情 报 学 报 ISSN 1000 - 0135
第33卷第10期1022 - 1031,2014年10月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000 - 0135
Vol. 33 No. 10, 1022 - 1031 October 2014

doi:10.3772/j.issn.1000-0135.2014.010.002

领域专家库系统构建研究[1)]

杜 晖[1] 邱均平[2]

(1. 武汉商学院商贸物流学院,武汉 430056; 2. 武汉大学信息管理学院,武汉 430072)

摘要 领域专家库系统的建设是符合我国人才发展战略规划中"建设人才信息网络,评价人才与发现人才相结合"指导思想的重要应用研究。文中从特征描述、科学评价和关联聚合三个维度构建了一个通用的领域专家指标体系,并详细阐述了各指标的计算方法,然后在此基础上通过对学术数据库的信息采集、整理、分析与挖掘的自动化处理,筛选指定领域专家,智能识别其研究领域,从而构建领域专家库系统。特别是基于专家间的各种关联关系实现了对相关领域专家资源的深度聚合与可视化揭示,能够为科研工作者和科研管理部门提供相应的信息服务和决策支持。

关键词 知识网络 专家库 馆藏数字资源 资源聚合 可视化

Research on Construction of Domain Expert Database System

Du Hui[1] and Qiu Junping[2]

(1. *School of Business &Logistics*, *Wuhan Business University*, *Wuhan*, 430056;
2. *School of Information Management*, *Wuhan University*, *Wuhan*, 430072)

Abstract Construction of domain expert information systems is an important application research, which that "talent information network should be built so as to combine the evaluation of talents, is in align with the statement in China's strategic plan for talent development, with idontification of talents". From the three dimensions of feature description, aggregation based on association and evaluation, a general expert indicator system was constructed in the scientific&research field, and expounded the calculation method of each indicator in detail. Then an intelligent information service platform of scientific research evaluation was constructed based on the evaluation system, which achieved a variety of academic databases information collection, collation, analysis and mining automation. By automatically filter specified domain experts, intelligently identify their field of study, so as to construct domain expert database system. Especially the implementation of the depth of aggregation of domain experts resources and visualization based on various association between expert, and can provide information services and decision-making support to the scientific researchers and the department of research and development management.

Keywords knowledge network, expert database, library digital resource, aggregation of information resource, visualization

收稿日期:2014 年 3 月 25 日

作者简介:**杜晖**,男,1979 年生,讲师,武汉商学院商贸物流学院,主要研究方向:信息系统与信息计量。E-mail:33526559@qq.com。**邱均平**,男,1947 年生,博士生导师,武汉大学信息管理学院,主要研究方向:信息计量与科学评价。

1) 基金项目:本文系国家社会科学基金重大项目"基于语义的馆藏资源深度聚合与可视化展示研究"(项目编号:11&ZD152)的研究成果之一。武汉商学院 2014 年度重点科研课题"基于知识网络的领域专家聚合系统研究"(编号:2014A007)。

1 引 言

国务院2010年6月颁布的《国家中长期人才发展规划纲要(2010~2010年)》中明确提出,为了加强人才工作基础性建设,推进人才工作信息化建设,必须建立人才信息网络和数据库,为此政府需要积极支持社会各方力量建设完善面向市场的各类专业人才数据库和人才管理系统。同时纲要还提出,在体制机制创新中必须创新人才评价发现机制,要完善人才评价标准,注重靠实践和贡献评价人才,把评价人才和发现人才结合起来[1]。本文首先提出了一个通用的科研领域专家描述模型,并在此基础上构建了一个基于知识网络的领域专家库系统,在微观层面通过专家间的各类关联实现对广泛分布的专家知识的有效揭示和深度聚合。该系统依托各类网络学术数据库构建各领域的专题文献库,通过识别、筛选出各领域的杰出专家,并分析得出专家的各项指标数据(包括特征描述、关联聚合、科学评价三类),在此基础上构建领域专家库。各领域专家库的建成有助于各领域研究者加深对同行的了解,也有助于科研管理部门和社会大众把握各领域研究人员分布情况和遴选相关领域的专家,能为各科研机构和高校的人才选拔、招聘以及项目团队组建提供定量依据。

2 方法论:指标体系及测度

本文提出了一个针对科研领域专家的通用描述模型,其指标分为三大类,分别是:描述性指标(提供专家基本特征数据但不参与评价)、聚合指标(用来测度专家间的各种关联程度并聚合成相关社区不参与评价)以及评价性指标(用来评价专家科研绩效生成排序的指标),后文将基于此模型实现领域专家系统的构建。表1列出了设计的全部三类指标及其基本含义,专家库系统的后续功能(如聚合和评价等)的实现将灵活的基于表中的部分指标的定量测度。

表1 专家指标体系

一级指标	二级指标	指标说明
描述指标	度中心度	中心性指标反映学者所在领域的地位和社会影响力
	中间中心度	节点的中间中心度测量的是该点在多大程度上控制他人之间的交往。测量的是该学者对资源控制的程度
	结构洞约束系数	该指标是衡量节点控制信息资源能力的指标。结构洞的存在使得连接两点的第三者扮演中间人的角色,拥有越低的结构洞约束系数的节点,越可能成为结构洞,越具有获取多样化知识的能力,是潜在的创新节点[2]
	媒介角色系数	也是对节点创新潜力的评价,不同的是媒介角色系数适用于对已经分群的节点进行评价。通过识别在子群内部或子群之间起到不同媒介作用的节点,其中包括对边界跨越者所作贡献的测度[2]
	H 指数	*H* 指数是一个衡量科学家终身成就的指标,该指标的奥妙在于通过一种简单的测量,且仅用一个单一指标值达到了描述生产率和影响力两方面的效果[3]
评价指标	发表论文数	该指标能在一定程度上反映学者在某领域的科研产出水平
	被引用程度	该指标能在一定程度上反映学者在某领域的科研产出被认可程度
	被关注程度	该指标能在一定程度上反映学者研究领域的热门程度该学者研究成果的吸引程度
	高被引论文数	该指标反应了学者的成果产出质量
	影响因子系数	该指标反应了学者的成果影响力,也反应了成果质量
	成果利用率	该指标反应了学者的学术水平的被认可程度
聚合指标	作者关键词耦合强度	该指标从学者的标注关键词(知识产出)的行为建立学者之间的潜在关联关系
	作者引文耦合强度	该指标从学者的引用参考文献(知识吸收)的行为建立学者之间的潜在关联关系
	作者兴趣耦合强度	该指标通过建立学者的兴趣向量空间模型
	作者合作强度	该指标从学者的合作次数反应学者间的直接关联
	作者共被引强度	该指标以两名学者共同被第三方作者引用来体现两者的某种关联,是实现学术社区聚合探索学科知识结构的常用分析方法

2.1 专家描述指标及其测度

(1)度中心度

度中心度指标代表在科研合作网络中共有多少个直接合作者。也可以采用加权合作中心度(即总合作次数)来测量。通过计算合作网站中该学者所代表节点的度数(Degree),即连接该节点的边的数目,节点 i 的度:$K_i = \sum_{j=1}^{n} A_{ij}$。该指标能衡量结点在网络中的重要性和影响力[4]。

(2)中间中心度

在科研合作网络中某学者所代表节点 i 的中间中心度为每对节点经过点 i 的捷径数除以每对节点间的捷径数。节点的中间中心度(Betweenness Centrality)计算的是某节点占据其他两个节点之间最短路径上的能力,即节点作为信息枢纽的能力。中间中心度的计算公式为:$B_u = \sum_{ij} \frac{\sigma_{(i,u,j)}}{\sigma_{(i,j)}}$,其中 $\sigma(i, u, j)$ 表示节点 i 与节点 j 之间经过节点 u 的最短路径的数量,$\sigma(i, j)$ 表示节点 i 与节点 j 间最短路径的总数量[4]。

(3)结构洞约束系数和媒介角色系数

这两项指标是基于知识流通的评价指标,它们和中间中心度指标都可以作为学者合作网络中介性的评价指标,其中,中间中心度指标用于评价那些对知识的快速流动起到重要作用的节点,而结构洞约束系数和媒介角色系数用于评价节点的创新性[2]。目前这两项指标的自动化计算尚处于研究中。

(4)H 指数

当且仅当某学者发表的 N 篇论文中有 $h(h <= N)$ 篇论文每篇至少获得了 h 次的引文数,其余的 $N-h$ 篇论文中各篇论文的引文数 $<= h$ 时,此 h 值就是该学者的 H 指数。计算方法为将某位学者在领域文献库系统中的全部文献按被引用次数降序排列,当某篇文献的序号(从 1 开始依次递增)正好等于该文献的被引用次数时,该值即为该学者在领域专家库系统的 H 指数。

2.2 专家聚合指标及其测度

(1)作者关键词耦合强度

关键词集合(去重)交集法采用了刘志辉的定义[5](不考虑词频区别,直接计算两位学者在文献库中的相同关键词数),是绝对耦合强度,将作者 A 和作者 B 发表的论文的所有关键词看成一个集合 $S1$ 和 $S2$(关键词无重复),取 $S1$ 和 $S2$ 的交集,交集的数目即为两者的耦合强度值。公式为:$NAKC = S1 \cap S2$。最小耦合值加权算法则考虑到了关键词词频分布对作者间关系的影响,借鉴了马瑞敏关于作者耦合强度的最小耦合值算法[6],两位作者 A 和 B 在文献库的发文集合中,每个耦合词 i 在两位作者的关键词集合 S_A 和 S_B(没有去重)中出现频次值的最小值(即最小耦合值)的累加而成,公式为 $DAKC = \sum \min(S_{Ai}, S_{Bi})$。

专家关键词耦合强度还可以采用相对量的测度,如采用 Jaccard 相似度通过计算交集的大小来获得集合之间的相似度,学者 A 和学者 B 的关键词集合分别用 C_A 和 C_B 代表,那么学者 A 和学者 B 的 Jaccard 相似度为 $|C_A \cap C_B| / |C_A \cup C_B|$(取值范围在[0,1]之间)。还可以通过对作者耦合次数进行标准化运算(克服了关键词数目较多的作者之间耦合强度也可能较高的意外情况),转化为取值在[0,1]之间的一个相似性系数(AKC_strength)。计算公式如下:

$$\text{AKC\_strength} = \frac{couple\_ab}{\sqrt{(num\_a * num\_b)}}$$

其中,$couple\_ab$ 表示作者 a 和作者 b 的关键词耦合频次,num_a 表示作者 a 的关键词总数,num_b 表示作者 b 的关键词总数[7]。

(2)作者引文耦合强度

作者引文耦合分析 ABCA(Author Bibliographic Coupling Analysis)通过作者间共同引用的参考文献数目来表示其耦合强度,从学者的知识吸收行为(参考文献引用)来揭示学者间的潜在关联。作者引文耦合强度的计算完全类似于作者关键词耦合强度,也可以采用绝对耦合强度(包括参考文献集合交集法和最小耦合值加权算法)和相对耦合强度(Jaccard 相似度和作者引文耦合标准化系数)两种算法[7]。

(3)作者兴趣耦合强度

通过借鉴文本表示的方法,运用向量空间模型技术构建专家研究兴趣模型并通过夹角余弦来测度专家间的兴趣相似度,从而实现专家基于兴趣的聚合。采用集合来描述学者的特征,由于集合的无序性属性,导致集合里面的每个关键词所起的作用是一样的,但事实是每一个关键词对代表一个学者所起的作用是不同的,因此必须给每一个关键词一个权重,以反映这个元素在描述资源特征方面的不同重要性,所以我们采用向量空间模型来描述学者。向量空间模型将每一个作者用一个关键词组成的向

量来表示,关键词的权重采用 TF/IDF 计算,作者之间的相似度(即作者兴趣耦合强度)通过余弦相似性算法来计算。首先利用 TF/IDF(Term Frequency/Inverse Document Frequency,词项频率/逆文档频率)对作者关键词分布进行归一化处理,然后利用 Salton Cosine 计算作者之间关键词耦合强度构成耦合矩阵[7]。

(4)作者共被引强度

该指标属于可以直接采集到数据的指标,通过两位作者共同被引用的次数来衡量两位学者之间的潜在关系。直接计算两位学者在领域专家库系统文献库参考文献列表集合中共同出现的次数,即可作为两位学者的共被引强度。

(5)作者合作强度

该指标属于可以直接采集到数据的指标,通过作者间的直接合著论文关系,来直接反应学者间的学术关系。直接计算两位学者共同出现在领域专家库系统文献库作者列表中的次数,即为两者的合作强度。

2.3 专家评价指标及其测度

(1)发表论文数

该指标可以根据不同情况采取不同的计算方式:如发文总数(统计周期内发表论文的数量,反应了成果数量)、独立发文数(统计周期内该作者独自撰写论文数量,反映了自主科研能力)和第一作者发文数(只统计一定周期内以第一作者身份发表的文章数量,反映了一定的学术地位或者原创能力)。

(2)被引用程度

该指标可以采用绝对量,统计某位学者发表的论文被统计周期内系统中其它论文引用的总次数(代表了认可程度);该指标也可以采用相对量来测度,如采用被引率(Citations per Paper),能克服有些发文较多的学者相对来说被引用次数也多的问题,计算方法为:Citations per Paper = Times_cited / Num_issue,其中 Times_cited 代表被引用次数,Num_issue 代表该作者发文总数。

(3)被关注程度

该指标可以采用绝对量,统计某位学者发表的论文在统计周期内被下载的总次数(代表作者研究领域的热门程度);该指标也可以采用相对量来测度,如采用下载率(Download rate,反应了该作者文章的吸引程度),能克服有些发文较多的学者相对来说被下载次数也多的问题,计算方法为:Download rate = Times_down / Num_issue,其中 Times_down 代表被下载次数,Num_issue 代表该作者发文总数。

(4)高被引论文数

该指标是通过统计周期内该人员发表的高被引论文(被引用频次在其所在学科内排名前 5% 的论文)的数量来反映的。

(5)影响因子系数

该指标能反映学者在一定周期内学术产出成果的影响力水平(影响因子系数,Impact Factor Coefficient),采用的计算方法为:$IFO = \sum IF / Num\_issue$,其中 IFO 代表该学者的发文影响因子系数,IF 代表每篇文献的影响因子,Num_issue 代表该作者发文总数。

(6)成果利用率

该指标的计算采用一个相对量的测度(成果利用率,Use_ratio),通过测度某位学者在学术数据库系统中被下载的论文的被引用水平,来反应学者的学术水平的被认可程度,计算方法为:Use_ratio = Times_cited / Times_down,其中 Times_cited 代表被引用次数,Times_down 代表被下载次数。

3 专家库系统的构建

3.1 系统思路

唯物辩证法认为客观事物都是相互联系而不是孤立存在的,任何事物之间都存在一定的联系(或关联)。事物之间的关联通常是一个复杂的和多向的网络关系。知识网络理论认为知识网络是由知识节点(知识单元)和知识关联构成的知识体系,反映的是知识单元之间的时空变化关系和逻辑结构关系。知识关联是指知识单元之间存在的各种联系的总和。知识单元采用了广义的概念,不仅包括知识内容本身,还包括各类知识载体及其集合。因此,知识网络中的知识节点可以有多种理解,可以是不同的知识单元,如文献、引文、作者、机构、期刊、国家、地区、学科、主题词和关键词等[8]。本文认为在某一科学研究领域,专家就是一种知识载体,可以认为是一种知识节点,学者间的知识关联及其强度(比如合作关系、共被引关系、耦合关系等)可以看成是联系专家节点的边,这样就构成了一个基于学者关联的知识网络(知识网络也是专家库系统的存在形式)。通过各种领域专家间关联强度的测度,可以实现专家关联的推荐,并且在知识网络中发现紧密

联系的学术社区,从而实现对相关学术信息资源的聚合。

为了构建领域专家库系统,首先要选定特定的研究领域,通过学术期刊系统的数据接口批量导出文献题录,然后解析并截取题录各字段内容构建专题文献库。在专题文献库基础上,进行各种文献信息的分析挖掘,通过调用基本统计分析生成的数据生成各类关联矩阵,然后通过矩阵运算得出有意义的社会变量,同时也可以将各种矩阵数据导出为Excel格式、UCINET专用格式,可以方便的进行更深入的数据分析与挖掘。在专题文献库的各种分析基础上,识别特定领域专家及其研究领域,按照一定规则进行筛选,同时计算专家各个评价指标数据,并调用综合评价模型生成各类专家排行榜;根据专家间的各种关联(如合作强度、关键词耦合、引文耦合等),通过定量分析,提供专家检索、推荐、聚合及可视化展示等相关辅助性信息服务,从而满足科研管理部门及学者对隐藏在海量文献下的领域专家信息的深度揭示需求。

3.2 系统构建

系统为基于Web的*N*层体系架构,采用LAMP平台开发实现,共划分为信息采集、文献分析与挖掘(基本计量分析、网络分析挖掘)和专家库三大模块,其中信息采集和分析模块是基础功能,专家库是核心功能。系统模型如图1所示。

信息采集模块主要完成以各科技文献数据库(目前,本系统的数据源主要来自CNKI中国学术文献网络出版总库、Web of Knowledge,以后将逐步扩展)为数据源的自动采集、入库和标准化处理,自动建立各领域的专题文献库,以满足深入分析与评价的需要,目前已经建立了涵盖基础科学、医药科学、农业科学、工程科技、人文与社会科学的共29个领域的专题文献库共计80万条记录。采集模块同时也支持用户通过交互接口自行上传待分析的文献数据集。文献分析与挖掘模块包括基本文献计量分析和高级网络分析。基本分析功能基于文献计量学、统计学原理,实现对某领域文献库的基本分析,比如论文、作者、期刊、机构、基金等各种基本量的统计及其分布特征描述等,还包括一些基本的社会网络指标的统计。高级文献分析模块主要是发现科研工作者的相互关联,解释其合作模式与规律。专家库采用了专家关联网络(将专家库中的专家按照其各种关系组成一个相互链接的专家网络)、专家地图(以Google地图作为媒介的方式可视化呈现专家信息、区域分布与相互联系)和专家排行榜(综合排行榜和各单项指标排行榜)三种形式来聚合丰富的专家关联及相关文献信息。包括专家识别与筛选、专家检索、专家推荐、专家聚合、专家可视化和专家评价等6个子模块。

4 领域专家库系统实现

4.1 专家识别与筛选

专家识别的实现,首先在领域文献基本分析的基础上,如果作者分布满足洛特卡定律,则按照普赖

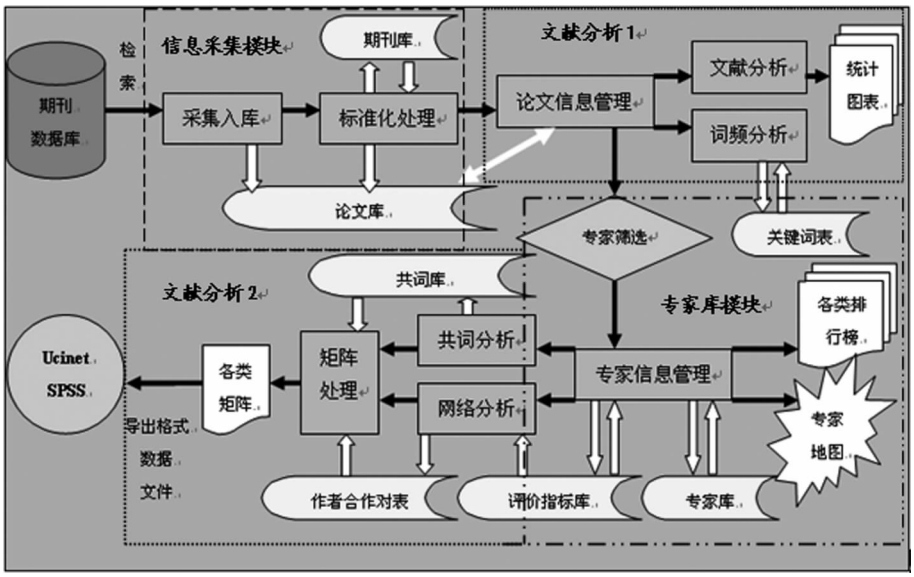


图1 系统模型

斯的理论(核心作者群体的最低发文数等于所有作者中的最大发文数的二次开方的0.749倍)确定核心专家发文下限[9],并提供各种组合筛选条件(如年均发文数、被引频次、下载频次等)由用户定制,然后从专题文献库抽取专家个人基本信息(姓名、职务、所在单位等)和研究专长领域(采用基于词频词典的机械中文分词引擎SCWS,通过对文献标题、摘要等内容进行分词,然后结合关键词一起匹配内置主题词表和禁用词表,并结合文献库的词频统计分析来抽取,构建的专家研究兴趣向量模型可以通过夹角余弦来测算专家的研究兴趣相似度,实现专家的聚类分析[10])等,最后计算该专家的相关特征数据,包括发文指标、引文指标、社会网络指标三大类,经过归一化标准处理后将得分及排名数据一起存入该领域专家库。每一个专家都有其自动生成的专家个人主页(代表了专家网络中的一个节点,整合了其个人学术、社会关系等特征信息及科研评价信息),包括:个人基本信息、研究主题及时间分布、在线沟通工具、各单项指标数据及其排名、合作人脉网络、潜在合作关系挖掘与展示、相关文献列表等。专家个人主页实现了微观层面某位专家及相关文献信息资源的聚合。

4.2 专家检索

专家检索包括常规的多字段(包括按题名、作者、机构、关键词等)组合检索和作者关联网络的检索与呈现(包括作者间的直接合作关系和各种潜在关系的揭示)。任意指定一位专家,系统将输出该专家各项基本信息及排名,并实现相关专家推荐;如任意指定一对专家,系统则输出他们之间的关联程度,并输出相互间联系的所有最短路径(图论的最短路径Dijkstra算法只能输出一条节点间的最短路径,本文参照有关文献对遍历方式和中间数据的结构改进后可以同时输出所有路径[11]),包括联系他们的相关文献信息。

实现方法是指定作者对后,按照三个条件分支进行判断,即是否有直接合作关系(即合作网络距离为1,则输出合作文献列表)、是否有共同合作者(即合作网络距离为2,两者的合作者集合交集非空,则输出共同的合作者列表)、是否合作网络距离大于2(无共同合作者,则构建合作矩阵,计算并输出所有可能最短路径)。如图2所示,揭示两位专家的学术关联,并聚合相关学术信息资源。

4.3 专家推荐

专家间的关联取决于其相似程度,可以从共现或耦合视角来研究其相似程度。共现现象的依据是心理学的邻近联系法则和知识结构及映射原则,比如是否在同一篇论文署名,是否共同被其他作者引用,代表了两者某种程度的相似性;耦合现象代表在某种程度上具有相似的属性特征,也代表了两者之间的某种关联,比如使用了相同的关键词,引用了的相同的参考文献等,通过对因时空障碍而无直接联系的专家之间进行关键词或引文耦合分析,可以发现有共同研究兴趣的潜在合作者,有助于科研项目团队的组建。


专家检索

邱均平

王菲菲

查看专家关联信息 取消

专家:邱均平 专家个人主页 关联推荐
所在单位:武汉大学中国科学评价研究中心;武汉大学信息管理学院;
研究领域:社会网络分析|知识交流|研究热点|文献计量|网络学术信息
专家排名:第2位
合作伙伴:苏金燕|贺颖|余凡|刘艳玲|姜冠图|宋艳辉|杨思洛|曾倩|李慧|温芳芳|王菲菲|瞿辉|秦鹏飞|胡文君|罗力|马凤

专家:王菲菲 专家个人主页 关联推荐
所在单位:武汉大学信息管理学院;
研究领域:社会网络分析|知识交流|学术共同体|作者文献耦合分析|作者研究活力与影响力
专家排名:第81位
合作伙伴:李晶|张晋朝|邱均平

【table1】专家关联信息

两者有直接的合作关系,合作网络距离为1

专家1	专家2	关键词耦合数	关键词耦合度	关键词耦合系数Jaccard	IAKC1(词权重)	IAKC2(Ochiia)	是否合作者	直接合作次数	合作者耦合数	网络距离
邱均平	王菲菲	9	10	0.112500	3.500000	7.914214	是	2	0	distance

相关文献	期刊	作者	机构
基于博客社区好友链接的知识交流状况分析——以科学网博客为例	图书情报知识	邱均平;王菲菲	武汉大学信息管理学院;
基于引证关系的国内情报学领域作者研究活力与影响力分析	图书馆论坛	邱均平;王菲菲	武汉大学科学评价研究中心;武汉大学信息管理学院;

图2 专家检索及相关文献揭示

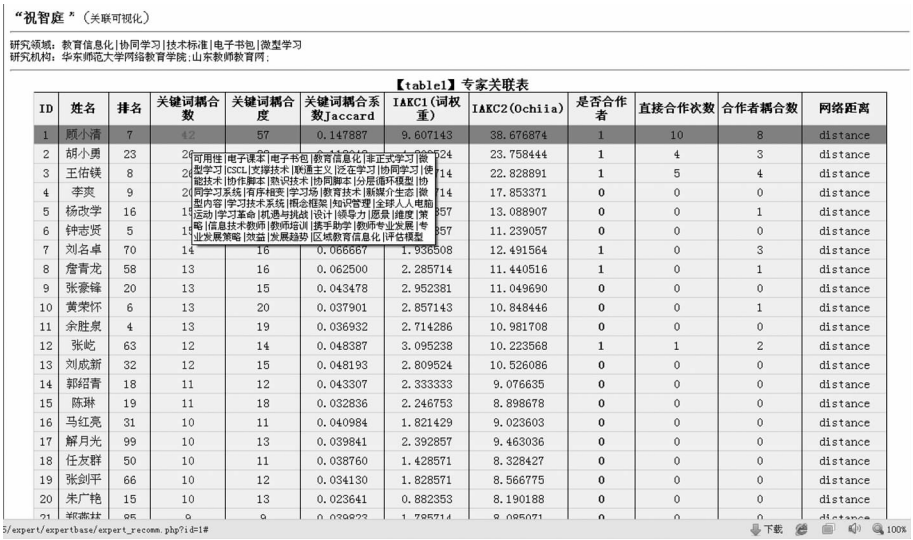
本系统实现了专家合作强度(合作次数、网络距离)、专家关键词耦合(一定程度上代表了专家间共同的研究方向和学术兴趣的相关程度)、专家引文耦合、专家共被引等几个相似度的计量。基于作者的各种相似性度量,在此基础上可以实现各种知识推荐和专家检索功能,可以从微观层面实现以某位专家为中心的相关学术资源聚合。图 3 为根据一个专家节点,按照相关性降序给出的关联专家列表。首先从构建的领域专家库中抽取核心专家,构建专家间的两两耦合对,并生成耦合专家对之间的各种耦合强度指标值。然后根据专家之间的引文耦合强度(反映了知识继承/吸收相似性)、关键词耦合强度(反映了知识产出的相似性)或混合耦合强度(结合前两种耦合强度并加权的一种计量指标)来判断其相似程度,并在用户访问一个专家学术主页时给出按相似程度递减顺序排列的推荐专家列表及其链接,同时可以通过显示隐藏提示层的方式提示学者间相似的解释(如给出两者耦合的对象,如使用相同的关键词或相同的参考文献)。通过对因时空障碍而无直接联系的专家之间进行关键词或引文耦合分析,可以发现没有过直接合作关系却高度耦合的专家,由此识别出有共同研究兴趣的潜在合作者,有助于科研项目团队的组建时参考。

4.4 专家聚合

基于作者的各种关联强度度量指标,通过实现所有专家对的关联程度测度,构建相关矩阵,通过矩阵运算实现网络学术空间的相关作者聚合,从而形成虚拟学术社区,可以从宏观层面实现领域专家相关学术信息资源的聚合。

本文要实现的是,在整个专家库构成的专家知识网络中,实现基于不同耦合指标(如 AKCA)在不同耦合强度(阈值)上面的深度聚合,将相关专家节点聚合成一个个重叠的全耦合网络(派系或称为最大完备子图)。每一个全耦合网络内部各成员节点都是完全耦合关联的(邻接),而派系外的节点都与派系内部各节点无法关联(耦合程度低于阈值)。在此基础上,实现不同聚合深度(即在不同耦合强度阈值)下,专家派系的聚合可视化展示,并通过派系内专家间的关键词集的逐个交集来计算标签(给每个派系生成一个合理的标签)。在特定聚合深度(指定阈值)下,发现文献耦合网络中的最大耦合子网络(派系,至少三个节点且派系中任意一对节点间的耦合强度不小于阈值)。最后提炼出整个派系共同耦合的关键词作为该派系的标签。以关键词耦合数为聚合指标,通过设置不同的聚合深度值,发现不同聚合深度下的最大耦合网络(即派系或最大完备子图)[7],如图 4 和图 5 所示。

专家之间的聚合阈值增加,反映了构成派系的条件要求更高,所以满足条件的派系数量减少,人员构成也缩减了。派系耦合词能够反映这个派系的共同研究兴趣。该功能进一步的研究是构建科研领域本体模型,通过本体模型来标注专家社会网络的节点和边,将更为丰富的语义关系注入专家关联网络,使得网络具有一定智能性,可以利用描述逻辑进行推理和查询,从而可能挖掘出丰富的内容,比如某些隐藏的重要关系和重要节点的精确发现[12]。


“祝智庭”(关联可视化)

研究领域:教育信息化|协同学习|技术标准|电子书包|微型学习
研究机构:华东师范大学网络教育学院;山东教师教育网;

【table1】专家关联表

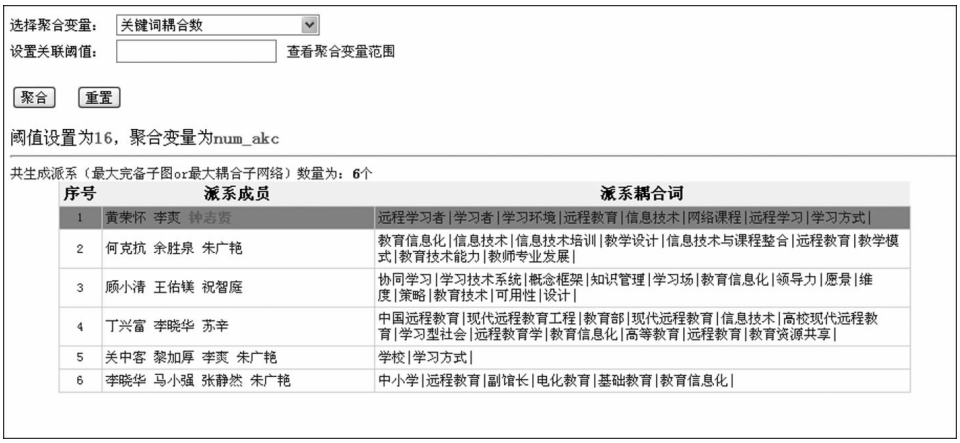
ID	姓名	排名	关键词耦合数	关键词耦合度	关键词耦合系数Jaccard	IAKC1(词权重)	IAKC2(Ochiia)	是否合作者	直接合作次数	合作者耦合数	网络距离
1	顾小清	7	42	57	0.147887	9.607143	38.676874	1	10	8	distance
2	胡小勇	23	26			524	23.758444	1	4	3	distance
3	王佑镁	8	26			714	22.828891	1	5	4	distance
4	李爽	9	20			714	17.853371	0	0	0	distance
5	杨改学	16	15			857	13.088907	0	0	1	distance
6	钟志贤	5	19			857	11.239057	0	0	0	distance
7	刘名卓	70	14	16	0.066667	1.938508	12.491564	1	0	3	distance
8	詹青龙	58	13	16	0.062500	2.285714	11.440516	1	0	1	distance
9	张豪锋	20	13	15	0.043478	2.952381	11.049690	0	0	0	distance
10	黄荣怀	6	13	20	0.037901	2.857143	10.848446	0	0	1	distance
11	余胜泉	4	13	19	0.036932	2.714286	10.981708	0	0	0	distance
12	张屹	63	12	14	0.048387	3.095238	10.223568	1	1	2	distance
13	刘成新	32	12	15	0.048193	2.809524	10.526086	0	0	0	distance
14	郭绍青	18	11	12	0.043307	2.333333	9.076635	0	0	0	distance
15	陈琳	19	11	18	0.032836	2.246753	8.898678	0	0	0	distance
16	马红亮	31	10	11	0.040984	1.821429	9.023603	0	0	0	distance
17	解月光	99	10	13	0.039841	2.392857	9.463036	0	0	0	distance
18	任友群	50	10	11	0.038760	1.428571	8.328427	0	0	0	distance
19	张剑平	66	10	12	0.034130	1.828571	8.566775	0	0	0	distance
20	朱广艳	15	10	13	0.023641	0.882353	8.190188	0	0	0	distance
21	郑燕林	85	9	9	0.039823	1.785714	8.095071	0	0	0	distance

可用性|电子课本|电子书包|教育信息化|非正式学习|微型学习|CSCL|支撑技术|联通主义|泛在学习|协同学习|便能技术|协作脚本|熟识技术|协同脚本|分层循环模型|协同学习系统|有序框架|学习场|教育技术|新媒介生态|微型内容|学习技术系统|概念框架|知识管理|全球人人电脑运动|学习革命|机遇与挑战|设计|领导力|愿景|维度|策略|信息技术教师|教师培训|携手助学|教师专业发展|专业发展策略|效益|发展趋势|区域教育信息化|评估模型

5/expert/expertbase/expert_recomm.php?id=1#

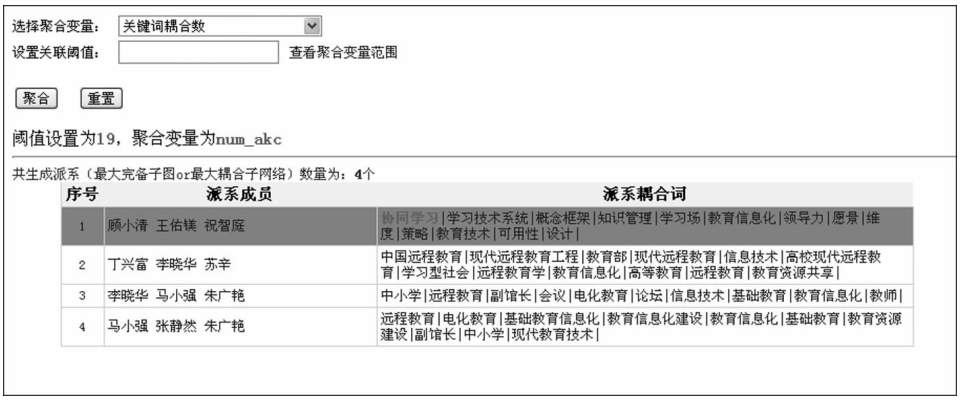
下载 100%

图 3 基于关键词耦合关系的专家推荐


选择聚合变量：关键词耦合数
设置关联阈值：
查看聚合变量范围
聚合 重置
阈值设置为16，聚合变量为num_akc
共生成派系（最大完备子图or最大耦合子网络）数量为：6个

序号	派系成员	派系耦合词
1	黄荣怀 李爽 钟志贤	远程学习者|学习者|学习环境|远程教育|信息技术|网络课程|远程学习|学习方式|
2	何克抗 余胜泉 朱广艳	教育信息化|信息技术|信息技术培训|教学设计|信息技术与课程整合|远程教育|教学模式|教育技术能力|教师专业发展|
3	顾小清 王佑镁 祝智庭	协同学习|学习技术系统|概念框架|知识管理|学习场|教育信息化|领导力|愿景|维度|策略|教育技术|可用性|设计|
4	丁兴富 李晓华 苏辛	中国远程教育|现代远程教育工程|教育部|现代远程教育|信息技术|高校现代远程教育|学习型社会|远程教育学|教育信息化|高等教育|远程教育|教育资源共享|
5	关中客 黎加厚 李爽 朱广艳	学校|学习方式|
6	李晓华 马小强 张静然 朱广艳	中小学|远程教育|副馆长|电化教育|基础教育|教育信息化|

图 4　专家聚合(聚合深度阈值设为 16)


选择聚合变量：关键词耦合数
设置关联阈值：
查看聚合变量范围
聚合 重置
阈值设置为19，聚合变量为num_akc
共生成派系（最大完备子图or最大耦合子网络）数量为：4个

序号	派系成员	派系耦合词
1	顾小清 王佑镁 祝智庭	协同学习|学习技术系统|概念框架|知识管理|学习场|教育信息化|领导力|愿景|维度|策略|教育技术|可用性|设计|
2	丁兴富 李晓华 苏辛	中国远程教育|现代远程教育工程|教育部|现代远程教育|信息技术|高校现代远程教育|学习型社会|远程教育学|教育信息化|高等教育|远程教育|教育资源共享|
3	李晓华 马小强 朱广艳	中小学|远程教育|副馆长|会议|电化教育|论坛|信息技术|基础教育|教育信息化|教师|
4	马小强 张静然 朱广艳	远程教育|电化教育|基础教育信息化|教育信息化建设|教育信息化|基础教育|教育资源建设|副馆长|中小学|现代教育技术|

图 5　专家聚合(聚合深度阈值设为 19)

4.5　专家可视化

专家聚合可视化部分主要采用了自编可视化功能、Google Maps API(实现专家区域分布的展示)和社会网络可视化软件 Netdraw 的批处理接口(实现专家合作网络的呈现)。

通过可视化方式揭示专家聚合结果,要根据数据库存储的实体关联程度数据,绘制网络结构图。功能的实现基于 PHP 的 GD 绘图库。可视化流程为首先创建画布(背景图层)。所有图形处理都基于这个背景图层进行操作。画布实际是在内存中开辟了一块临时的内存区域,用于存储图像信息。第二步为绘制图像。基于背景图层使用各种图像函数设置图像颜色、填充色以及画各种图形。如何合理设置节点、连线等的位置、颜色、大小、粗细、样式等参数,以及相关文本标签的显示,节点的空间布局等各种参数如何与节点本身代表的实体的特征指标取值关联,还要考虑空间布局的合理性,简洁性。最后输出图像并释放资源。完成图像绘制后,需要将图像以某种格式保存到服务器中或者直接生成在线浏览图片,支持格式包括:GIF、JPEG、PNG、BMP 图像被输出后,存储在内存中的存储画布信息的区块没有了,需要及时清除其所占用内存资源。

专家地图是专家库系统的一个用于可视化展示专家相关信息的功能,通过 google 地图加载专家基本信息(姓名、所在机构、研究领域等)和评价信息(发文数、引用次数、度中心度等指标的原始数据及其排名),直观呈现本学科领域的专家的地域分布。实现方法是基于专家的基本信息和各项指标原始数据,首先进行各项指标排名的生成处理,然后按照预置的模板将这些信息动态生成 XML 文档,然后利用 XMLDOM 组件把相应专家各项数据解析出来并利用循环语句在地图上添加 GMarker 地标即可。由于专家数目可能比较多,因此采用了 GMarkerManager 地标管理器对象通过导入地标数组来实现批量地标数据的加载[13]。专家地图的实际应用效果如图 6 所示。

图 6 专家地图效果

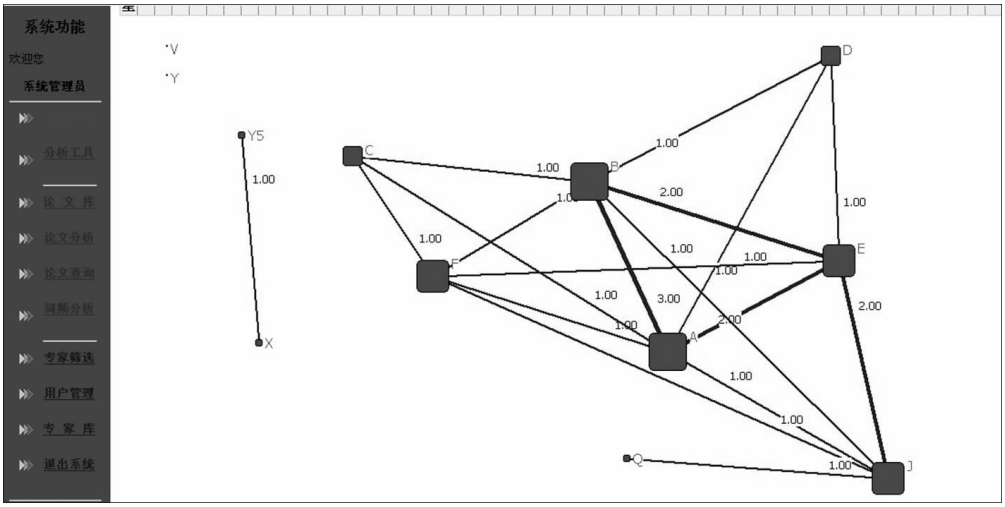


图 7 专家合作网络可视化

基于各种关联关系的专家聚合的实现是通过专家库的专家关联数据，动态生成专家耦合关联矩阵（合作关系、关键词耦合、引文耦合等各种关联指标），然后调用 Netdraw 的绘图接口，实时动态地生成基于耦合关系的专家聚合网络图[14]。图 7 是部分专家的合作网络可视化结果（专家姓名用字母代替），连线的粗细代表作者的关联强度。

4.6 专家评价

评价的本质是一种价值判断，就是根据评价的综合标准和指标体系来确定被评价对象有无价值及价值的大小[15]。该模块主要基于文献计量的各指标来进行评价，可以实现专家的综合评价与各单项指标评价。评价的指标权重确定采用了层次分析法。层次分析法是将定量和定性相结合，将人的主观判断用数量形式表达和处理的应用广泛的综合评价方法。评价模型首先通过层次分析法构建专家评价指标体系并且采用 1 ~ 9 标度方法为每个指标设置权重。通过专家判断并构建指标两两比较的判断矩阵，通过方根法求得矩阵对应的最大特征根和特征向量，对特征向量进行一致性检验通过后，再对特征向量进行归一化处理，可以得到各指标的权重数据[16]。然后调用评价指标库的指标及其权重数据，同时调用专家筛选阶段生成的专家各指标标准得分（指标原始分数归一化后生成的分数，去除了量纲的差异）实现各一级指标排序和综合排序，最后自动生成各种评价报表。考虑到部分指标设置的争议性和局限性，评价指标管理特别提供了交互界面，供用户自主选择评价指标及设置相应权重，系统据此输出不同的评价结果，具有一定的用户需求适应性。

5 系统建设现状及应用效果

目前,本系统已经在武汉大学中国科学评价研究中心各评价项目组(大学及科研机构评价、期刊评价、人才评价等)投入了实际使用,大大提高了科研中的文献数据处理的效率和精确性,同时通过各种统计图表、可视化网络图的输出,可以为科研管理部门提供全面的决策支持;本系统已经为一些高校的某些领域高端人才的引进与评价提供了咨询服务。本文通过对馆藏数字资源中的期刊数据库的分析和挖掘,从作者间各种关系的揭示角度,实现了馆藏文献资源的聚合与展示,通过 Web 形式来满足广大科研工作者、科研管理部门各种信息需求。由于很难获得科研人员在具体某一学科的专利、各类国家基金项目、科技奖励等数据,因此专家评价主要采用期刊论文数据。本系统的国内数据源主要来自 CNKI 中国学术期刊出版总库,国外数据源主要来自美国 ISI 的 Web of Science,而没有涉及博士、硕士的学位论文、会议论文、专利、标准和其他科技成果数据库以及科技奖励和科研项目信息,数据源的多元性和动态性尚需要不断完善。关于专家个体内在特质的各种心理特质和道德水准的衡量,因为采用的测量及评价方法属于间接测量和心理映射,并没有纳入综合排名计算中,而是作为单独的人才测评模块提供用户使用[17],目前也得到了众多教育和科研领域用户的良好反馈。

6 结 语

本文基于知识网络和知识关联理论构建了一个领域专家库系统,在文献分析挖掘基础上,通过揭示专家的各种特征信息及关联,实现了相关专家及文献资源的深度聚合及可视化展示,并给出各种评价结果。目前在支持数据源的多样性方面尚有待提高,今后将进一步扩展数据源的种类,包括对非结构化的 Web 数据源的采集。由于专题文献库的规模并不大,所以关于网络分析的各种算法(如最短路径、社团分析等)还需要针对大规模的数据集进行测试和优化。对专家关联的各种测度指标设计的合理性还需要寻求理论和实验的进一步验证。

参 考 文 献

[1] 中央人才工作协调小组办公室,中共中央组织部人才工作局. 国家中长期人才发展规划纲要(2010 - 2020年)学习辅导百问[M]. 北京:党建读物出版社,2010.

[2] 宋歌. 社会网络分析在引文评价中的应用研究[J]. 图书情报工作,2010,54(14):16-19,115.

[3] 赵基明,邱均平,黄凯,等. 一种新的科学计量指标——h 指数及其应用述评[J]. 中国科学基金,2008(1):23-32.

[4] 朱天,吴斌,王柏. 科研合作网络的重要作者发现[J]. 数字图书馆论坛,2010,75(8):29-35.

[5] 刘志辉,张志强. 作者关键词耦合分析方法及实证研究[J]. 情报学报,2010,29(2):268-275.

[6] 马瑞敏,倪超群. 作者耦合分析:一种新学科知识结构发现方法的探索性研究[J]. 中国图书馆学报,2012(3):4-11.

[7] 杜晖. 基于耦合关系的学术信息资源深度聚合研究[D]. 武汉大学,2013.

[8] 文庭孝,汪全莉,王丙炎,等. 知识网络及其测度研究[J]. 图书馆,2009(1):1-6.

[9] 邱均平,王菲菲. 基于 SNA 的国内竞争情报领域作者合作关系研究[J]. 图书馆论坛,2010,30(6):34-40,134.

[10] 张学义,胡兴雨,吴俊,等. 基于兴趣的科研合作网络演化模型[J]. 计算机工程与应用,2010,46(30):104-107,111.

[11] 徐凤生,李天志. 所有最短路径的求解算法[J]. 计算机工程与科学,2006,28(12):83,84.

[12] 刘臣,张庆普,单伟,等. 基于语义的社会网络关联路径评价及其应用[J]. 情报学报,2011,30(2):172-182

[13] 江宽,龚小鹏. 程序天下 Google API 开发详解:Google Maps 与 Google Earth 双剑合璧[M]. 北京:电子工业出版社,2008.

[14] 陆伟,韩曙光. 组织专家的检索系统设计与实现[J]. 情报学报,2008,27(5):657-663.

[15] 邱均平,文庭孝. 评价学理论·方法·实践[M]. 北京:科学出版社,2010.

[16] 邱均平,程妮. 中国重点大学的网络影响力评价研究[J]. 科学学研究,2009,27(2):190-195,175.

[17] 邱均平,杜晖,党永杰. 基于心理测量的人才评价系统研究[J]. 科技进步与对策,2012,29(10):99-103.

(责任编辑 车 尧)

情 报 学 报 ISSN 1000 - 0135
第33卷第10期1032 - 1040,2014年10月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000 - 0135
Vol. 33 No. 10, 1032 - 1040 October 2014

doi:10.3772/j.issn.1000-0135.2014.010.003

基于作者合作的数字馆藏资源聚合研究

——以法学学科数字文献资源为例

邱均平 季元魁

(武汉大学 中国科学评价研究中心(RCCSE),武汉 430072)

摘要 随着大科学的发展,"信息过载"现象已经日益严重,资源聚合成为提高数字馆藏资源利用效率的重要途径。本文以CSSCI中收录的法学学科数字文献资源作为对象,将作者间的合作为研究切入点,利用社会网络分析中的点度中心性分析、中间中心性分析和子群分析等方法,不局限于对高产作者群体的研究,从而揭示作者间合作关系形成的过程、分布特点及基于作者合作作为聚合方式的独特作用;研究结果也展现了基于作者合作的聚合研究直观、简洁、深入的特点,和对于促进作者间的学术交流、提高资源利用效率的重要意义。

关键词 数字馆藏资源聚合 作者合作 社会网络分析法 法学学科

Integration of Digital Collection Resources Based on Author Collaboration: The Case of Digital Literature Resources in Legal Field

Qiu Junping and Ji Yuankui

(*RCCSE*, *Wuhan University*, *Wuhan*, 430072)

Abstract With the development of information technology, the phenomenon of "information overload" has been increasingly serious, resource aggregation is one important way of improving the efficiency of digital collection resources. This article included in CSSCI law subject digital literature resources as object, based on cooperation between the author study the breakthrough point, using social network analysis in some degree centrality analysis, intermediate centrality analysis, subgroup analysis, showing the author cooperation law subject structure, reveals the relationship between the author formation process and the characteristics of different types of cooperation. At the same time, this paper studies also showed the polymerization research based on the author cooperation intuitive, concise, and the characteristics of the deep, and to promote academic exchange between the author and the significance of improving the efficiency of resource utilization.

Keywords cooperation research of digital collection resources, author polymerization, the method of social network analysis, the law subject

1 引 言

随着信息和电子技术的发展,人们获取信息和加工信息的门槛在不断降低[1]。人们已不满足于仅仅做信息的消费者,也成为信息的生产者。Web 2.0的发展正满足了人们的需求,在这种情况下,互联网上的信息资源正在日益的膨胀。面对网络信息资源的泛滥,各种旨在提高信息资源利用率的研究已成为热点。

收稿日期:2014年5月26日

作者简介:**邱均平**,男,1947年生,武汉大学中国科学评价研究中心教授,博士生导师。主要研究方向:信息计量与科学评价、知识管理与竞争情报,E-mail: jpqiu@whu.edu.cn;**季元魁**,男,1989年生,武汉大学中国科学评价中心硕士研究生。