基于Python技术和TF-IDF算法的科技专家库建设案例研究

杨好[1*] 周长海[2*]
1. 中国科学院科技战略咨询研究院 北京 100190
2. 中国科学院发展规划局 北京 100864

摘 要：本研究以中国科学院科技专家库建设为案例，探讨了在信息安全环境下利用智能技术完善并更新专家库信息、运用综合指标遴选专家的路径，在此基础上总结了中科院科技专家库信息系统的设计与实践。首先，基于Python大数据网络爬虫技术和文献情报分析相结合的方式，补充专家基础数据，并定期更新专家信息；其次，建立专家信誉度评价指标体系；第三，在遴选专家时，运用TF-IDF算法对项目和专家信息进行关联分析，并结合学科分类标准对专家研究领域分类，以提高项目-专家研究领域的匹配度；第四，综合各项关键指标遴选确定最终候选专家；最后，在此基础上设计并开发了中科院科技专家库信息系统，有效提升了专家库管理和专家遴选的工作效率。

关键词：科技专家库 Python技术 TF-IDF算法 专家遴选 专家库信息系统
DOI:10.11842/chips.20220324001

0 引言

同行评议是科技评价的重要方法，遴选高水平、符合要求的同行专家是开展同行评议的重要前提，科技专家库是遴选专家的基础性工具。《国务院办公厅关于完善科技成果评价机制的指导意见》（国办发〔2021〕26号）指出，要创新科技成果评价工具和模式，利用大数据、人工智能等技术手段，开发信息化评价工具。充分利用智能技术，优化完善科技专家库建设，不断提升同行评议的质量和效率。

国内相关研究学者在专家数据资源获取、专家动态淘汰机制建立、专家遴选等方面做了大量研究工作，通过对专家库建设的相关文献调研分析发现，在专家数据资源获取方面运用最多的是大数据的网络爬虫技术，但因为基础数据的海量性、多类型性以及分散性，数据处理困难程度较高，需要引入数据挖掘分析技术实现对海量数据的有效分析和挖掘，找出其中高价值的信息资源；评审专家的淘汰机制方面有研究学者利用数据挖掘聚类方法把信誉度较差的专家从评审专家库中逐渐淘汰；专家遴选方面，评审专家智能推荐算法和遴选系统被运用，例如国家自然科学基金委基于科研大数据、AI等技术构建了科学基金项目管理系统。

中国科学院科技专家库自建立以来，有力支撑了多项评估评审工作。本文基于专家库建设和使用的实践经验，分析存在的不足，开展在信息安全环境下通过智能技术优化科技专家库建设的方法研究，不断提升遴选

* 杨好，硕士，研究方向：计算机科学技术在科技评价中的应用，数据分析、数据库与信息管理技术以及研究所评价等；周长海（通讯作者），博士，副处长，研究方向：科技评价与科技政策，科研机构评价，科技成果评价和奖励评审等。

专家的质量和效率。文中介绍了通过文献情报分析和大数据网络爬虫技术相结合的方式定向获取所需数据，有效避免了海量数据处理的复杂性和不确定性；利用Python技术针对指定网址抓取目标字段内容，存储最近信息以实现专家信息的更新；探索建立专家信誉度评价指标体系的方法；借鉴专家-项目关键词相似度计算方法，并结合学科分类标准对专家研究领域进行分类，以提高项目-专家研究领域的匹配度；综合各项关键指标确定最终候选专家。在此基础上设计并开发中科院科技专家库信息系统。

1 中科院科技专家库现状分析

1.1 建设现状

中国科学院科技专家库（以下简称“专家库”）建立于2005年，由中科院第三方评估研究中心维护。截至2022年，共有国内专家1万余位，国外专家4000余位。专家学科覆盖了40个一级学科，449个二级学科，1613个三级学科，研究领域包含了数理与化学、天文与空间、地球与环境、生命与健康、农业与生物多样性、信息、材料、制造与工程、能源、海洋、前沿交叉11个领域。基本形成了学科领域布局全面、国内外专家兼顾、专家信息定期更新、数据不断扩充的高水平专家库，有力支撑了多项评估评审工作。

1.2 存在不足

目前，专家库建设中还存在一些不足。

首先，部分研究领域专家数量较少。随着科技新方向、新技术的快速发展，对相关研究领域的专家需求日益旺盛。国家重点发展的部分基础核心领域、国家重大科技基础设施类、行业和用户等方面专家数量相对较少。

其次，专家库信息更新不及时。现阶段主要通过人工定期核查等方式更新专家信息，存在信息更新不及时的情况，影响遴选推荐专家的效率和准确性。

专家库尚未有效建立专家信誉度评价机制。对专家的信誉度评价很重要，需要在评审实践中建立并完善专家信誉度评价指标体系，判断专家是否符合评审要求，对信誉度较差的专家降低其遴选的优先级，直至最后淘汰出专家库。

专家库缺乏信息安全技术保障。专家库中涉及到专家的个人信息，需要对专家数据进行严格地保密管理。尤其在专家信息传输过程中，要对传输信息进行加密处理，即使出现文件被盗取的现象，盗取人员没有密匙也无法得到其中的真实信息。另外建设科技专家库信息系统，必须要保障系统的网络安全环境，对防火墙技术、VPN网关设置、登录人员访问权限等都需要有严格要求。

2 利用智能技术扩充专家库基础数据

2.1 数据分析

针对专家库目前现状，需增加部分研究领域的专家，国家重点发展的基础核心领域如人工智能、量子信息、集成电路、生命健康、脑科学、生物育种、空天科技、深地深海等前沿领域；再如一些国家急迫需要和长远需求出发的关键核心技术，如新发突发传染病和生物安全风险防控、医药和医疗设备、关键元器件零部件等；国家重大科技基础设施如高能同步辐射光源、硬X射线自由电子激光装置、转化医学研究设施等大科学装置类；形成系统解决方案、产生重大社会经济效益的应用研究成果类，需要重点对这些领域的专家开展信息采集和补充。

2.2 数据抓取

大数据技术包括大数据采集、大数据预处理、大数据存储、大数据分析。数据采集常用的方式是大数据采集，即对各种来源的结构化和非结构化海量数据，所进行的采集，包括网络数据采集、数据库采集、文件采集。常用的是网络数据采集，指借助网络爬虫或网站公开API，从网页获取非结构化或半结构化数据，并将其统一结构化为本地数据的数据采集方式。

采用大数据网络爬虫技术去网页抓取专家信息时，如精准抓取，即直接抓取专家信息列表内容，因为每个目标网址结构类型都不一样，网站编程是基于html+css+js编写，程序员写爬虫程序时匹配id和class有一定难度，精准抓取需要定位id和class。如模糊抓取，可以从body开始全部抓取，但因为数据的海量性和复杂性，获取的数据需通过机器学习、数据挖掘等大数据处理与分析技术进行数据分析，会花费大量的时间，难以快速、准确定位高水平专家信息。

因此，本文通过利用文献情报分析技术和大数据技术有机结合的方式予以解决。首先借助文献情报工具分析目标研究领域处于世界领先位置的大学、科研机构和人员。目前常用的文献分析工具有Citespace、TDA、Vosviewer等，可以通过生成多种基于文献计量关系的图谱，分析某个领域的研究现状和研究进展，进而锁定某些研究领域处于领先位置的机构。然后利用大数据网

图1　数据抓取路线图

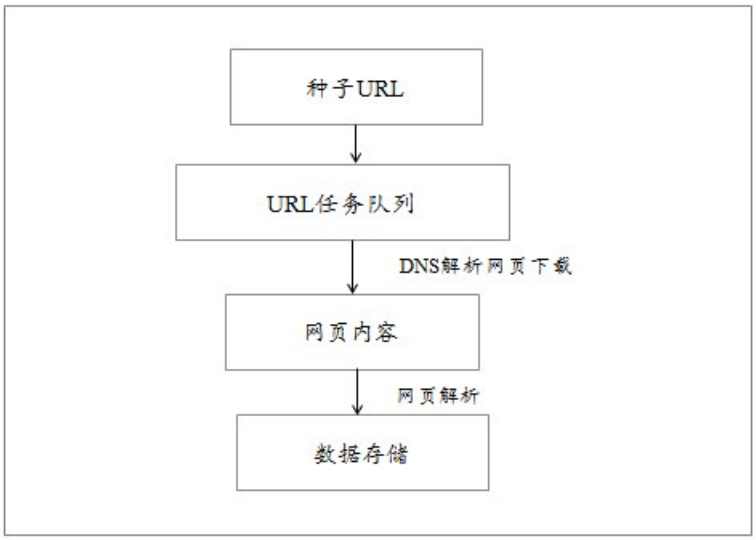

图2　网络爬虫的基本工作流程

络爬虫技术在不突破目标网站访问权限，保障网络数据信息安全的前提下，有针对性地抓取专家信息。最后，通过人工校验方式核查数据抓取的有效性和准确性。

图1用以阐述数据抓取的主要工作路线。

3　利用Python技术更新专家库基础数据

本文主要介绍通过网络爬虫Python技术[1][2][3]定向抓取网页的专家信息，即利用Python技术针对指定网址，抓取目标字段内容并存储到EXCEL中，通过再次抓取的信息与前一次抓取记录对比，将更新的信息存储而不重复存储，实现专家信息的更新。

Python是面向对象的程序设计语言，由于Python有很多可用于数据抓取的第三方工具库，所以是数据抓取的主要语言。Python常用的工具库包括：(1)Request库。自动网络请求提交，抓取HTML页面。(2)Beatiful Soup库。解析HTML页面，提取页面信息。(3)Re库。正则表达式库，提取页面关键信息。大数据网络爬虫技术常用的爬虫框架有Scrapy(Python)、WebMagic(Java)、Crawler4j(Java)。

网络爬虫工作首先要明确抓取对象及定向内容，本文的抓取对象是专家的个人信息网址，存储在专家库“数据来源”字段中，定向内容共有8个字段，包括姓名、单位、职称、研究领域、职务、荣誉称号、电话号码、邮箱。设计出适合抓取对象的专门爬虫工具，抓取所需信息后将数据存储到EXCEL，存储在数据库中。基本工作流程如图2所示：

程序设计思路：

(1)用Python加载专家库文件，用for循环语句读取数据，放到爬虫任务队列中。

(2)通过request库发送URL请求，解析Web页面。

(3)通过Beautiful Soup库解析返回的网页内容，提取数据。

(4)从第三方库引入xlwt模块，将数据写入Excel中。

最后，专家库文件命名规则以文件名和日期流水号命名，可对比历史抓取记录，实现专家信息的更新。

4　建立专家信誉度评价指标体系

4.1　探索专家信誉度评价指标内容

专家信誉度评价指标内容应包含5个维度：评审有效率[4]、评审离散率、评审态度、参与评估工作次数、缺席次数。

4.1.1　评审有效率

评审有效率是反映专家评审有效程度的指标，该指

标可以一定程度上反映出专家在评审过程中是否认真填写评审意见、敢说真话等内容，而并非说好话、空话较多，这些内容可以记录在专家库中作为遴选专家的重要参考依据。

为量化该指标，将评审有效率分为命中率和成功率两个要素。命中率是指专家评审结果与最终结果一致的项目数占项目总数的百分比。式(1)为命中率计算公式：

$$P_j = \frac{G}{E} \times 100\% \quad (1)$$

其中，P_j表示评审专家j的命中率，G表示专家的评审结果与最终结果一致的项目数，E表示专家j所评项目总数。式(2)为专家评审命中率得分：

$$H_{P_j} = P_j \times 100 \quad (2)$$

其中，H_{P_j}表示第j位专家评审命中率得分。

成功率是指被评上的项目中经实施成功的项目数占被评上项目总数的百分比，它反映了专家评审结果的正确性及整体评议水平的高低，一般用在项目或机构的咨询论证评议，或一些奖励的预评审工作中。

4.1.2 评审离散率

评审离散率是指不同项目评审结果间的差异程度，一般包括横向离散率和纵向离散率两个要素，该指标能较好地反映出专家的评审水平及评审公正性，本文主要介绍横向离散率。同行评议一般按照评审对象的学科特点采取分类评审，横向离散率指通过对某专家与其他专家的比较来衡量其对某一个项目的评审质量。将专家对项目的评分结果收集、统计后可建立矩阵(公式3)。其中m为专家个数，n为项目总数，$X_{ij}(1\leqslant i\leqslant n;1\leqslant j\leqslant m)$表示第j位专家在第i个评审项目上的评分(下同)。

$$[X]=\begin{bmatrix} X_{11} & \cdots & X_{1m} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{nm} \end{bmatrix} \quad (3)$$

基本原理是利用统计学中偏差的概念，即用个别测定值与测定的平均值之差来衡量测定结果的精密度高低，偏差越小，说明测定结果精密度越高。

取$X'_{ij}=\left(X_{ij}-\frac{1}{m}\sum_1^m X_{ij}\right)$，则可得到评分的偏差矩阵公式(4)：

$$[X']=\begin{bmatrix} X'_{11} & \cdots & X'_{1m} \\ \vdots & \ddots & \vdots \\ X'_{n1} & \cdots & X'_{nm} \end{bmatrix} \quad (4)$$

由偏差矩阵可得到项目i的评分均值$\overline{X_i'}$(公式5)及其方差γ_i^2(公式6)：

$$\overline{X_i'}=\frac{1}{m}\sum_{j=1}^{m} X'_{ij} \quad (5)$$

$$\gamma_i^2=\frac{1}{m}\sum\left(X'_{ij}-\overline{X_i'}\right)^2 \quad (6)$$

方差描述了随机变量对于数学期望的偏离程度，若γ_i^2数值大则表示专家组成员在该项目得分上存在较大分歧。

以某奖评审举例说明。评审按学科领域分为A、B、C、D共4组，每组10个项目，每个项目约1位小同行专家，共10位专家，根据上述公式可计算出某位专家在本组中对不同项目的得分偏差，偏差越小说明对某个项目的评分结果精密度越高，由偏差计算出专家在本组中的评分均值。

此外如果专家组成员在某个项目上γ_i^2数值较大，科研管理者应给予重视，认真研判是否专家的研究领域方向存在非共识的可能，包括可能存在专家结构不合理，或者项目本身存在分组不合理的情况，但评审结果与申报项目类别和数量、学科分布均有很大关系，如何找到更好的方式，需要在工作中进一步研究和探索。

4.1.3 评审态度

专家的评审态度对评审结果的客观性和公正性有着重要影响。

现场评审或线上函评可尝试通过Zabbix[5][6]监管专家登录系统的评审状态，Zabbix提供了数据存储功能，可以通过自定义监控项实现对多种数据类型的采集，其扩展能力强，可查看历史数据并进行二次分析，进而为判断专家评审态度提供参考。例如可以通过Zabbix记录专家查看评审材料花费的时间、反馈意见是否在规定的时间范围内等信息，进而判断专家评审态度是否认真。将两者对应的理想时间和实际时间分别赋值。

函评中评审专家未在规定时间内完成评审任务的情况很常见，所以该项内容的监测对提升评审效率也很重要。

4.1.4 参与评估工作次数

参与评估工作次数是指专家多次积极参加评审工作的次数，可量化统计并给该指标赋值，同上转化为得分制。

4.1.5 缺席次数

缺席次数是指专家接受邀请后，临时通知评审机构取消评审的次数，可量化统计并给该指标赋值，同上转化为得分制。

4.2 建立专家信誉度评价指标体系

依据上述赋分规则，假设专家的评审有效率分值为

M，评审离散率分值为N，参与评估工作次数分值为O，评审态度分值为P、缺席次数分值为Q。专家进入评审专家库时，信誉度初始值f定为100，专家每参加一次评审工作后，其信誉度就相应更新一次，信誉度的更新机制如下：

（1）信誉度计算公式为：

本次信誉度＝a×M-b×N+c×O+d×P-e×Q（a、b、c、d、e为5个维度的权重，0<e<d<c<b<a<1且a+b+c+d+e=1）（7）

（2）当前信誉度的计算公式为：

当前信誉度＝历史信誉度+本次信誉度 （8）

由于专家的信誉度数值是动态的，需根据专家每次参加评审工作后进行调整，

对得分较低的专家可降低其遴选的优先级，以进一步保障评审工作的客观性和公平性，提升评审质量。

5 综合指标遴选专家

遴选专家是同行评议的关键环节，其中最重要的步骤是根据项目关键词匹配该研究领域的专家，再结合专家的信誉度评价指标、学术水平、利益回避原则等综合信息确认最终候选专家。

5.1 建立项目-专家关键词匹配机制

建立专家与评审对象的匹配关联机制，保障评审质量。本文以评审项目为例展开介绍，机构和个人的评审，方法同理。通过对项目信息和专家信息数据的分析，进而产生项目标签和专家标签，根据两者的相似度初步计算出匹配程度。然后根据综合标签最终确定项目候选专家排序。

5.1.1 构建标签的算法

通过TF-IDF算法[7][8][9]提取项目任务书中的关键词，以此作为该项目的标签。基本原理是，TF-IDF算法为词频-逆向文本算法，TF代表词频，该参数随着词语出现的增多而增长，IDF代表逆文本频率，表示包含关键字的文档在文本中的分布。利用此算法可以计算关键词的权重，即某个关键词的重要性，可以对文本进行分类。

$$\text{TF-IDF}=\frac{m_{kj}}{\sum_k m_{kj}}*\log\frac{M}{m+1} \quad (9)$$

式（9）中m_{kj}代表关键词j在项目k中出现的次数，$\sum_k m_{kj}$代表关键词j在k个项目中出现的总数；M代表项目的总数量，m代表与关键词j相关的项目数量，其基本思想是，如果一个词语在该项目中出现的次数多，但是在该组项目中出现的次数少，说明这个词语对此项目的重要性很高，可以作为关键词。项目任务书的关键词提取是推荐专家是否成功的主要因素。

5.1.2 构建科研项目和专家的标签矩阵

分别构建出科研项目（研究领域）-标签矩阵和专家（研究领域）-标签两个矩阵。

在科研项目（研究领域）—标签矩阵式（10）中，计算某项目使用某个标签的次数，a_{1n}代表项目1使用标签n的次数，以此类推。

$$\text{User\_tags}=\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \quad (10)$$

利用TF-IDF算法可以计算标签对某一项目的重要程度。科研项目u与标签t的关联程度R(u,t)可根据式（11）来计算，其中N用来表示选择过标签t的所有项目数量。

$$R(u,t)=\frac{a_{n,t}}{\log(1+N)} \quad (11)$$

构建专家（研究领域）—标签矩阵。其中b_{11}代表专家1选过标签1的次数；b_{1n}代表专家1选过标签n的次数，以此类推。如式（12）所示。

$$\text{item\_tags}=\begin{bmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{n1} & \cdots & b_{nn} \end{bmatrix} \quad (12)$$

同时结合TF-IDF算法，计算出某一标签对于专家（研究领域）的重要程度。则标签t对专家i的重要程度R(i,t)可根据式（13）计算，M表示选过标签t的专家总数。

$$R(i,t)=\frac{b_{i,t}}{\log(1+M)} \quad (13)$$

最后，根据式（14）预测专家i对科研项目u的匹配程度，记为P(u,i)，并根据计算结果得出按科研项目研究领域关键词所匹配的相近研究领域的专家。

$$P(u,i)=R(u,t)*R(i,t)=\frac{a_{n,t}}{\log(1+N)}*\frac{b_{i,t}}{\log(1+M)} \quad (14)$$

5.2 利用学科分类进一步完善项目-专家研究领域匹配度

科学技术部的国家标准《学科分类与代码》，共包含58个一级学科，573个二级学科，近6000个三级学科。为进一步提高项目和专家研究领域的高匹配度，避免因不同学科之间关键词相近带来的问题，根据《学科分类与代码》对所有专家的研究领域按照一级学科、二级学科（及三级学科）进行分类，此项工作需要研究人员认真查阅专家库中每一位专家近几年来的科研经历、论文、专利等成果信息，才能对专家的研究领域进行准确、科学的学科分类。例如某专家研究领域是“原子、分子反

表1 专家研究领域学科分类

专家	研究领域	一级学科	二级学科	三级学科
专家1	原子、分子反应动力学	化学	物理化学	化学动力学(包括分子反应动力学等)
专家2	低纬强关联系统及自旋系统的性质	物理学	凝聚态物理学	超导物理学

应动力学”,属于一级学科“化学”,二级学科“物理化学”,三级学科“化学动力学(包括分子反应动力学等)”,如表1所示。

利用上述标签算法进行专家和项目关键词的相似度计算之后,依据专家研究领域学科分类将所遴选的专家进一步迭代筛选,以确保项目-专家研究领域匹配度的准确性。

5.3 综合各项关键指标确定专家

确定专家的研究领域后,需要综合考虑专家的信誉度排序、学术水平、合作关系[10]、年龄要求等其他关键指标,最终确定候选专家。图3为遴选专家工作流程图。

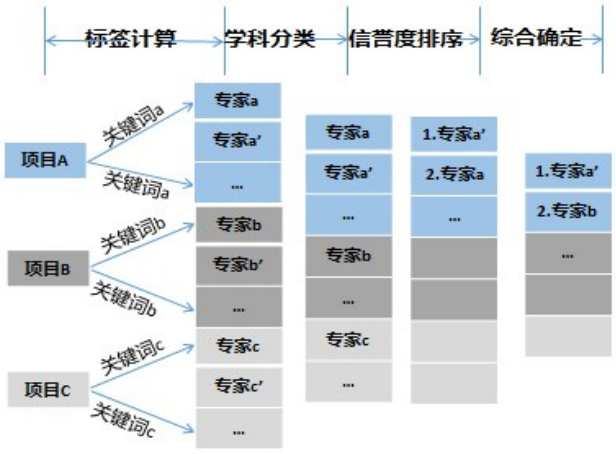

图3 遴选专家工作流程

6 中科院科技专家库信息系统的设计与实践

本章在专家数据库的基础上,基于第2节介绍的利用智能技术扩充专家库数据、第3节利用Python技术更新专家基础信息、第4节构建的专家信誉度评价指标体系、第5节介绍的综合指标确定专家,设计并实现了基于信息安全[11]环境下的中科院科技专家库信息系统。

该系统采用华三(H3C)F1050硬件防火墙,可过滤不安全的服务,只有经过精心选择的应用协议才能通过该防火墙,极大地提高了网络内部的安全性;系统登录采用auth2科技云通行证登录,中国科技云通行证登录账号是使用邮箱绑定注册,并使用https安全连接来保护账号在网络传输过程中的安全性;系统访问权限采用Apache Shiro权限框架,执行身份验证、授权、密码和会话管理等功能,本系统设置只有1位超级用户,为系统的核心管理人员,其他人员登录需经过核心管理人员的授权才能访问数据;数据传输过程中采用base64加密技术。

6.1 系统总体框架设计

本文所构建的中科院科技专家库信息系统结构框架分为3层,自下向上分别是数据存储层、数据分析层、应用层,具体结构框架如图4所示。

6.2 系统主要结构的功能描述

6.2.1 数据存储层[12]

数据存储层是系统的基础层,分为数据采集和数据更新模块,作用是利用网络爬虫Python技术定向抓取目标网址的专家信息,并对采集后的数据进行数据清洗,过滤掉专家信息中不符合标准规范或无效的数据、对数据进行重复性辨别,并对重复数据进行合并或清除处理;格式转换,主要针对不同来源的数据按照统一规则进行转换,包括字符集转换、数据格式转换和代码转换;数据分区、数据分片、构建索引,主要对数据进行数据资产管理,提供各区数据资源的管理,处理后的数据存入对应的数据库。

6.2.2 数据分析层

数据分析层是系统的支撑层,分为信誉度评价指标体系模块、项目-专家领域(方向)匹配模块、综合指标模块。

信誉度评价指标体系模块包含对评审有效率、评审离散率、评审态度、参与评估工作次数和缺席次数等指标的管理,作用是计算专家的信誉度得分并对数值进行由高到低排序。

项目-专家领域(方向)匹配模块的作用是利用标签算法将项目库中项目的关键词和专家科研成果库中专家的领域(方向)关键词进行匹配,找出同领域的专家,并利用学科分类标准进一步对专家进行筛选,规避因不同学科关键词相近错误匹配专家的问题。

综合指标模块的作用是在上述通过信誉度评价和匹配领域(方向)遴选专家的基础上,进一步通过专家基础信息库中荣誉称号分类(如院士、长江学者、杰青、千人等),专家科研成果库中近年来所获奖励、专利等表征专家学术水平的信息,根据不同评审需求进行筛选;同

图4　科技专家库信息系统结构框架

时去除与被评对象有利益合作关系的专家，本文只考虑专家的单位回避及项目合作关系的回避，确定最终遴选专家名单。

6.2.3　应用层

数据应用层是系统的功能应用模块层，主要分为专家信息管理和评审专家遴选模块。

专家信息管理模块分为数据补充、数据更新和数据查询。

评审专家遴选模块根据不同场景的评审分为奖励遴选、专项遴选和机构遴选。因为不同的评审对专家的需求不同，比如奖励评审需要小同行专家较多，机构或专项评审需要一些大同行专家，要对被评审对象的发展前景以及有可能涉及到的交叉学科的发展方向有一定判断。

6.3　系统使用情况分析

中科院科技专家库信息系统自2020年开发以来，抓取专家数据32573条，经筛选有效增加了目标数据4351条，截至目前专家库共有10468位国内专家，4584位国外专家，库中专家信息依托系统开展了常态化更新。系统近年来有力地支撑了中科院的多项重要评估评审工作，专家遴选的成功率由72.5%增至92.1%，为提高评估工作效率、提升专家专业锲合度做出了贡献。

7　结论

数据是核心竞争力，数据的准确性和完整性是构建一个科学高效的科技专家库的根本所在。专家库建设中常见的问题包括以下几点内容：

（1）利用大数据网络爬虫技术可以很快获取大量专家信息，但数据的准确性有待核查和验证；

（2）专家库数据信息更新不及时；

（3）专家信誉度指标体系的内容需进一步完善；

（4）利用专家智能推荐算法遴选的专家，是否为高质量的专家，需要进一步验证；

（5）专家库的信息安全建设需要重视。

本文针对上述问题提出以下解决方案：

（1）利用大数据网络爬虫Python技术，并结合文献分析手段定向获取目标数据，有效提高了获取数据的效率；

（2）实现了定期更新专家库基础数据；

（3）建立了专家信誉度评价指标体系；

（4）遴选专家时，在完善专家研究领域学科分类的基础上，借鉴关键词匹配智能算法，并结合其他重要关键指标综合遴选专家，进一步保证了同行专家的高匹配度，遴选出符合条件的优秀专家；

（5）在信息安全环境下建设专家库信息系统，保障专家信息的安全性。

总之，恰当引入计算机智能技术有目标地补充每年新增选的优秀科技人才，并定期更新专家库数据，打通专家库建设的数据流，建立专家信誉度评价机制，完善专家库闭环管理，保证专家库建设数据的准确性和完整

性，是持续优化和完善专家库建设、遴选和推荐高水平评审专家的重要依据，对科技工作的质量、水平和方向真正发挥指导和把关作用，是保障科技评价质量的关键所在。

参考文献：

[1] 黎妍,肖卓宇.引入Scrapy框架的Python网络爬虫应用研究[J].福建电脑,2021,37(10):58-60.
[2] 任洛漪.基于Scrapy的商务网站数据抓取[J].信息与电脑,2018(19):56-57.
[3] 罗安然,林杉杉.基于Python的网页数据爬虫设计与数据整理[J].电子测试,2020(19):94-95.
[4] 张莹,王志浩.基于层次分析法的科技项目同行评议专家综合评价体系构建研究[J].昆明理工大学学报（社会科学版）,2021,21(5):88-96.
[5] 孙卫真,王詠,向勇.基于工作流的监控系统灵活性增强方法[J].计算机工程与设计,2019,40(9):2704-2711.
[6] 甘丽,胡昊.基于Zabbix的高校数据中心云监控系统设计与实现[J].计算机应用,2021(5):87-89.
[7] 曹滔宇.科技项目评审专家智能遴选系统的研究与实现[D].北京邮电大学,2021.
[8] 鞠亚军.基于智能推荐算法的科研管理系统的研究与开发[D].扬州大学,2021.
[9] 张雯.基于知识图谱的领域评审专家推荐[D].北京信息科技大学,2021.
[10] 戴石钰.基于论文合作关系的科研项目专家回避模型研究[J].图书情报工作,2020(18):125-132.
[11] 齐邦强.信息安全技术在电子信息工程中的应用与解析[J].信息与电脑(理论版),2019,31(19):197-198+201.
[12] 周炜翔.面向信息安全领域的项目评审专家推荐方法研究[D].北京信息科技大学,2020.

Case Study on the Construction of Science and Technology Expert Database Based on Python Technology and TF-IDF Algorithm

Yang Hao[1], Zhou Changhai[2]

1. Institutes of Science and Development, Chinese Academy of Sciences, Beijing 100190

2. Bureau of Development and Planning, Chinese Academy of Sciences, Beijing 100864

Abstract: By taking the construction of the science and technology expert database of Chinese Academy of Sciences as the case, the study explored the path of using intelligent technology to improve and update the expert database information and using the comprehensive index in the information security environment, and on the basis of this, summarized the design and practice of information system of the science and technology expert database, Chinese Academy of Sciences. Firstly, based on the combination of the Python big data web crawler technology and literature information analysis, the expert basic data is supplemented and the expert information was updated regularly. Secondly, the expert credibility evaluation index system was established; Thirdly, in the selection of experts, TF-IDF algorithm was used to analyze the correlation between project and expert information, and the expert research field was classified according to the discipline classification standard to improve the matching degree between project and expert research field. Fourthly, the final candidate experts were selected based on various key indicators. Finally, on this basis, science and technology expert database information system of Chinese Academy of Sciences was designed and developed, which effectively improved the efficiency of expert database management and expert selection.

Keywords: science and technology expert database; python technology; TF-IDF algorithm; expert selected; expert database information system

（责任编辑：何岸波； 责任译审：毛子英 张述庆）