·专题·

面向知识服务的引文索引数据组织研究(Ⅱ)*

——引文索引数据架构与编码设计

朱云霞[1,2] 苏新宁[1]
(1.南京大学信息管理学院 江苏南京 210093)
(2.南京邮电大学计算机学院 江苏南京 210023)

摘 要: 文章以知识服务为视角,分析了引文知识服务的类型,阐述了新型引文索引的构建思路,在此基础上对引文索引基础数据层的组织架构、数据库结构以及公共字典库的编码设计进行了详细的介绍,探讨了引文索引中数据组织对知识服务的支撑作用。

关键词: 知识服务 引文索引 数据组织 编码设计
中图分类号: G254 **文献标识码:** A **文章编号:** 1003-6938(2013)05-0007-05

Research on Data Organizations about Cited Index for Knowledge Service(Ⅱ)

——*Database and code design for citation index*

Abstract In this article, the author analyzed the type of citation knowledge service, stated how to construct a new citation index in the perspective of knowledge service, then introduced the basis of the data structure, database design and code design for the citation index system in detail, and discussed the role importance of data organizations for knowledge service.
Keywords knowledge service; citation index; data organization; code design

1 引言

文献通过引用建立关联,这种关联蕴含着丰富的知识,对引用关系进行分析可以揭示知识的关联,帮助发现隐藏的知识与科学规律。信息技术的发展推动了引文数据的开发利用,早从20世纪60年代开始,就研制出了以SCI为代表的一系列引文索引系统。当前人们对知识服务的需求不断提升,引文索引已不再是简单的检索工具,人们希望能从中获取更多的知识。如何借助引文索引实现知识服务?如何从引文索引中发现科学研究规律和和潜在学术价值?这就需要我们对引文索引结构与组织进行深入探讨,使之充分体现引文索引价值,满足知识服务对引文索引的要求。

2 研究背景

文献间的引证关系始于19世纪西方科学界形成的严格科学传统[1],引文索引正是利用这种引证关系创建而成。国外最早出现的是1961年计算机编制的《遗传学引文索引》,其后在尤金·加菲尔德的带领下,先后诞生了SCI、SSCI、A&HCI等一批优秀的引文索引。国内对引文索引的研究始于80年代末期,陆续诞生了CSCD、CSTPC、CSSCI等一批引文索引系统。郭丽芳[2]、王婧[3]对中外引文索引的功能进行了比较研究。从大量文献可以看到,国内对于引用关系的研究多集中于引文数据的分析利用,而对于引文索引及其数据组织关系的研究则凤毛麟角。南京大学苏新宁教授撰写多篇文章详细介绍了CSSCI的数据组织结构与应用价值[4-5],为国内引文索引的设计与研究工作奠定了良好的基础。在此基础之上,也陆续产生了一些针对专业领域的引文索引系统[6-7]。

传统的引文索引以文献为单位,强调的是文献的检索,对于文献内部蕴含的知识以及知识间的关联不能全面、深刻的进行反映,从而不能满足广大用户的知识获取需求。本文以知识服务为视角,阐述了新型引文索引的构建思路,并在此基础上对面向知识服务的引文索引的架

*本文系国家自然科学基金项目"面向知识服务的知识组织模式与应用研究"(项目编号:71273126)和江苏省高校2012年度高校"青蓝工程"优秀青年骨干教师人才项目研究成果之一。
收稿日期:2013-09-27;**责任编辑:**魏志鹏

构设计、数据库结构以及索引编码设计进行了详细的介绍。

3　面向知识服务的引文索引构建思考

文献之间的引用本质上是知识间的关联，这些关联知识也正是提供知识服务的前提与基础。引文索引是一种典型的关系类知识工具，在文献引用过程中，各类实体间的关联是广泛而复杂的。知识服务是一种用户目标驱动的服务，是面向知识内容、面向解决方案的服务，贯穿于用户进行知识析取、集成、创新全过程的服务[8]，因此引文索引的数据组织也应当以科学研究的需要、学者的需求为目标。

3.1　引文索引的知识服务类型

科学、有效的数据组织是提供知识服务的有利保证，知识服务是数据组织的最终目的。为了更深刻的理解引文索引功效，发挥引文索引在知识服务中的重要作用，我们归纳了引文索引能提供的知识服务类型(见表1)。

表1　引文索引知识服务类型及应用说明

序号	服务类型	引文索引的知识服务应用说明
1	检索统计型	检索某一主题、某一学者的文献，检索某一期刊或作者的被引等。
2	特征分析型	提供学科的研究特征服务，如学科成长性、学科国际化程度等分析。
3	资源评价型	提供期刊、论文、图书、学者、机构、地区等学术影响力分析评价。
4	知识发现型	某领域重要学术成果的发现，如根据各类引用网络探索发现重要成果。
5	学术预测型	根据关键词分析以及引用网络的分析总结领域热点与发展趋势等预测。

传统的引文索引以检索型服务为主，以文献作为信息传递单元。虽然大多索引都具有分类统计功能，也提供了较多的检索途径，但知识服务功能相对较弱，对于更宏观、更全面的分析、评价和预测功能则却鲜见。

根据上述五种知识服务类型，我们按照知识需求的层次从低到高进行划分：检索统计型提供最低级的知识服务，其次是特征分析型和资源评价型，知识发现型和学术预测型是最高层次的知识服务类型。不同类型的知识服务对引文索引的设计要求也不同，层次高的知识服务需要有更大规模的数据和更先进的分析技术作为支撑，同时也希望基础的数据组织架构能够表达实体间更多的关联，为知识服务提供更好的数据基础。根据对不同类型知识服务的需求分析，我们对新型引文索引系统的设计目标总结为：①结构科学合理，发挥各数据属性功用，增加检索途径；②科学组织数据，呈现科学特征、规律，为数据挖掘和知识发现打下基础；③实现数据代码化，为科学地、多角度地统计分析提供精准数据；④注重数据间的关联，为展现对象间的多重关联提供途径和实现手段；⑤数据的组织能够易于系统功能的扩展。

3.2　知识服务引文索引系统模型

为达到上述系统目标，按照数据工作流程，我们将整个引文索引系统组织分为五大层次，依次为基本业务层、基础数据层、数据模式层、知识服务层和用户层(见图1)。

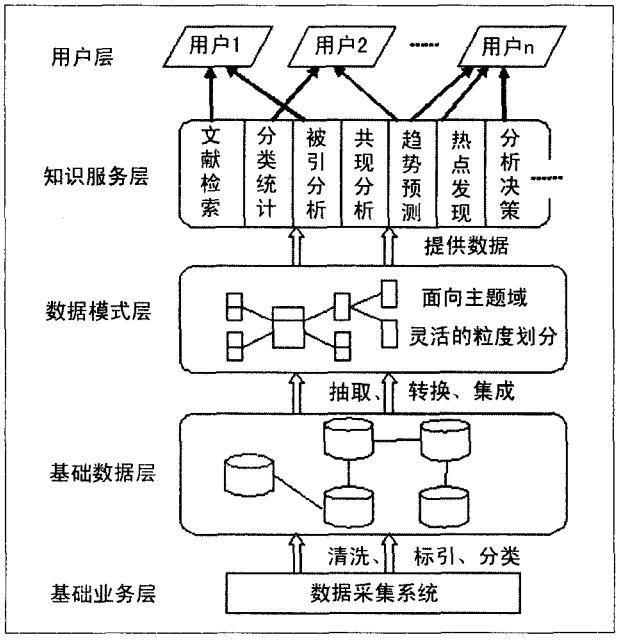

图1　知识服务引文索引系统模型

基本业务层的主要工作是相关数据的采集。包括：资源的选定(如期刊引文索引中的来源期刊的选定)，对采集的数据输入、整理、清洗、标引和分类工作等。

基础数据层是引文索引的实体部分，主要提供文献检索和一般性知识查询服务。这一层重点关注数据库的架构、细节化的库结构设计以及元数据的表达等，它是整个索引系统提升知识服务的基础数据来源，也是一般性统计分析的重要基础。

在数据模式层中，主要建立数据中各类关联，为知识服务奠定基础。该层的数据组织主要依据用户需求，建立面向主题域的知识仓库。知识仓库的数据来自于基础数据库，其数据关联来自于用户需求和科研领域的需要，并能够充分体现对象间的多维关联。

知识服务层由是完成知识服务功能的系统组成，它根据用户需求，并对基础数据层和数据模式层提供的数据进行统计、分析、挖掘等工作，并提供用户知识服务。在这一层面，要求功能模块可以根据需要扩展，系统功能的

开发可以完全独立于数据库的物理存储结构，提升系统的逻辑独立性。

用户层的作用是对用户的信息需求进行分析，将用户的需求分解成对应的知识服务功能模块，由知识服务层启动相关功能模块为用户提供知识服务。

上述分析可以看到，引文索引必须以满足用户的需求为目标，以更深入的分析和辅助决策功能为特色，以良好的数据组织和架构设计为基础，以信息知识化、方法科学化、分析智能化为核心动力，是知识服务又一新型载体。

4 引文索引的基础数据组织

4.1 框架构建

基础数据的组织是整个系统的基础，它直接关系到索引系统的执行效率和系统的稳定性。我们将基础数据分为9个组成部分，分别是：来源文献库、被引文献库、来源作者信息库、期刊目录库、期刊沿革库、机构信息库、机构变化库、关键词索引库、公共字典编码库(见图2)。

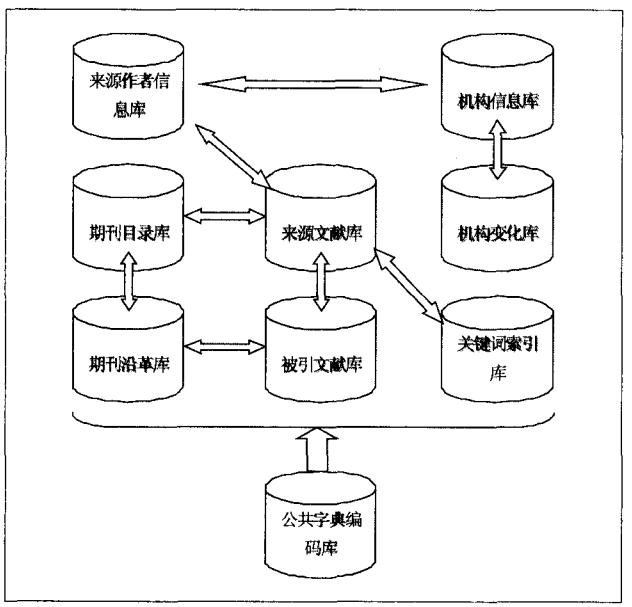

图2 基础数据组织架构图

在引文索引系统的开发中，数据库设计应遵循必要的数据库范式理论，以减少冗余、保证数据的完整性与正确性，如此设计的引文索引组织架构特点在于：

(1)提供更多的检索途径。将来源文献与被引文献分不同的库存放，用户既可以从来源文献角度追踪其被引文献，也可以从被引文献角度回溯其来源文献，提高了服务的灵活性。

(2)减少数据的冗余。由于一篇论文往往有多个作者，一个作者又常常会标注多个机构，为了节约存储空间，减少数据冗余，在来源文献数据库中只描述文献信息，对作者及其机构信息单独建立作者信息库。

(3)将机构、期刊名称进行统一。改革开放以来，特别是近十几年来，我国高校的机构名称发生了很大变化，另外很多地区的高校进行了合并，一些原有机构已不复存在，这对以机构名称为单位的统计工作造成了很大影响，因此特别增加了机构变化库，详细记录各机构名称的变化情况。同样的情况，对于期刊的历史变迁则专门建立了期刊沿革库，以记录期刊名称变化的情况。

(4)编码知识化。在设计过程中专门设置公共字典编码库，用于存放各项类型的编码，如地区编码、机构类型编码、引文类型编码等，公共字典编码库作为代码化的知识工具对各数据库数据起到统一、规范和关联的作用。

(5)提供关键词方面的检索和分析。设计的关键词索引能够极大提高检索的效率，同时对基于关键词频的学科热点分析、关键词共现分析等，均提供了良好的数据基础。

4.2 库结构及关系描述

(1)来源文献库：用于记录引文索引所收录的每一篇文献的详细信息，字段包括：文献号、中文篇名、英文篇名、中文关键词、英文关键词、期刊代号、语种代码、发表年份、卷期、页码、各种分类号、文章类型、基金类型代码、基金内容、出版日期、参考文献数量等。对于一些更为完善的引文索引，可能还会增加人工标引的主题词、中英文摘要等信息。

(2)来源作者信息库：用于记录来源文献每一位作者的基本信息，主要字段应有：文献号、作者序号、作者姓名、机构名称、机构类型编码、地区代码、通讯地址、备注等。设置作者序号，有利于进行“第一作者”检索，备注字段用于存储作者的个人情况，比如性别、出生日期、研究方向等，这些信息都能够为深层次的知识服务提供分析用数据。

(3)被引文献库：用于记录被引文献的基本信息，字段包括：文献号、引文序号、引文篇名、引文语种代码、引文类型代码、被引作者、引文期刊名称、出版社、出版年、卷期、起止页码、被引形式、被引角色、备注等。

(4)期刊目录库：用于记录被收录期刊的基本信息，该库主要与来源文献库相关联的关键字是期刊代号。主要字段有：期刊代号、期刊中文名称、期刊英文名称、ISSN号、国内刊号、出版周期、出版单位、主办单位、期刊分类、创刊时间、邮发代号、通信地址、邮政编码、网址等。

(5)期刊沿革库：用于记录期刊名称变化的情况，主

要用于统计中的数据归并处理，也可用于期刊引用网络构建时的刊名统一化处理。该库的主要字段有：期刊代号、期刊中文名称、之前名称、更名时间等。

(6)机构信息库：用于记录发文作者所在机构的基本信息，主要与来源作者信息库相关联，该库的主要字段包括：机构名称、机构英文名称、机构类型代码、国家代码、地区代码、通讯地址、邮政编码等。

(7)机构变化库：用于记录机构的名称变化信息，该库的主要字段包括：机构名称、机构类型代码、变化原因、之前名称、更名时间等。

(8)关键词索引库：用于存储所有收录文献的关键词，并建立倒排索引，主要目的用于检索和进行关键词统计，通过对关键词的统计，可以分析学科研究热点和发展趋势。该库的主要字段包括：关键词、关键词词频、来源文献号集合等。

(9)公共字典编码库：几乎和引文索引中的所有库都具有联系，该库包括7张编码表(国家与地区编码表、国内地区编码表、机构类型编码表、基金类型编码表、语种类型编码表、分类体系对照表和引用类型公共编码表)，是整个基础数据架构的连接和纽带，同时其特有的编码设计也为知识服务提供了有力的保证。

4.3　公共字典库的代码设计

编码是对数据进行知识化和规范化的过程，编码本质上是对象的抽象表达，优秀的编码规则能够使系统发挥更大的效能[5]。我们在公共字典库中设置了7种编码，使各项数据之间建立起知识化的关联，下面介绍几种主要的编码。

(1)国内地区编码。地区编码主要针对作者的机构所在的地区信息进行的编码，编码的目的是能够以省或市为单位对国内各地区进行成果的统计和分析，甚至可以进行同级别城市(如各省会城市)的科研成果数量比较。通过地区编码，我们既可以很方便地以省为单位集中统计，也可以很方便地对各省内的城市进行相关统计分析。更重要的是，地区编码能够提升地区统计的准确性和效率。

由于目前各期刊对作者的地区标引没有统一的规定，文献中作者的地区数据显得非常凌乱，有的只标注省份，有的标注省和市，有的仅给出了城市名，有的甚至没有地区标注。另外，随着中国城镇化建设步伐的加快，地区名称也常常出现变更的情况，因此编码为解决这类地区名称变化带来的统计上的困难提供了有效途径，它实际上起到了地区名称规范化、统一化的作用。地区编码需要注意地区间的从属关系以及地区信息表示的粒度问题，既要能突出表现将省内城市聚合成以省为单位的粗粒度，也要能够表示省内所辖城市为单位的细粒度。以江苏省及其所辖市为例，既要能够通过编码从大量的文献中提取出江苏省作者发表的论文，也能够统计江苏省内各地级市作者发表的论文。这样的编码设计可以依据不同用户的需求，灵活的设置地区级别。此外，通过区县级的编码还能够从细节上发现小区域的经济特色和优势产业，为科技服务产业提供了良好的数据资源。

具体的地区编码策略为：采用6位数字编码方式，其中1-2位为省级行政区域编码(包括省、直辖市、自治区和特别行政区)，3-4位为市级编码，省会城市的编码统一为"01"，若是直辖市这两位代表它们的区县，第5-6位为区县级编码，包括下属的县级市、区和县。依据这样的编码规则，北京的编码为"010000"，江苏南京的编码为"160100"，依次类推。

(2)机构类型编码。发文机构数量庞大、类型众多，对机构编码的过程，实质上是对众多发文机构进行归类整序的过程，同时也是知识化的过程。经过机构编码，引文索引可以根据用户不同的知识需求，对编码进行组配查询，来达到特殊要求的统计分析结果。例如欲分析对比全国"211"师范院校的科研实力，在没有编码前需要罗列这些学校，然后逐一检索获得相关数据。而科学的编码，可使这类繁琐的工作变得简单高效。

从科学分析角度针对科研机构分类，我们可以将所有机构划分：高等院校、科研院所、党政机关、文化团体、企业、军队系统、非高校教育单位和其他8种类型，分别用数字1-7、9表示(我们在整个编码系统中，对于其他类别统一都用数字9表示)，对于每种类型再根据具体情况编制下级机构类型代码。

例如：对于高校，可以将其划分为"985"工程院校、"211"院校、教育部直属院校、中央其他部门所属院校、地方本科院校、高职(专科)院校6种类别，分别用数字1-6表示；再下一级可以表示高校的专业特征，如：综合性院校、师范类院校、医学类院校、工科类院校、体育类院校、艺术类院校、军事类院校、财经类院校和其他专业院校9种类别，用数字1-9表示。当然，也可以根据实际情况增加院校类别。有些高校可能会对应多个编码，本编码规定一律高靠，比如，南京大学可能对应的机构类型编码有"111""121"和"131"三种，本编码系统只为南京大学取"111"作为编码，可在进行统计分析时，通过一定算法来区分。在对某一类学校进行统计对比时，我们只要借助编码就可以方便的进行统计分析。例如，当对全国财经类院校进行统计分析时，我们就可以利用代码"1x8"抽取相关数据完成统计，其中"x"可以为任何数字，也可以是指定

某几个数字。

对于科研院所也可以划分三级进行编码，如：国家所属、省部所属、专业学会所属、其他等；然后再将其细分：自然科学类、工程科学类、医药学类、社会科学类、其他，等等。如，中国社会科学院为国家所属的社会科学院科研机构；再如，江苏省中医药研究院为江苏省所属的医药学类研究所。

其他类机构都可以按纵横两个方面再划分两级编码，纵是指行政上的上下级或所属关系，如，国家级、省市级等等；横是指机构的类别或属性，如高校按学科划分等。所有机构类型编码均由三位数字组成。在实际的机构编码过程中，可以根据资源涉及的机构状况，根据实际需要来设置自己的机构编码。

(3) 基金类型编码。其他中研究院术情报所化论、转化论与融合轮将用户的需求分解成基金类型编码用于为各类资助基金进行编码，以反映各类基金的科研成果状况，进而分析各类基金在科学研究中发挥的作用。目前，基金项目类型众多，从纵向看，有国家级、部委级、省市级、以及各单位的资助项目；从横向看，有攻关计划、重大项目、重点项目、一般项目、青年项目、国际合作与交流项目等等，所以基金项目类型编码同样采用分层编码的方式。如，纵向分别用1-3,9表示国家级、省部级、市级、其他；对国家级项目再划分：国家自然科学基金项目、国家社会科学基金项目、国家863和973项目、国家其他项目计划(如科技基础条件平台建设计划、政策引导类科技计划等)，并分别用1-3,9表示；第三层再划分重大项目、重点项目、专项项目、一般项目、青年项目、其他等。

同样，部委级、省市级基金项目也可进一步划分为：自然科学基金项目、社会科学基金项目等；第三层再划分重大项目、重点项目、专项项目、一般项目和青年项目等。通过分层编码，将众多繁杂的基金项目进行了有序的归类和标引，为进一步的分析比较以及深度的知识服务做好数据基础。例如，通常情况下重大项目代表了各个研究领域的最前沿技术、最高的科研水准或国家、地区急需解决的科研问题。

通过对基金项目的编码，我们可以很方便的调出有关基金项目的成果。例如，若想通过引文索引查找或统计所有国家级重大项目的科研成果，只需要利用基金类型编码"1x1"对引文索引中来源文献进行检索即可以获得。

5 结语

任何知识的创新都是在前人研究基础上进一步努力的结果，没有继承就没有创新。文献间的引用填补了知识沿时间和空间的互补性需要。引文索引是对文献信息资源进行管理的有效工具之一，经过50年的发展，引文索引对整个科学领域的研究产生了重要的影响，引文索引系统也逐步摆脱信息检索的单一形象，在对期刊的评价、科研成果评价、人才的培养、学科的发展过程中都起到了关键的导向作用。

通过对国内多个引文索引系统的使用分析可以看到，目前我国引文索引建设还存在着明显的不足：一是功能简单，一般只提供检索与简单的统计分析功能；二是数据来源不同，检索的结果差距较大，各索引系统因开发单位不同，存在重复劳动情况，没有能够形成类似于WOS的统一检索平台；三是深层次的知识服务功能还非常稀缺，数据的统计分析都是以文献为单位，没有能够深入具体引用的内容部分，缺乏语义层面的分析。

有效的数据组织是进行知识服务的前提，引用关系是文献间最普遍最直接的联系，我们的任务不仅仅是表示这种关联，更要能从引用中发现更多知识，提供深层次的知识服务。数据的组织是知识服务的基础与关键，引文索引的研究不仅仅是个别人、个别单位的事情，应该是全社会集思广益、不断深入的过程，引文索引也不能是一个固定僵化的系统，要能够适应社会的发展和人们知识需求的变化。

参考文献：

[1]马智峰.参考文献的引用及影响引用的因素分析[J].编辑学报,2009,21(1):23-25.

[2]郭丽芳.中外五大引文索引系统比较分析[J].现代图书情报技术,2005,(1):36-39.

[3]王婧,华薇娜.国内外文科引文索引数据库检索功能比较[J].新世纪图书馆,2011,(1):42-44,73.

[4]苏新宁.中国社会科学引文索引设计[J].情报学报,2000,19(4):290-295.

[5]苏新宁.中文社会科学引文索引(CSSCI)的设计与应用价值[J].中国图书馆学报,2012,(38):95-102.

[6]纪蔚蔚.基于Web引文索引数据库建设方略[J].现代图书情报技术,2004,(12):45-50.

[7]陈建青等.中文生物医学期刊引文数据库(CMCI)的研制特色[J].现代图书情报技术,2005,(3):63-65.

[8]柴永红.论信息服务与知识服务[J].情报杂志,2004,(4):74-75,78.

作者简介：朱云霞，女，南京邮电大学副教授，南京大学信息管理学院博士研究生；苏新宁，男，教育部长江学者特聘教授，南京大学信息管理学院博士生导师。