

·馆藏与出版论坛·

# 面向文献建设需求的学科核心作者数据库构建策略研究\*

朱轶婷

(中国民航大学图书馆 天津 300300)

〔摘要〕以关系型数据库为基础,通过Web信息抽取技术从主流数据库中采集基础数据,利用数据挖掘技术进行数据整合、查重、消歧,然后根据发文量、h指数和 $h_m$ 指数综合判定核心作者及作者排序,从而构建学科核心作者数据库,为图书采访工作提供客观数据支持。最后以飞行技术学科核心作者数据库为例,说明数据库的实际效果。

〔关键词〕核心作者 h指数  $h_m$ 指数 关系数据库 Web信息抽取

〔分类号〕G253

## 1 构建学科核心作者数据库的必要性

当前,高等院校不断加大力度推进专业结构优化和重点学科建设工作。由此,对图书馆的文献资源建设提出了更高的需求,即图书馆采购的学科文献应该紧跟专业发展变化、切合教学科研需求。但是高校图书馆的传统文献采购方式往往是根据书商的供书目录进行勾选、订购,在文献采购到馆前,采访人员无法看到文献的实际内容。即使在采购过程中征求相应学科专家的意见,也往往因为缺乏客观依据,无法成功预测文献的学术价值。而且,图书馆采访人员很难深入了解学校的每一个重点学科,并且追踪该重点学科的发展变化。因此,如果能有客观数据辅助采访人员判断文献质量和学术价值,将有助于提高文献采购的客观性和科学性。

采访人员在采购过程中,通过供书目录可以掌握文献的以下特征数据:①作者,即完成创作、编写、编译该文献内容的个人或者团体;②出版信息,即出版社、出版年;③版本,即该文献是初版还是再版,再版次数等;④语种,即该文献的写作语言;⑤载体信息,主要有该文献的开本尺寸、页码、装订情况等;⑥价格。这些客观性数据有助于判断文献的学术价值。比如,作者是否为该文献涉及学科的专家学者或者权威研究机构,出版社是否为该学科的核心出版社,修订并出版多次的文献更被读者认同等等。因此,采访人员应该关注并利用这些特征数据,从而了解和确定选购文献的学术价值,确保满足学校和读者提出的文献需求。

研究选择文献作者为切入点,通过构建学科核心作者数据库,探索以信息技术手段辅助图书馆采访人

员提高学科文献选购科学性、客观性的新方法。

## 2 构建学科核心作者数据库的基础

### 2.1 图书情报学界关于核心作者的研究

核心作者是指那些在(某)学科领域研究较深入、造诣较高、研究成果较多从而具有较大影响力的作者,对学科发展具有引领作用,不断将研究水平推向新的高度。<sup>[1]</sup>图书情报学界的学者们运用文献计量学理论对核心作者作了很多深入研究。例如,方太强、周蓉等结合发文总数、被引次数、核心期刊发文数等因素,利用维普数据库测定图书情报学领域的核心作者;<sup>[2]</sup>赵基明等运用h指数方法,利用CSSCI引文数据库1998-2006年的数据,测定《中国图书馆学报》的核心作者;<sup>[3]</sup>龚舒野运用发文量、h指数和 $h_m$ 指数方法,利用CNKI数据库2001-2009年的数据,测定了《情报科学》的核心作者,并分析这些作者的年龄、职称、地域等特征信息;<sup>[4]</sup>邱均平等运用发文量和h指数相结合的方法,利用CSSCI引文数据库的数据测定图书情报学领域近30年的核心作者。<sup>[5]</sup>

亦有学者将核心作者的测定运用到实践中,推进图书馆工作开展。例如,苏志芳等运用发文量、h指数和主题研究连续数相结合的模糊综合评判法,测定学科领域核心作者,并提出以核心作者为主要依据的中文社科图书决策系统;<sup>[6]</sup>蔡璐运用层次分析法测定高等教育学科的核心作者,作为判断图书学术价值的依据之一;<sup>[7]</sup>沈艳红、吴信岚等利用CNKI数据库,确定食品学科的核心作者,作为制定采购该学科核心书目的依据之一。<sup>[8]</sup>

这些研究与尝试,探讨了核心作者的不同测定方

\* 本文系中国民航大学校级科研项目“日本民航网络资源典藏库的构建与研究”(项目编号:2010kyh03)及中国民航大学2015年中央高校基本科研项目“基于大数据分析的多馆制文献资源管理策略研究”(项目编号:31220157006)研究成果之一。

法,并以实证研究方法研究实际效用,为我们研究构建学科核心作者数据库提供了文献计量学方面的理论支持。

## 2.2 人物数据库的研究现状

人物数据库指利用信息技术记录和管理人物信息,并且实现便捷查询和数据共享的数据库。核心作者数据库也属于人物数据库范畴。

在国外,比较有影响的人物数据库有英格兰圣公会神职人员数据库(The Clergy of Church of England Database)<sup>[9]</sup>,该数据库记录了1540~1830年期间英国神职人员的任职、职务等信息;还有ASP世界历史人物索引库<sup>[10]</sup>,该数据库记录了世界上历史事件发生时所涉及的第一个人物,内容包括信件、日记、口述史与其他个人叙述等。

在国内,有中国科学技术协会牵头、北京理工大学图书馆主要承办的老科学家学术成长资料数据库,收集300位80岁以上的院士或96岁以上有突出贡献的非院士科学家的资料;<sup>[11]</sup>也有各高校图书馆基于学科研究或特藏建设需求而建立的人物专题数据库,如暨南大学图书馆的留学人物数据库、嘉兴学院图书馆的嘉兴名人数据库等等;还有公共图书馆建立的地方特色人物数据库,如湖南图书馆的湖南近代人物资源库、广州图书馆的广州人物数据库等等。

无论国外还是国内的人物数据库,均重视建立设计完备、字段丰富的数据库系统,以便较好地汇集、组织和揭示人物信息;注意建立人物信息与文献信息之间、异构信息之间的关联;尝试运用知识地图、本体论等理论方法,对人物信息中的知识进行深层次挖掘,以期提供针对性强的高层次知识服务。这些数据库的有益尝试,对我们研究构建核心作者数据库提供了实践支持。

## 3 学科核心作者数据库的构建设计

构建学科核心作者数据库的设计定位是挖掘、判定学科领域核心作者,将其提供给采访人员作为订购学科文献的辅助决策依据。根据文献计量学理论,判定学科领域核心作者需要一系列的基础数据,因此构建学科核心作者数据库的基本设计思路是通过网络抽取从主流数据库中获取的基本数据,然后根据文献计量学理论设计核心作者的判定算法,再结合云计算技术与元数据进行数据组织,储存并揭示核心作者的信息,最后利用动态网页开发技术将作者信息整合在一起,提供一个可视化的、便捷的数据呈现界面,方便采访人员进行采购决策。

### 3.1 学科核心作者数据库系统模型

构建学科核心作者数据库采用B/S架构,系统模型如图1所示,细分为四层:资源层、指标层、数据层和应用层。

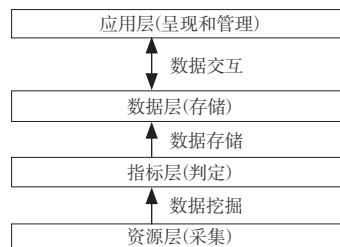


图1 学科核心作者数据库模型

资源层是获取基础数据的一层,属于四层结构的最底层,是构建学科核心作者数据库的数据基础。根据文献计量学理论,判定核心作者主要依靠发文量、被引频次、h指数等等,但是这些数据很难直接获取。因此在资源层,主要采集学科领域内所有作者的发文情况,如题目、刊名、关键词、摘要、出版日期等等,以及作者的个人属性数据,如单位、职称、主要研究方向等。获取方法以网络Web抽取为主,辅助以人工抽取。数据来源为主流数据库,如中国知网、万方、维普等。

指标层是完成核心作者判定的一层。首先对资源层的基础数据进行查重整合,然后将发文情况的整合结果提供给计算算法,得出发文量、被引频次、h指数等判定数据;再按照判定算法,给出核心作者的判定结果。如果某一作者被判定为核心作者,则将整合后的作者个人信息和判定数据一起储存在数据层中。

数据层是储存学科核心作者数据的一层。在这一层中,依照元数据的标准,建立数据表,对核心作者的个人属性数据进行静态数据标引,对作者发文情况和判定结果进行动态元数据标引。运用数据关联技术,将“作者—文献—学科”关联起来,为应用层的核心作者呈现和检索提供了基础。

应用层是直接面对用户的一层,主要提供人性化、便捷的Web交互界面。用户分成两类:普通采访人员和管理员。针对普通采访人员,应用层提供学科选择、时间段选择、核心作者浏览、核心作者检索等服务,支持关联作者发表文献,以方便采访人员进一步深入研究该核心作者;针对管理员,应用层提供数据维护、人工去重、专家判定等管理功能。

### 3.2 基础数据采集和查重

研究尝试采用一种基于Agent的中文Web信息检索平台,模拟正常用户访问主流数据库的流程,然后根据设定的检索表达式,进行数据检索,再把检索结果返回。这种做法能够规避大规模的人工检索和数据整合,

有效提高构建数据库的效率。

因为研究建设学科核心作者,所以在构筑检索式时,以学科主题词为检索词,生成相应的检索表达式。通过检索,可以直接采集以下数据:文章属性相关数据——题名、刊名、出版年、卷、期、页码和摘要;文献计量相关数据——单篇文章被引次数、下载次数;作者相关数据——姓名、单位、联系地址。

由于每个学科均有多个主题词,因此由 Agent 平台直接采集、返回的数据存在较多重复数据,因此在基础数据传递给指标层、用于判定核心作者之前必须进行查重。一是要合并相同的文章,主要通过比对文章题名、刊名和出版年卷期数据等,二是对于作者姓名的查重和消歧。可借鉴香港中文大学图书馆的 Chan 和 Yik<sup>[12]</sup>提出的用于机构知识库的作者姓名规范的概念模型,建立作者信息规范表,赋予每个作者 ID 编号作为唯一标识,将作者 ID 号、姓名、机构名作为一个集合进行考察,经过匹配完成作者姓名的查重和消歧。建立每个作者的唯一标识,就是赋予每个作者唯一身份,还可以将采集到的文章属性数据、文献计量数据和指标数据映射到这个唯一标识上,避免因作者姓名引起的文章归属冲突,使发文量的计算更加准确。

### 3.3 学科核心作者的判定

资源层的基础数据经过整合、查重和消歧后,可以得到每一位作者的文献计量学指标:发文量、总被引证篇(次)数、单篇被引证篇(次)数。发文量是指某一位作者总共发表了多少篇文章。在文献计量领域,曾根据这一指标评判作者的学术成就,但是发文量指标仅能说明该作者是该领域中写作活跃的作者,不能反映文章质量和该作者对该学科领域的影响力。同样,被引证篇数也是文献计量学评价作者学术水平的传统指标之一,论文被引用的越多,说明其观点和资料越被同行学者认可,论文作者的水平也越高,但被引次数同样也存在不足,比如论文自引现象。综合近几年文献计量学者的研究,较少根据单一指标判定核心作者,很多高质量的研究论文都是采用多个指标综合评估、判定核心作者。因此,可根据基础数据的采集情况和文献计量学的研究成果,采用发文量、h 指数和  $h_m$  指数综合判定学科核心作者。

首先,根据发文量数据,运用普赖斯定律进行核心作者的初选。普赖斯受社会学的卢梭定律启发,经过研究后发现,在同一主题中,半数的论文由一群高生产能力作者撰写,这一作者集合在数量上约等于全部作者总数的平方根,具体公式为: $m \approx 0.749(\sqrt{n_{\max}})$ 。其中, $n_{\max}$  是指发文量最多的作者的发文总数。也就是对于

某一学科领域,只有发文量超过  $m$  的才能被列为高产作者,可以被初步选为候选核心作者。

然后,运用 h 指数,进一步判定学科核心作者。h 指数是美国统计物理学家 Hirsh 于 2005 年提出的,其核心思想是一位作者至多有 h 篇论文分别被引用了至少 h 次。h 指数同时考察作者的发文数和引文数,并把这两项指标合二为一,兼顾了作者文章的“量”与“质”。h 指数可以根据作者的发文量和单篇被引次数计算得出,然后根据给定的阈值,在候选核心作者群中,确定学科核心作者。

最后,运用  $h_m$  指数对学科核心作者进行修正和序次建议。h 指数在反映高质量论文上有很多优势,但是仍有不足。经过实践,在同一学科中会出现很多学者的 h 指数相同的现象,在需要根据核心作者对图书进行采购决策时,容易出现难以取舍的情况。因此,可以引入  $h_m$  指数。 $h_m$  指数是我国学者赵学梅提出,并已经经过实证研究证明可行<sup>[12]</sup>。 $h_m$  指数引入修正因子  $1 + \frac{1}{n}$ , 对 h 指数进行一次修正,公式为: $h_m = h + \frac{h}{N_{c,\text{tot}}}$ , 其中  $N_{c,\text{tot}}$  为该作者的总被引篇(次)数。通过  $h_m$  指数的公式,可以看出: $h_m$  指数是一个介于 h 和 2h 之间的小数,且总被引次数越高, $h_m$  指数越接近 h 指数。也就是说, $h_m$  指数和 h 指数差值越小,该作者的影响力越大。除非某两位学者的 h 指数和总被引次数完全相同,他们的  $h_m$  指数才会相同。这样经过 h 指数判定为核心作者的学者,在绝大多数情况下都会有一个自己独特的  $h_m$  指数,能够给采访人员更加准确的决策依据。

### 3.4 数据库的数据结构表示

学科核心作者数据库虽然从逻辑上分为四层,但是采集、判定、呈现、检索等应用全部围绕数据展开。因此,在构建学科核心作者数据库时,选择关系型数据库 SQL Server 为数据中心,向判定、检索等上层应用传递资源层 Agent 采集的基础数据,充分利用 SQL Server 服务器的并发和处理能力,将数据分析交给数据库服务器的存储过程,简化了上层应用的设计复杂程度。

因为以 SQL Server 数据库为中心,所有的数据和对象均映射到数据库中,数据结构的设计非常重要。根据学科核心作者数据库的各类数据性质,可以分为两大类:静态数据和动态数据。静态数据是指在数据库的整体框架下,用户能够直接获取、描述、标引的数据,如作者的个人属性特征和单篇文章的特征数据。动态数据是指在数据库的整体框架下,需要经过数据分析、演算才能得到的数据,如判定指标。根据这一分类,设计学科核心作者数据库的数据结构如图 2 所示。



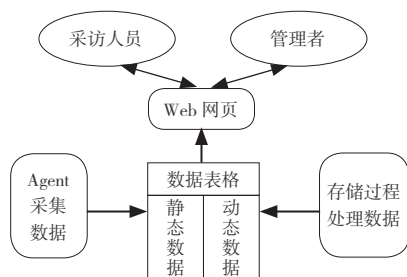


图2 学科核心作者数据库的数据结构

#### 4 应用分析

考虑到航校学科建设的需要,可尝试建立飞行技术专业学科核心作者数据库。由于主要为中文图书的采购决策提供依据,数据库的基础数据采集对象选择中国知网数据库。首先,采集近10年与飞行技术相关的文献,共计2635篇。经过数据整合、查重、消歧后,得到341名作者的相关数据。其中最高产作者的发文量是28篇,根据普赖斯定律,确定候选核心作者的最低发文量为4篇,则满足这一条件的候选核心作者为64名。

通过计算h指数和 $h_m$ 指数,可以发现这些候选核心作者中h指数最高为11,最低为0。考虑到飞机技术学科属于理工科,与图书情报等社科学科不同,经咨询专家,确定h指数为2及以上的作者为核心作者,共计36名。同时,计算这些作者的 $h_m$ 指数,给出作者排序,作者的排序可以为采访人员进行图书采购决策提供依据。

#### 5 结语

学科核心作者数据库的构建研究在国内尚处于起步阶段,研究以关系数据库为基础,以发文量、h指数、 $h_m$ 指数等文献计量学标准作为判定算法的依据,融合Web信息抽取、数据挖掘技术,形成学科核心作者数据库的整套构建策略。学科核心作者数据库的建立,以IT技术代替了人工数据整理,具有自动化、高效率的特

点,可以为采访人员的采购决策提供客观性的数据支持,从而使采访人员摆脱学科知识的局限,提高采访工作的质量和效率。同时,学科核心作者数据库的建设对于图书馆的特色馆藏建设、专业特色数据库建设和机构知识库建设也有一定的帮助。

(来稿时间:2014年12月)

#### 参考文献:

1. 杜秀杰,葛赵青,刘杨等.基于著者索引的高校学报核心作者群分析.编辑学报,2006,18(5):366-368
2. 方太强,周蓉,胡英等.我国图书馆学情报学核心作者分析.图书情报工作,2005(1):69-73
3. 赵基明,舒明全等.基于CSSCI的《中国图书馆学报》h指数及核心作者测定.中国图书馆学报,2008(2):98-102
4. 龚舒野.基于h指数和 $h_m$ 指数的《情报科学》核心作者分析.情报科学,2013(1):82-85,95
5. 邱均平,周春雷.发文量和h指数结合的高影响力作者评选方法研究.图书馆论坛,2008(6):44-49
6. 苏志芳,张建中,胡惠芳等.基于模糊综合评判的中文社科图书“核心作者”决策研究.图书情报工作,2010(1):42-45,41
7. 蔡璐.基于学科分类的高校图书馆核心馆藏规律的实证研究——以高等教育学科为例.图书情报知识,2012(4):106-110
8. 沈艳红,吴信岚等.学科馆员如何利用cnki开展采访工作——以食品学科为例.图书馆,2012(3):105-106,109
9. The Clergy of Church of England Database. [2014-07-13]. <http://www.theclergydatabase.org.uk/index.html>
10. In the first person. [2014-07-14]. <http://www.inthefirstperson.com/firp/index.shtml>
11. 王晓山.科技名人数字图书馆的实践与探索——以老科学家学术成长资料数据库建设为例.图书情报工作,2013(2S):79-82
12. 张学梅. $h_m$ 指数——对h指数的修正.图书情报工作,2007(10):116-118,16

## Research on Building Strategy of Subject Core Author Database for the Demand of the Literature Construction

Zhu Yiting

(Library of Civil Aviation University of China)

**[Abstract]** Based on relational database, acquisition of basic data from the mainstream database through the web information extraction technology, data integration, checking, disambiguation by data mining technology, and then determining the core authors and authors sort according to the quantity of published articles, h-index and  $h_m$ -index, finally, subject core author databases are constructed in this paper. The aim is that providing objective data support for the book acquisition work. Besides, in order to illustrate the practical effect of database, the paper also takes the subject core author database on flight technology as an example.

**[Keywords]** Core authors H-index  $H_m$ -index Relational database Web information extraction

**[作者简介]** 朱轶婷(1979-),女,硕士,中国民航大学图书馆副馆长。