to cross-domain cooperation among knowledge subjects and to achieve a win-win situation.

Keywords： artificial intelligence； bibliography for scholars； organization of information； knowledge management

分类号：G257

***基金项目**：本文系国家社会科学基金项目“面向国家安全的情报研究”（18FTQ005）、国家社会科学基金重大项目“南海疆文献资料整理中的知识发现与维权证据链构建研究”（19ZDA347）的成果之一。
作者简介：郑杨（1997-），男，硕士研究生，研究方向为智能目录学、保密信息安全；石进（1976-），通讯作者，男，副教授，博士生导师，研究方向为大数据分析、学术评价。

0 引言

知识经济时代，以共赢为导向的同领域和跨领域合作现象已越来越普遍。以高校为例，高校学者与企业的产业合作以及不同课题组成员间的跨领域合作程度不断深化，这使得知识主体之间的联系日益密切。随着信息化时代的脚步不断推进，学者信息在互联网上的公开度不断提升，信息需求者能够快速了解特定领域学者的相关资料，这进一步推动了知识主体之间的合作。

现有的学者信息来源途径包括以学者名录、人物传记、学者个人主页、人物信息搜索引擎等为代表的信息检索工具。诸如“中国科学家在线”、“AMiner”、“Wilson 人物传记图文数据库”、“Gale 著名传记数据库”等学者信息检索工具的推出，实现了学者信息的多维度展示，促成了知识需求者与学者之间的沟通。

然而，目前所存在的这些学者信息检索工具存在诸多弊端：只是将学者的基本信息进行简单罗列，缺乏对学者信息的结构化、全方位、立体化组织，无法切实推动知识需求者与学者之间的合作；人物传记工具书上关于被传者的资料主要是以颂扬为主，有的材料是被传者本人按统一的格式直接提供的，有的包含政治偏见，其客观性和准确性需要仔细审视，需要与其他材料对比、分析，才能做出比较公正的判断[1]；同一学者存在多个个人主页，更新不同步，导致单个主页揭示的学者信息较为片面；人物搜索引擎多为收费性质，公开度低，无法平衡好涉及学者隐私保护与学者信息公开的矛盾问题。

为了解决上述问题，本文创新性地提出了学者智能目录体系的构建设想。学者智能目录体系从信息需求者的角度出发，以非结构化的学者信息为基础，通过信息收集、信息分析、信息著录、信息标引、信息排检，将杂乱无章的原始学者信息进行组织，构建一个有序的、优质的信息集成平台。学者智能目录体系的目标是为信息需求者提供特定领域学者的信息，促成知识主体之间的同领域或跨领域合作，实现共赢。

1 相关研究工作

学者是指掌握某一研究领域的学识、能表达具有学术影响力的观点、能提出学术见解的人。虽然国内学术界对于学者目录的研究起步较早，但研究量不足、研究的深度也不够。现有的学者信息来源途径包括以学者名录、人物传记、学者个人主页、人物信息搜索引擎等为代表的信息检索工具。

1.1 学者目录相关研究

我国对于学者目录的研究可以大致分为单一学者信息描述和多维度刻画两个阶段。单一学者信息描述阶段的学者目录构建可以追溯到 1993 年，张京生等编纂了回族学者、史学家杨志玖先生的著述目录 [2]，这一阶段的学者目录编纂特点是以成文形式概述了学者的成长轨迹，内容记载详细、全面；但目录所记载的内容单一、体系紊乱，信息精确度也有待考证。

到了 21 世纪，学者目录的发展步入多维度刻画阶段，目录在囊括学者生平、学者学术成果等基础信息之上，加入了对学者的学术评价以及各维度指标体系的衡量。张前从学者业绩与境界层面丰富了《岸边成熊博士业绩目录》一书的内涵[3]；李相勋构建了韩国船山学学者研究成果目录[4]。目录所包含的内容呈现出多元化的趋势，信息的精度也有所提升，目录致用性也有所改善，但目录的编目维度依旧杂乱，用户在查找学者相关属性信息时依旧耗时久、效率低。

上述弊端的暴露，表明现存的学者信息检索工具无法满足知识需求者的需要，急需构建一个完善的学者智能目录平台来对学者信息进行整合。

1.2 学者信息检索工具发展现状

随着信息技术的发展，目前用户对于学者相关信息的搜集主要借助于在线学者信息检索工具。学者信息检索工具的形式多样，主要包括学者名录、人物传记、学者个人主页、人物信息搜索引擎。学者信息检索工具为用户提供了查找关于人物生平、研究领域、科研成果等方面信息的途径。

表 1.1 国内外学者信息检索工具概况

范围	名称	功能	不足
国外	俄亥俄州立大学主页“People Search”	查找教职员工和学生信息	只供单位内部人员使用
	People in U.S.	利用查找功能检索美国各级政府部门的官员信息	由于个人信息保护和个人信息搜索的商业化，检得结果越来越少
	Free People Search	可在同一页面调用多个人物检索网站数据库的信息，绝大多数情况下可满足检索要求	需要等候的检索时间较长
	Gale 著名传记数据库	整合了多个传记数据库与多种全文刊物，报道世界范围内各个学科领域有重大影响力的学者机器评传型资料	收费昂贵、公开性差
	Wilson 人物传记图文数据库	收录了公司出版的 100 多种人物传记工具书及其相关的期刊文献	编目紊乱、检索效果差
国内	中国人物传记网	可以检索或浏览从古至今的著名人物的传记，包括：生卒日期、职业信息等	以颂扬为主，客观性和准确性需要仔细审视
	中国科学家在线	依托于科学家在线网站的专家数据库，科学家在线提供学术专家的智能发现引擎，可以通过姓名、研究成果、研究领域、研究项目、研究专长等方面的关键词发现专家，并了解专家的研究成果	学者信息收录不完整、收费、不开源
	AMiner	学者档案管理及分析挖掘、专家学者搜索及推荐、技术发展趋势分析、全球学者分布地势图、全球学者迁徙图、开放平台等	学者信息收录不完善、更新慢、重名学者区分度差

1.3 智能目录体系发展

1.3.1 目录学发展的趋势

大数据时代，人工智能技术的广泛使用赋予目录学全新的时代意义。目录学已不单纯只是“辨章学术，考镜源流”的工具，当代目录学呈现出数字化、智能

化、全面化和实用化的研究发展趋势，铁路物资目录、医疗大数据目录、地质资源目录等新型目录体系的出现，在进一步扩大现代目录学学科边缘的同时，也凸显出我国目录学发展的实用性。目录学处理的对象也从传统的文字信息转换为电子信息资源，王蕾认为信息的精准提取、搜索的快捷方便、信息的有效控制、资源的服务共享、知识的存储提取等已成为当代目录学研究的重点[5]。

新时代下目录学与网络资源研究的结合也更加密切，目录学的功能拓宽到了智能检索、个性化推荐等方面，数字目录学的应用可以为网络信息资源提供导航与评价[6]，超文本、搜索引擎、指引库技术，以及内容方面的元数据和图书馆编目与目录学的关联性也越来越强[7]。

1.3.2 智能目录学研究工作

智能目录学作为传统目录学与人工智能技术结合的产物，已逐渐成为当代目录学的一个研究热点。智能目录学的研究起源于上世纪末，这一阶段的显著特征是数字技术在目录学领域得到了广泛的应用，具有代表性的成果包括美国图书馆公司与 1987 年研发了一款作用于多感官、配有声音及图像信息、带有人工智能软件的 BiblioFile 新型智能目录[8]；中央档案馆于 1992 年研发了《计算机档案资料管理智能软件系统及革命历史档案目录数据库》并获得国家级科技进步奖[9]。

到了 21 世纪，智能目录学步入快速发展阶段，这一阶段的特征是目录学与电子商务、消费者行为学的研究紧密相关。这一阶段的相关成果主要包括：丁峰提出了一套完整的基于本体映射的电子目录智能服务体系，同时设计了电子目录以及相应的本体映射关系描述方法和存储方法[10]；陆楠、梁正平等开发了基于商业智能兴趣度的顾客目录分割算法，实现了对向不同顾客的目录个性化定制服务[11]；席磊、郑光等紧随其后，构建了一套基于个性化特征的无公害农产品目录智能服务系统，将个性化目录定制的理论付诸实践[12]。

虽然目前已存在大量智能目录产品，但是对于智能目录理论体系的研究依旧缺乏，对于智能目录学的定义、作用、功能和构建学术界尚未形成共识。石进、

胡雅萍等率先给出了智能目录工作的定义，智能目录工作指应用计算机、大数据、人工智能等技术提高索引、文摘、参考咨询等工作的效率并尽量满足工作人员和用户的各种需求[13]。

在此基础之上，我们给出学者智能目录体系的定义：从信息需求者的角度出发，以非结构化的学者信息为基础，应用计算机、大数据、人工智能等技术，通过信息收集、信息分析、信息著录、信息标引、信息排检，将杂乱无章的原始学者信息进行组织，从而构建的一个有序的、优质的信息集成平台。

相较于传统的学者目录，学者智能目录体系的创新点主要包含以下几点：从用户需求出发，完善学者的社交网络图，将学者各个维度的信息进行自动关联，以便给予知识需求者一个直观的概念；建立更为完善、合理的学者学术评价体系，供知识需求者进行筛选，并提供相似学者的推荐；量化学者的合作偏好度，帮助知识需求者以较快的效率找到合适的学者。

2 学者智能目录体系的功能

学者智能目录体系在功能方面与传统的学者信息检索工具有本质区别。传统的学者信息检索工具以提供学者基本信息为主要功能，对于用户来说附加值低、实用性差；而学者智能目录体系以合作为导向，通过提供学者信息检索、学者推荐、学科导览、学者信息导读、学者信息关联、学术评价等功能，提升信息需求者的检索满意度，拉近知识主体之间的距离。

2.1 信息检索

与传统目录检索系统相比，学者智能目录系统的信息检索功能更具高效性，其高效性体现在高效的信息著录、标引与排检过程和高效的检索语言两个层面。

（1）高效的信息著录、标引与排检过程：基于 RDF 模型的元数据描述框架能够使学者信息的著录更为概念化和直观化，以 XML 作为 RDF 数据模型的语法，能通过非常规则的方式表达数据模型的全部功能，使描述数据的语法形式更为简

洁；借助语义网技术，在涉及学者相关信息资源主题词主动标引的过程中，结合主题词表的分类体系构建语义网，对标引过程中所得到的主题词在语义网框架下进行语义逻辑推理，得到具有语义意义的标引词；智能目录体系将以搜索引擎的形式呈现给用户，综合运用 PageRankTM 技术、超文本匹配分析技术和内容相关度评价技术，并基于信息关键词的排检能够将学者各种信息有序存储在信息系统、方便用户检索。

（2）高效的检索语言：与传统学者名录检索工作相比，学者智能目录在检索语言上可以支持自然语言检索、多媒体检索和超文本检索，以实现对文字、图像、视频和音频的检索，进一步提升了用户友好性。

2.2 学者推荐

学者智能目录以促进知识主体之间的跨领域合作为构建目的。为响应用户的合作需求，目录将学者合作偏好度的考量融入学者推荐功能模块。以促成知识需求者与学者之间的合作为导向，实现相关学者的推荐。

学者合作偏好度能用来衡量学者与他人在科研或工程项目等方面进行合作的喜好程度[14]。构建学者合作偏好度指标体系，能够量化学者的合作喜好程度，帮助知识的需求者精准、高效地找到相关领域的专家，促成合作。我们可以通过用户需求与学者研究领域相似度、学者的 H 指数、论文合著情况等指标来量化学者合作偏好度，具体可见本文第三章中的详细描述。

2.3 学科导览

不同于传统学者名录的信息展示方式，学者智能目录以演化趋势图的形式将学科的发展脉络呈献给使用者，这大大地提升了信息的可塑性与价值性，起到了开化的作用。通过对学者信息中相关知识资源的挖掘、分析、总结与关联，能够将内含的隐性知识外化，构建起学科内各个学者群的关联脉络体系。具体可以表现为运用知识图谱、语义分析、聚类分析等文本处理技术，从现有的学者知识资源中归纳、总结、研究某个领域内学者的关系，辨识出具有高学术影响力的学者，

并对学科的演进趋势和研究发展方向做出预测，揭示学科前沿趋势。

2.4 学者信息导读

学者智能目录能够为用户提供优质化的学者信息导读，剔除了冗余、过时信息对用户检索效果的影响。为剔除冗余信息的影响，学者智能目录将结合大数据技术从源头对收录的信息进行筛选、清洗和评估，以挖掘出高质量、高附加值的学者信息。为保证学者信息的时效性，学者智能目录体系的构建需要引入高效的触发器机制与更新机制。触发器最早是一种应用在数字电路上具有记忆功能的循序逻辑元器件[15]。

在学者智能目录体系构建的过程中，学者信息是多维度、多变的，学者信息的更改会刺激触发器使之产生新的脉冲，提醒平台进行信息的更新。

此外，针对传统学者信息检索工具所暴露出的信息重复收集、收录不精确、回溯性差等问题，学者智能目录体系引入高效的溯源机制。受区块链技术的启发，学者智能目录体系将引入去中心化、可溯源性的数据存储模式，构建一套适合科研行为溯源的模型及方法。

2.5 学者信息关联

在学者信息获取过程中，受信息渠道非单一的影响，获取的学者信息通畅具有异构性、多样性、大规模等特征，同一用户在不同数据源中的信息关联成为了学者智能目录体系构建过程中的重点。

目前，国内外主流的身份关联方法可分为基于表示学习的方法和基于身份匹配的方法。基于表示学习的方法是指在对学者信息进行特征抽取的基础之上，以映射的形式表示相关信息，从而判断信息是否关联，常见的基于表示学习方法有基于网络结构与嵌入式的方法、无监督式方法、基于特征建模方法等；基于身份匹配的方法是指利用概率模型，结合学者姓名、工作机构、论文、专利等信息，计算用户的相似度及身份关联概率，从而实现学者信息的关联，常见的基于身份匹配的方法有基于用户名与用户头像方法、基于多种属性方法、基于拓扑结构方法等。

上述两种身份关联技术都有各自的优缺点。基于表示学习的方法普适性强，能够综合考虑学者的科研偏好，缓解由于学者信息量不足带来的关联不当问题，但是该方法会受先验知识的影响，信息特征抽取方式的优劣直接影响关联效果；基于身份匹配的方法利用学者信息中相对独特和稳定的数据，能够缓解先验知识不足带来的影响，但该方法的扩展性能差，普适性差，对应用领域的要求较为苛刻。

综上，基于表示学习的方法和基于身份匹配的方法均有各自的优缺点，各方法之间可以弥补各自的不足，因此在实现学者信息关联功能时可以将两种方法组合使用。

2.6 学术评价

智能目录系统的用户由于缺乏专业知识，对于如何选择合作对象，相似学者如何进行区别和选择，往往无从下手。在这种情况下，我们就需要考虑引入学者学术评价功能。

学术评价是综合运用定性或定量的研究方法，对学术实体之间在理论和实践角度相互影响程度的度量。定性方法以同行评议法最为普遍，学术大国多以同行评议作为学术界认同的主要评价方式，同行评议法具有一定的权威性和专业性，某种程度上更加规范和严谨，但该方法对评审专家的学术水平要求较高，且带有较强的主观性[16]。定量方法就是构建一套文献计量指标体系，学者的学术影响力进行量化，常用的定量指标有发文数量、被引数量、H 因子、G 因子、皇冠指数等，但是定量分析方法存在未区分不同层次学术期刊发文难度与被引难度、未区分不同领域发文和被引难度等问题。因此，急需构建一套完善、公平的学者学术水平评价机制。

学者智能目录体系拟将学术均衡理论引入学术评价研究，以求相对公平地评价研究者的学术水平。所谓的学术均衡价值度理论，是指充分考虑了每位学者的论文发文数量、对单篇论文的贡献程度、发文价值（即刊载期刊排名）和引用价值，经过综合计算后能够量化出每位学者的学术价值度，以确保学术评价体系的可靠性、可行性与价值性。

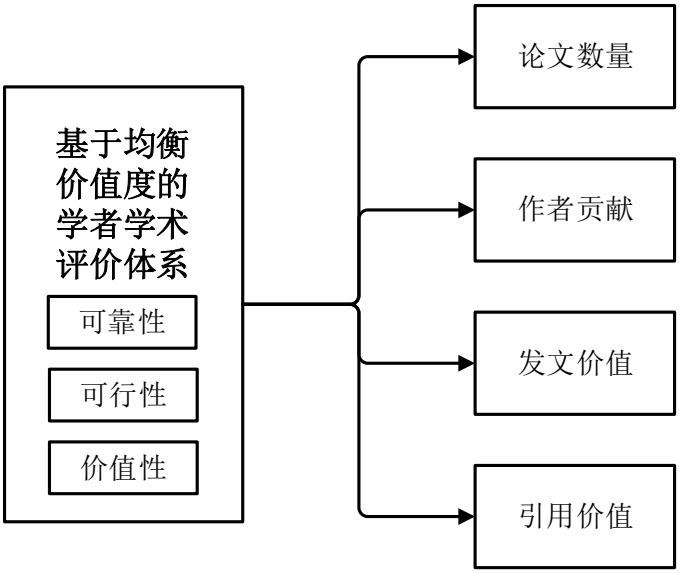

图 2.1 基于均衡价值度的学者学术评价体系

3 学者智能目录体系的构建

本章首先展示学者智能目录体系各个层级的构建，再从学者信息采集标准、学者入选标准和学者科研合作偏好度量体系四个部分出发，规范体系构建的流程。

3.1 体系结构介绍

学者智能目录体系的构建可以分为三个层级：学者信息收集层、学者信息处理层和学者信息应用层。信息收集层的重点在于参照学者信息入选标准，收集与学者相关的内源式和外源式信息；信息处理层的主要工作是基于信息收集层所采集的学者信息通过预处理、分析与汇总构建学者信息数据库，对学者信息数据库中的数据进行著录、标引、索引和排检，从而建成学者信息搜索引擎；应用层着重于在学者信息搜索引擎平台上实现学者信息检索、学者推荐、学科导览、学者信息导读、学者信息关联和学术评价六项功能。

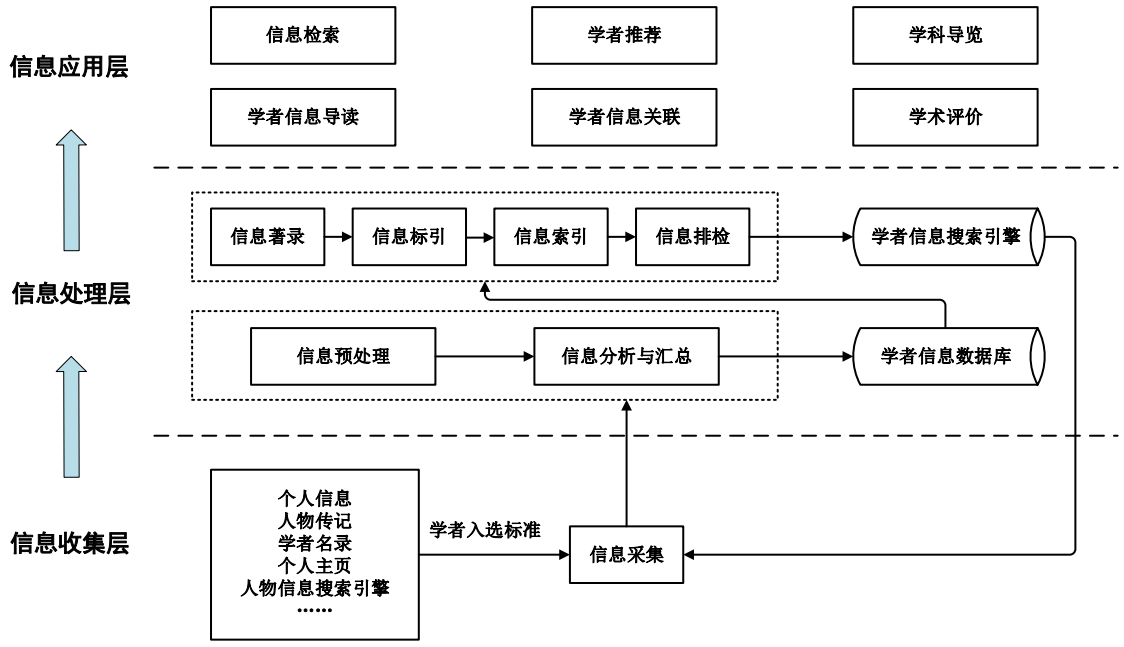

图 3.1 学者智能目录体系构建框架图

3.2 学者信息采集标准

学者信息的来源包括内源和外源两种，内源是指学者智能目录平台的用户自己上传的个人信息，这部分的信息真实度和可靠性高，但从自我角度出发往往缺乏对自身的客观评价；外源主要是指外部数据库。

学者智能目录体系在构建目标和功能上与传统学者目录存在较大的差异，因此需要构建一套适用于自身的信息采集标准。综合前文学者智能目录体系所要实现的功能，收录的学者信息应包括期刊维度的学者信息以及作者维度的学者信息。期刊维度的学者信息包括期刊载文量、期刊引用量、期刊影响因子和普莱斯指数等；作者维度的学者信息包括作者姓名、作者 ID、作者所在机构、发表论文数量、发表论文的被引频次、H 因子、G 因子等。

表 3.1 学者信息采集标准

收录维度	收录内容
期刊维度	期刊载文量
	期刊引用量
	影响因子
	普莱斯指数

	……
作者维度	姓名
	作者 ID
	机构
	发表论文数量
	发表论文的被引频次
	H-index
	G-index
	……

3.3 学者入选标准

本着“求精舍全”的原则，体系在入选学者时需要遵循一定的标准。结合 2.6 中提及的均衡价制度模型，我们提出学术价值度的概念，用以考量待入选学者的科研贡献。

学术价值度综合考量了学者发文数量、论文贡献、发文价值和引用价值。其中，发文数量衡量了学者学术产出量，但光靠数量难以客观反映学术价值，还需要结合其他维度的考量；论文贡献反映学者对单篇论文的贡献，我们可以用作者排名加以区分；发文价值体现了论文的学术影响力及对学科的贡献度，我们可以用刊登论文的期刊排名进行衡量；引用价值论文在业界的认可程度，我们可以利用引证文献的来源期刊排名衡量。

因此我们得出了学者的学术价制度计算公式：

$$学术价制度=\sum_{i=0}^{n}(论文初始分值*作者贡献*发文价值*引用价值) \quad （2）$$

（2）式中 n 表示该学者的发文总数量。单篇论文的初试分值为 1，作者贡献、发文价值和引用价值对应的取值情况如下所示。

表 3.2 单篇论文作者贡献值对照表

人数 \ 排名	第 1	第 2	第 3	第 4	第 5	第 6
1 人	1/1	-	-	-	-	-

2 人	2/3	1/3	-	-	-	-
3 人	3/6	2/6	1/6	-	-	-
4 人	4/10	3/10	2/10	1/10	-	-
5 人	5/15	4/15	3/15	2/15	1/15	-
6 人	5/15	4/15	3/15	2/15	1/15	0/15

表 3.3 单篇论文发文价值和引用价值赋值对照表

类别	来源	分值
期刊论文	《中国社会科学》、《中国科学》	9
	一流期刊	3
	《新华文摘》长文转载	3
	CSCD、CSSCI、TSSCI	1
	《中国社会科学文摘》、《高校文科学报文摘》长文转载	1
	《新华文摘》摘要转载	0.5
	CSSCI 扩展版	0.3
	UTD-24 种期刊	9
	SCI、SSCI、A&HCI 一区	9
	SCI、SSCI、A&HCI 二区	3
	SCI、SSCI、A&HCI 三区	1
会议论文	Ⅰ类	1
	Ⅱ类	0.5
	Ⅲ类	0.3

通过上述计算公式我们可以计算出每位学者的学术价值度。通过学术价值度排序我们能够筛选出具有高学术影响度的学者纳入学者智能目录体系。

3.4 学者合作偏好度量体系

学者智能目录体系的构建目标是促进学者与知识需求者之间的合作。为量化

学者科研合作偏好，我们构建了基于层次分析法的指标体系以及合作偏好度计算公式。

如表 3.4 所示，学者合作偏好可以从定性与定量两个维度进行考量，定性层面可以围绕着学者评职称需求以及个人学术追求入手；定量层面可以从研究相似度、H 指数、论文合著情况等入手。学者合作偏好度的计算公式可以归纳为：

$$F(\text{degree of interest of cooperation})=\alpha/(\alpha+\sqrt{f(x)}) \quad (1)$$

其中 $f(x)=\beta_1C_1+\beta_2C_2+\beta_3C_3+C_4$，(1)式表示学者合作偏好度的量化计算公式，其中 α、β_1、β_2、β_3 为调节因子，需要根据用户的需求进行调节，C_1 表示用户需求与学者研究领域的相似度，C_2 表示 H 指数除以该学者总发文量，C_3 表示论文合著率，C_4 为误差项常量。

由上述公式计算得到的数值能直观反映出学者的合作偏好，数值越大说明学者更倾向于与他人进行合作，需求者谋求合作的成功率越高。合作偏好度量方便了信息需求者对检得结果进行筛选，提升了检索效率。

表 3.4 学者合作偏好度指标层次结构

目标层	准则层	指标层	评价标准
学者合作偏好度	定性	评职称的需求	根据该学者目前的职称进行评职称需求的量化赋值，低级别职称的学者评职称的需求较高
		个人学术追求	个人学术追求可以与发文量以及 H 指数相关联，发文量越多说明学者的科研兴趣较高；H 指数越大的，在同领域内的学术影响力越大，学术追求往往越高
	定量	研究相似度	利用文本相似度算法，量化学者研究偏好与科研项目研究内容之

		间的相似度
	H 指数	H 指数是用来评估学者学术产出数量与学术产出水平的指标。
	论文合著情况	论文合著率=合著论文数量/论文总数量
		一作二作比率=一作二作论文数量/论文总数量
		非一作二作比率=非一作二作论文数量/论文总数量
		是一作二作且有其他作者的比率=是一作二作且有其他作者的论文数量/论文总数量

4 学者智能目录体系构建的具体问题

学者智能目录构建了一套全新的学者信息检索、导览、关联和评价体系，具备新颖性与价值性。学者智能目录体系在构建的过程中必然会面临一系列的问题，具体包括海量学者数据的爬取与处理存在难度、同名学者消歧手段不成熟、信息展示技术欠佳等。

4.1 海量学者数据的抽取存储

学者数据包括学者所发表的论文、申请的专利、个人信息、奖项成就等，如果要将所有学者数据全部抽取，那么数据量过于庞大。然而，若收录的数据量过少，整个学者智能目录体系的展示效果则会欠佳。因此，学者智能目录体系应当事先对学者进行分类，在收录学者信息时针对不同类型的学者构建不同的信息采

集维度标准，以规范体系数据库的构建。学者智能目录体系本着“求精舍全”的原则，对相关领域内的学者进行筛选，入选学者应当满足体系的入选标准。

借助网络爬虫和自动化脚本处理技术，我们能够实现对海量学者数据的抓取与筛选。常见的网络爬虫框架包括 Pyspider、Scrapy 和 Scrapy-splash。我们将学者信息抓取到后，借助自动化脚本实现入选学者的筛选。借助 Redis 关系型数据库，我们完成了入选学者信息的存储。

在建立学者信息索引结构时，传统的全文检索存储结构是对信息字段中每一个词建立索引，统计该词在文章中出现的次数和位置。当用户键入检索需求时，检索程序摆弄根据事先建立的索引进行查找，并将结果反馈给用户。学者信息来源多样、形式复杂，利用全文索引结构存储学者信息会耗用大量的存储空间，不利于定期维护和更新，管理成本会大大提升。

因此，我们引入倒排索引的存储结构。倒排索引是建立信息字段与信息字段所处文档的对照关系，以便用户从信息字段或检索词出发检得所需要的信息。倒排索引存储结构更贴近信息需求者的检索思路，也大大提升了信息的存储效率，节省了大量内存空间，提升了用户的检索效率。

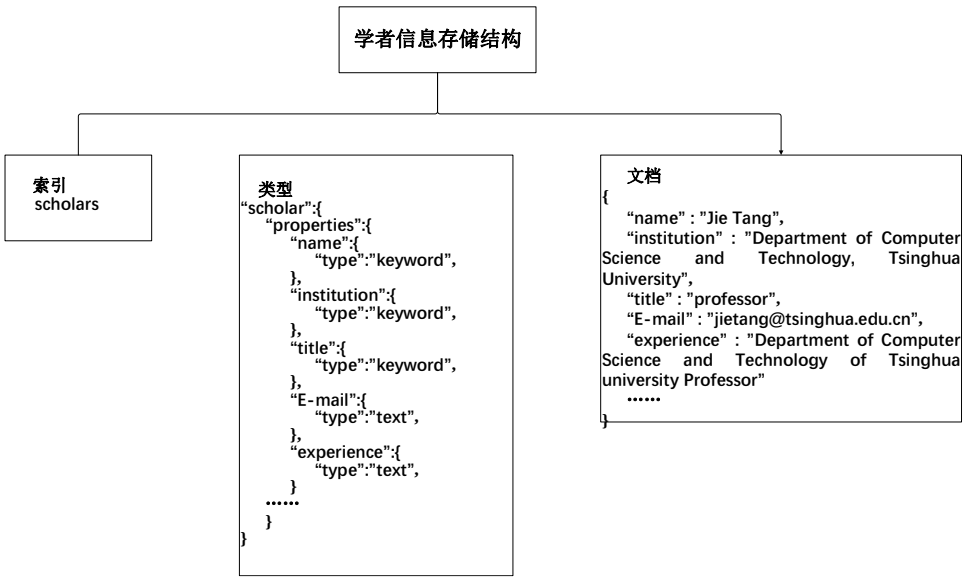

图 4.1 学者信息倒排索引存储结构实例

4.2 同名学者消歧

海量的科研信息进行作者的同名消歧是学者智能目录体系构建的重点和难

点，主要体现在：单篇文章的信息量有限，往往只有文章的作者名、题目、发表会议、期刊和发表时间。即便在文章中有关于作者基本信息的描述，但这些诸如学校或组织机构的信息会因为作者自身职位的变化而产生歧义。

学者同名消歧本质上是关系发现的过程，将关系较强的学者聚为一类。学者同名消歧可分为同名异人的消歧和同人异名的消歧。其中同名异人消歧大致包括特征抽取、相似度计算和聚类三个步骤；同人异名消歧是在特征抽取前加入信息映射这一步骤，主要是为了将可能是同一学者别名下发表的学术成果均映射到一个模块之上，在开展特征抽取、相似度计算和聚类的步骤。

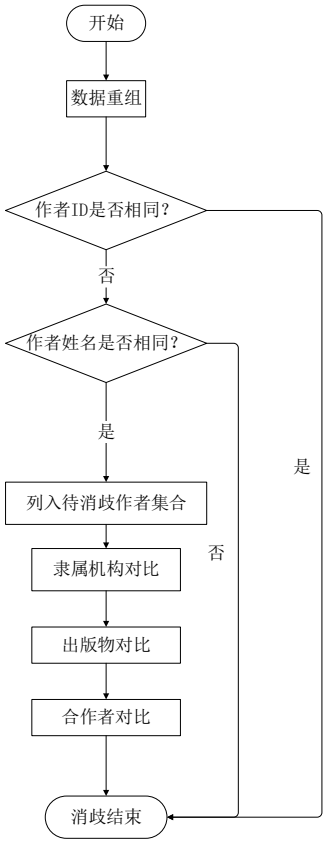

图 4.2 同名学者消歧流程图

学者智能目录构建过程中上述两种消歧方式都会涉及，本文提出基于合作作者、隶属机构及出版物信息的综合消歧方法，以提升排歧效果。如图 4.2-1 所示，基于合作作者、隶属机构及出版物信息的综合消歧方法将从论文数据重组开始。如图 4.3 所示，原始的论文信息被分解为多条数据，拆分后的每条数据对应论文中的每一位著者。其中原始数据的一些关键字段被保留，具体包含作者的姓名、ID 号、经过正则化后的隶属机构和合著者列表等字段。首先，我们要判断作者

ID 是否相同，若相同则归为同一学者，消歧结束；若不同则进行下一步作者姓名的判断。其次，若作者 ID 不同且作者姓名也不同，则归为不同学者，消歧结束。最后，若作者 ID 不同但姓名相同，则将其列入待消歧作者集合，依次对集合内作者的隶属机构、出版物、合作者进行对比，通过文本相似度算法计算相似程度以此衡量两者是同一人的可能性。

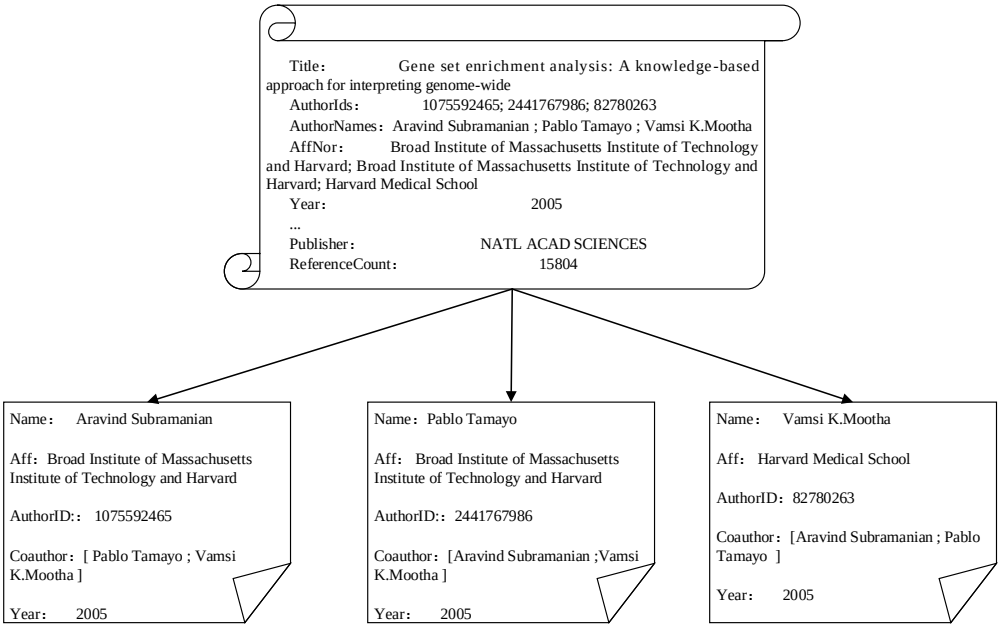

图 4.3 某篇论文学者数据重组与关键信息抽取

4.3 学者信息检索与展示

如何从海量学者信息中检索得到用户所需要的信息，这是学者智能目录必须要克服的难题。Elasticsearch 引擎为改善学者信息的检索效果提供了新方向。

Elasticsearch 是一个实时的分布式搜索和分析引擎，建立在全文搜索引擎 Apache Lucene 基础之上。4.1 节中提到学者数据的索引方式是倒排索引，建立了学者信息字段与包含其信息的文档列表的映射，再次基础上建立的 Elasticsearch 引擎应包含六个层面：应用通信层、传输层、服务层、索引层、分布 Lucene 层和网关层。

应用通信层通过 Restful Style 接口和 Netty 模块 Http 和 Netty 两种通信方式。传输层基于 JMX 管理扩展模块，实现 Thrift、Memchached 和 Http 等传输方式。服务层包括以 Zen 和 EC2 为主的服务发现模块、知识 mvel 和 js 等语言的脚本语言模块以及第三方插件模块。索引层包括生成学者信息倒排索引的索引模

块、负责关键字查找和文档获取等功能的检索模块、负责索引文档数据类型和域属性的索引映射模块。分布式 Lucene 目录层负责将多个索引段集中存储，将每个索引段的词典文件、词频文件、位置文件等包含在内。网关层连接本地文件系统、HDFS 文件系统等，同时负责文件的共享。

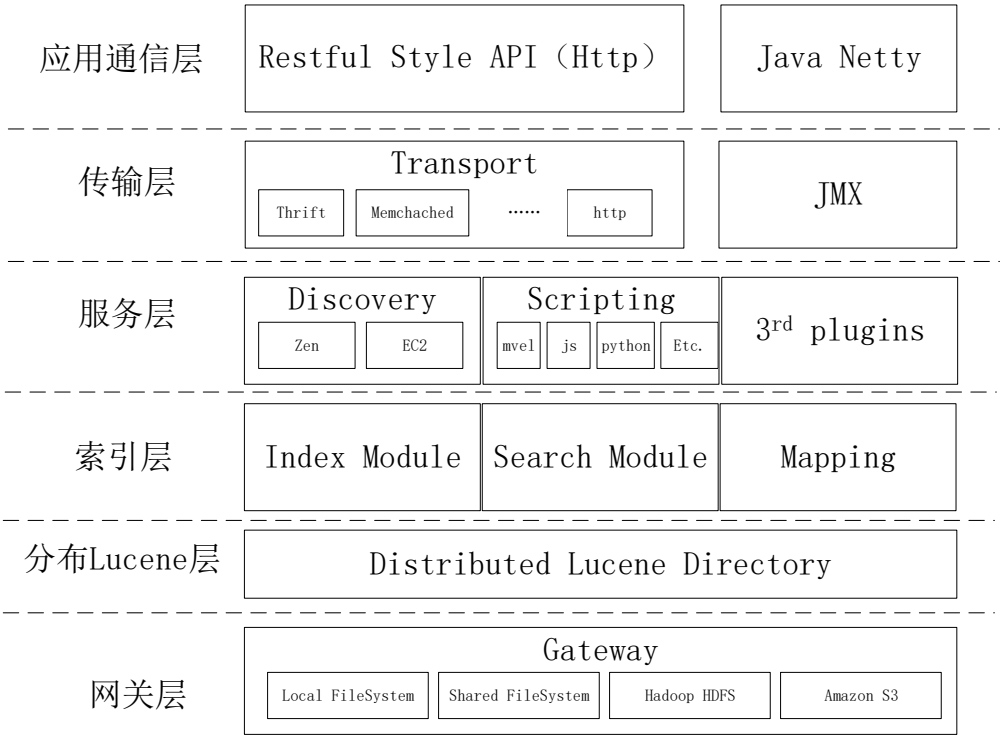

图 4.4 Elasticsearch 引擎结构图

Elasticsearch 相较于 HBase 等其他分布式数据存储系统存在一定的优势。首先，分布式的存储形式使得 Elasticsearch 具有更高的性能；其次，引擎所支持的接口更高级，支持的代码语言种类也更多，便于后续的学者信息展示；再者，引擎的数据可用度高、集群度高，能够支持 PB 级别的学者信息存储。

为方便使用者进一步了解相关领域的发展状况，体系在展示学者相关信息时，将采用关联技术与可视化技术，除了展示单个学者的学术信息外，也会显示作者群、作者单位等方面的关联，以此提升信息的可读性。

如图 4.5 所示，通过检索学者的姓名，我们可以获取从该学者涉足科研领域年份起至今的科研成果展示图，从图中我们可以更为直观地了解学者的整个成长轨迹、主要科研成果。基于成长演化图的学者成长轨迹研究也利于我们做出杰出学者及相关科研成就的预测。

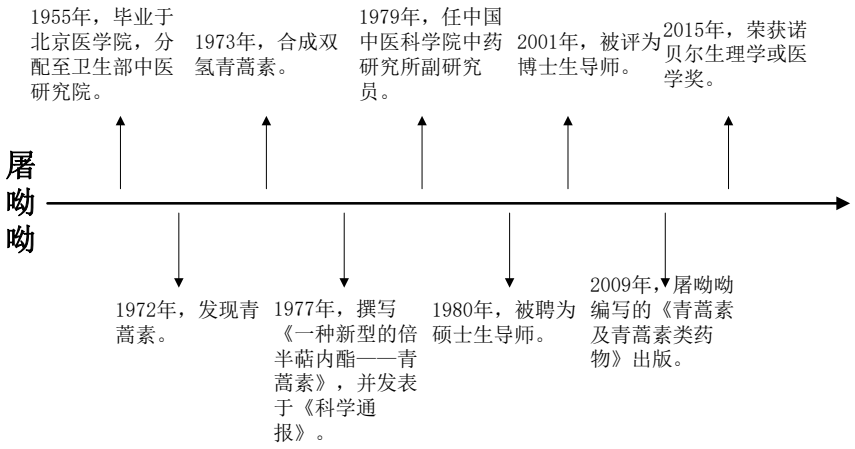

图 4.5 学者生长轨迹展示图

在使用学者智能目录进行学者信息检索的过程中，使用者除了对学者的研究领域和科研兴趣感兴趣，也会想去关注行业内部的科研热点和发展趋势。利用可视化技术我们可以展现学科发展的演化趋势。

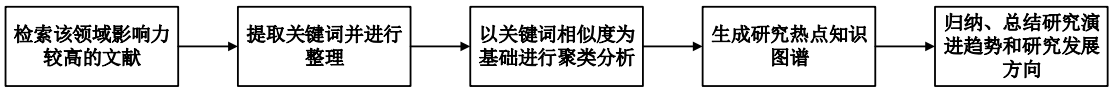

图 4.6 学科前沿趋势分析步骤

学者智能目录对于科学前沿趋势分析步骤为：首先，要检索学者在该领域影响力较高的文献。加菲尔德指出，科学研究的前沿是以被引频次最高的文献为核心，和引用这些核心文献的来源文献为基础的集合，而前沿的名称由出现频次最高的名字来表示[17]。接着，对这些文献进行关键词提取并整理，创建高频词汇表，并构建共词矩阵。在此基础之上，以关键词相似度为基础进行文本聚类分析，由此可以看出该领域不同时间阶段的研究热点。进而，可以构建研究热点知识图谱。最后，由共词知识图谱的展示，我们可以归纳、总结研究演进趋势和研究发展方向。

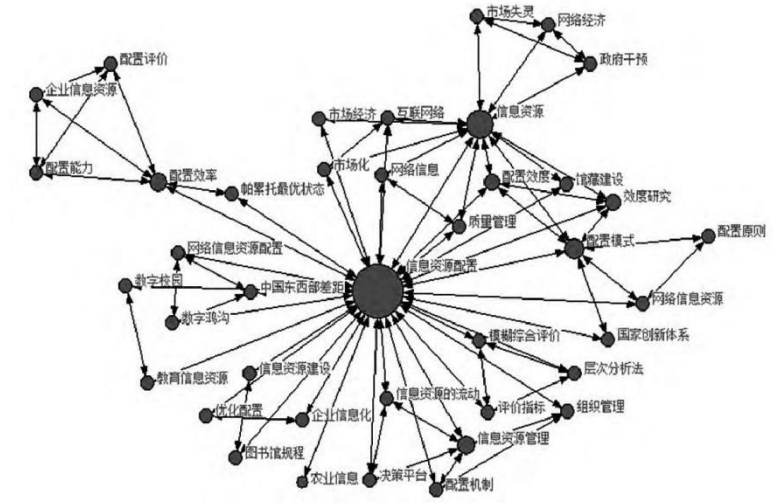

图 4.7 2004-2005 年我国信息资源配置研究热点知识图谱

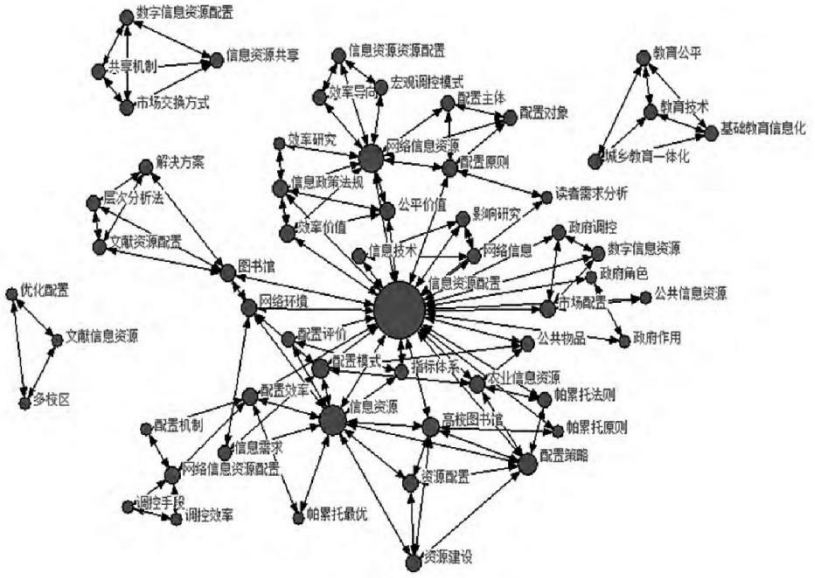

图 4.8 2006-2007 年我国信息资源配置研究热点知识图谱

5 总结与展望

本文从学者之间普遍的合作现象出发，基于智能目录体系的理论，对学者智能目录体系的构建提出了规划和建议。通过前期的文献调查与研究，我们可以发现传统的学者目录与学者信息检索工具普遍存在学者信息结构化程度不够；信息记载内容客观性不足；更新不及时等弊端。因此，构建学者智能目录体系能够弥补传统学者目录与学者信息检索工具的不足，迎合目录学的时代发展新趋势。

通过借鉴学者名录和在线学者导航平台的构建模式，构建一套以合作为导向，

全方位展现学者科研成果、研究兴趣、合作偏好的学者目录体系，并融入人工智能、大数据、深度学习、知识图谱等技术，有利于促进同领域学者或跨领域学者之间的合作，实现双赢。

学者智能目录体系是对传统学者信息检索工具的扬弃。较之传统学者信息检索工具，学者智能目录体系的优势主要体现在以下几个方面：一是以合作为导向，向用户提供全面、完善的合作方案，提高用户的知识获取效率；二是通过对学者相关信息中包含的显性知识和隐性知识的协调管理，建立了良好的组织方式，促进知识的传播和交流；三是学者智能目录体系摒弃了被动处理信息资源的工作模式，它与知识生产、分享、应用和创新全过程相融合，拉近了知识主体之间的距离，给传统目录学的智能化发展带来了新活力；四是以促进学者与知识需求者的合作为基础进行功能的设计，功能全面化、多元化，实用性强；五是学者信息的收集扩展至机构、地理位置、事件等多个维度；六是通过引入高效的触发器机制与溯源机制，确保学者信息更新及时、可溯源。

表 5.1 传统学者信息检索工具与学者智能目录体系的对比

	传统学者信息检索工具	学者智能目录体系
构建目标	为用户提供学者生平、研究领域、科研成果等方面信息	促进学者与知识需求者之间的合作
信息管理模式	以“信息管理”模式对学者信息进行管理	以“知识管理”模式对学者信息进行管理
信息资源工作模式	被动处理信息资源	主动融合知识生产、分享与创新的全过程
使用功能	单一、缺乏实用性	全面化、多元化
收录信息的维度	基本属性维度、文献维度、时间维度	基本属性维度、文献维度、时间维度、机构维度等多维度
信息更新与溯源	更新速度慢、溯源性差	基于高效的触发器机制与溯源机制进行学者信息的更新与溯源

本文创新性地提出了学者智能目录体系的构建设想，下一步工作将对体系中的数据挖掘、数据处理及信息展示等技术进行研究，实现学者信息的检索、导览、关联与评价，引领智能目录学发展的新方向。

参考文献

[1] 沈固朝. 《信息检索》[M]. 高等教育出版社, 2016.

[2] 张京生, 宋逦, 汤静芬. 回族学者、史学家杨志玖先生著述目录(1939年—1991年)[J]. 图书馆理论与实践, 1993(01):61-64.

[3] 张前. 学者的业绩与境界——写在《岸边成雄博士业绩目录》出版之际[J]. 中央音乐学院学报, 2003(04):7-8.

[4] 李相勋. 韩国学者船山学研究成果目录[J]. 衡阳师范学院学报, 2017,38(01):172-176.

[5] 王蕾, 郭芳茸, 王五选. 浅谈大数据环境下图书馆文献资源建设模式的变革[J]. 才智, 2019(28):244.

[6] 宫平. 数字目录学的功能拓展——网络阅读指导[J]. 图书馆学研究, 2007,10:73-75.

[7] 司莉彭斐章贺剑峰. 网络信息资源组织与目录学的创新和发展[J]. 图书情报工作, 2001,09:21-24.

[8] 都平平. 利用BiblioFile光盘编目系统实现西文套书快速著录的方法[J]. 图书馆界, 1998,04:3-5.

[9] 习删. 中央档案馆研制的《计算机档案资料管理智能软件系统及革命历史资料目录数据库》成果获国家级二等奖[J]. 档案学研究, 1993(01):31.

[10] 丁峰. 基于本体映射的电子目录智能服务理论研究[D]. 武汉理工大学, 2008.

[11] 陆楠, 梁正平, 杜文峰. 一种面向商业智能兴趣度的顾客目录分割算法[J]. 信息与电脑(理论版), 2011(03):100-101.

[12] 席磊, 郑光, 汪强, 等. 基于个性化特征的无公害农产品目录智能服务系统[J]. 农业工程学报, 2013,29(20):142-150.

[13] 石进, 胡雅萍, 李益婷. 大数据时代目录学的新使命[J]. 图书馆学研究, 2019(06):49-55.

[14] 胡伟, 徐福缘, 台德艺. 基于供需网的企业合作偏好度及其稳定性[J]. 系统工程, 2014,32(10):84-89.

[15] Schuler D A, Rehbein K. The Filtering Role of the Firm in Corporate Political Involvement[J]. Business & Society, 1997,36(2).

[16] 付伟棠. 我国学术期刊同行评议研究综述[J]. 中国科技期刊研究, 2019,30(08):819-826.

[17] 王瑜超, 卫武. 信息资源配置学科前沿演进趋势分析[J]. 图书馆理论与实践, 2016,05:44-49.