DOI:10.16644/j.cnki.cn33-1094/tp.2019.04.004

基于综合指数和知识图谱的水族文献核心作者群分析研究*

杨秀璋[1]，夏　换[2]，于小民[2]，项美玉[3]

(1. 贵州财经大学信息学院，贵州 贵阳 550025；2. 贵州财经大学贵州省经济系统仿真重点实验室；
3. 贵州财经大学贵阳大数据金融学院)

摘　要： 大数据时代，科研成果层出不穷，为了让科研工作者在海量文献中精准识别出文献的核心作者和科研群体，挖掘出作者间的合作关系，文章提出了一种基于综合指数和知识图谱的水族文献核心作者群识别方法。该方法采用Python抓取中国知网1953至2018年间990篇水族文献，结合发文量和被引用量构建综合指数遴选水族文献核心作者前20位，基于知识图谱和共现矩阵构建水族文献作者间的合作关系。据此梳理出我国水族文献的核心科研群体，明晰了水族研究的核心人物和团队现状，为水族文化研究提供了科学指引和参考依据，对传承与弘扬民族传统文化具有重要意义。

关键词： 水族文献；知识图谱；综合指数；核心作者群；普赖斯定律

中图分类号：TP391　　**文献标志码：**A　　**文章编号：**1006-8228(2019)04-13-05

Research on the core author group analysis of the Shui Nationality literature based on comprehensive index and knowledge map

Yang Xiuzhang[1], Xia Huan[2], Yu Xiaomin[2], Xiang Meiyu[3]

(*1. School of Information of Guizhou University of Finance and Economics, Guizhou, Guiyang 550025, China; 2. Guizhou Key Laboratory of Economics System Simulation of Guizhou University of Finance and Economics; 3. Guiyang Institute for Big Data and Finance of Guizhou University of Finance and Economics*)

Abstract: In the era of big data, scientific research results have emerged in an endless stream. To accurately identify the core authors and research groups in the vast literature, and to explore the cooperation between authors, this paper proposes a method for identifying the core authors of Shui literature based on comprehensive index and knowledge map. This method uses Python to capture 990 Shui documents from 1953 to 2018 in China, and combines the volume of publications and the cited quantity to construct a comprehensive index to select the top 20 core authors of Shui literature. Based on the knowledge map and co-occurrence matrix, the authors of Shui literature are constructed. On this basis, the core scientific research groups of China's Shui Nationality literature are sorted out, and the core figures and team status in studying Shui Nationality are clarified, which provides scientific guidance and reference basis for the study of Shui culture, and is of great significance for inheriting and carrying forward the national traditional culture.

Key words: Shui literature; knowledge map; comprehensive index; core author group; Price's law

0 引言

核心作者是学科研究的坚实基础[1]，决定着学术成果的质量。随着学术成果呈爆炸式增长，如何精准地识别出文献的核心作者和科研群体变得越来越困难。传统的核心作者识别方法是看发文量而忽视了论文的质量，缺乏利用知识图谱或社交网络技术构建核心作者间的关系，识别结果也往往比较片面[2]。

近年来，国内外学者致力于学术文献研究。姜春林通过文献计量历时法对《科学学研究》做出全面的计量分析[3]。梁永霞等基于CSSCI中国引文数据进行了分析和可视化研究[4]。黄晓斌等统计、分析我国情报学高被引论文，展示情报学的发展历程和学科主题[5]。蔡文伯等通过计量分析方法研究我国民族教育文献态势[6]。王宗水等基于1998-2014年中国社会科学引

收稿日期：2018-12-13

***基金项目：**贵州省教育厅青年科技人才成长项目(黔教合KY字[2016]172)；贵州省普通高等学校科技拔尖人才支持计划项目(黔教合KY字[2016]068)；贵州省教育厅青年科技人才成长项目(黔教合KY字[2016]178)；贵州省教育厅青年科技人才成长项目(黔教合KY字[2016]175)

作者简介：杨秀璋(1991-)，男，苗族，贵州凯里人，硕士，助教，主要研究方向：Web数据挖掘，知识图谱，数据分析。

通讯作者：夏换(1982-)，男，湖南永州人，博士，教授，主要研究方向：计算机仿真，大数据分析。

文数据分析社会网络范式的演化与发展[7]。徐庶睿等利用引文内容进行主题学科交叉类型分析[8]。同时，随着机器学习和人工智能技术迅速发展，知识图谱和社交网络技术也被运用来挖掘学科核心作者，分析学科发展脉络。罗双玲等提出了基于半积累引文网络社区发现的学科领域主题演化分析方法，并应用于“合作演化”领域[9]。马文博等通过文献计量方法和知识图谱分析《经济研究》近十年载文[10]。任晓松等归纳研究中国碳排放热点演化并构建知识图谱[11]。

水族是一个历史悠久和文化古朴的民族，具有重要的历史和文化价值[12]。1953年至2018年7月，中国知网共收录水族相关文献990篇，涉及水族文化、水族医学、水书文字、水族体育等主题。水族文献作为水族文化交流的重要载体，有效地推动水族文化的发展。当前水族领域的研究更多的是采用传统的查阅资料、现场考察及问卷调查的方法，核心作者识别仅考虑了发文量，没有采用综合指数和知识图谱来研究水族文献，缺乏对水族核心作者和科研团队深层次地挖掘。针对这些不足，本文依据普赖斯定律来确定水族文献核心作者候选人，提出了一种结合发文量和被引用量的综合指数方法遴选水族文献核心作者；基于知识图谱和共现矩阵构建水族核心科研群体及作者间合作关系。

1 研究方法

1.1 算法总体流程

本文旨在分析中国水族文献的核心作者及科研群体，具体流程如图1所示。

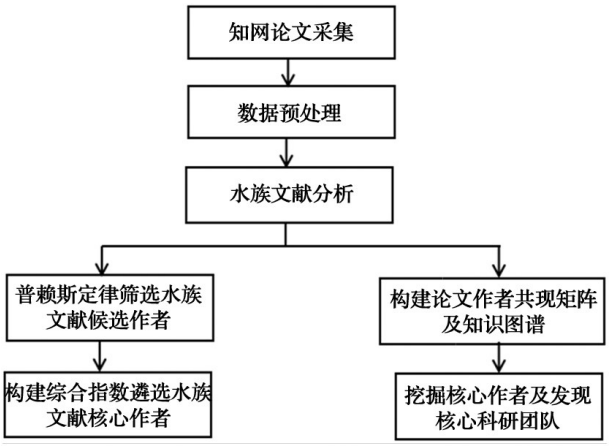


图1 水族文献核心作者群分析流程

通过Python和Selenium技术自定义爬虫抓取中国知网水族领域相关的文献。

接着对所抓取的语料进行数据预处理，包括数据清洗、数字提取、中文分词等处理。

利用普赖斯定律筛选文献作者候选人，通过综合指数法挖掘核心作者。

构建文献作者共现矩阵和知识图谱，发现水族文献核心科研团队。

1.2 数据采集及预处理

本文旨在挖掘中国知网水族文献的核心作者，分析水族科研群体及研究主题。抓取了中国知网1953年至2018年7月间990篇水族学术成果，详细信息如表1所示。对所抓取的文献进行数据预处理，这是为了得到质量更高和更完整的信息数据，从而为后续的实验提供有效支撑。本文的数据预处理操作包括中文分词、缺失值处理、停用词过滤。

表1 中国知网1953-2018年水族学术成果汇总表

学术成果类型	刊载成果数量	总下载数量	总引用数量	单篇最高下载数量	单篇最高引用数量	篇均下载数量	篇均引用数量
学术期刊	662	91478	1869	951	82	138.18	2.82
会议论文	106	4628	8	176	1	43.66	0.08
博士论文	7	7478	50	3664	18	1068.29	7.14
硕士论文	91	28324	192	1430	23	311.25	2.11
中国专利	5	0	0	0	0	0.00	0.00
科技成果	3	1	0	1	0	0.33	0.00
报　　纸	116	1589	5	65	2	13.70	0.04
总　　计	990	133498	2124	3664	82	134.85	2.15

2 基于综合指数的水族文献核心作者分析

结合文献的发文量和被引用量来综合确定核心作者候选人，再通过普赖斯定律计算核心作者候选人的最低发文量和最低被引用量，只要符合两者之一则

可以作为核心作者候选人进入测评样本[13]，再进一步计算水族文献的核心作者。步骤如下：

依据文献计量学中著名的普赖斯定律统计最低发文量。在本文的水族文献语料中，发文最多的作者是共计发文23篇。其计算公式如下：

$$Mp = 0.749\sqrt{Npmax} = 0.749 \times \sqrt{23} = 3.592 \quad (1)$$

Mp为普赖斯定律统计的最低发文量，Npmax为发表水族论文最多的数量。按照取整选择发表4篇或4篇以上的作者为水族文献核心作者候选人。

水族文献中作者发文被引用次数累计最高为176次，按照普赖斯定律确定核心作者候选人的发文累计最低被引用量，如公式(2)所示。

$$Mc = 0.749\sqrt{Ncmax} = 0.749 \times \sqrt{176} = 9.937 \quad (2)$$

Mc为普赖斯定律统计的最低被引用量，Ncmax为水族文献中作者发文被引用次数累计最高数量。发文被引用累计次数在10次或10次以上的作者可入选核心作者候选人。

筛选符合(1)或(2)的作者进行去重统计，最终确定水族文献核心作者候选人为151位，这些候选人共发表学术成果619篇，占全部水族文献的62.5%；候选人的总被引用次数为4089次，占水族文献总被引用的74.5%。

计算核心作者候选人的平均发文量和平均被引用量。平均发文量的计算过程如公式(3)所示，151位核心作者候选人共计发表学术成果619篇。

$$\bar{x} = \frac{619}{151} \approx 4.099 \quad (3)$$

平均被引用量的计算过程如公式(4)所示，151位核心作者候选人的学术成果共被引用4089次。

$$\bar{y} = \frac{4089}{151} \approx 27.079 \quad (4)$$

通过发文量和被引用量构建综合指数，从水族文献的数量和质量两个方面评估核心作者。综合指数的计算方法如公式(5)所示。其中，$score_i$表示第i个作者的综合指数分数，x_i和y_i表示第i个作者的发文量和累计被引用量，发文量和被引用量的权重均为0.5。

$$score_i = \frac{x_i}{\bar{x}} \times 0.5 + \frac{y_i}{\bar{y}} \times 0.5 \quad (5)$$

运用综合指数方法分别计算1953–2018年间水族文献151位核心作者候选人的综合分数，得出如表2所示的前20位核心作者。其中余跃生发表了水族相关的文章23篇，被引用量为174次，综合指数为6.018；顾晓艳发表了水族文献16篇，被引用量为176次，综合指数为5.201；王亚琼发表了水族领域的论文10篇，被引用量为132次，综合指数为3.657。

表2 中国知网1953–2018年间的水族文献核心作者

排名	作者	发文量	被引用量	综合指数	排名	作者	发文量	被引用量	综合指数
1	余跃生	23	174	6.018	11	谢渊	10	57	2.272
2	顾晓艳	16	176	5.201	12	石国义	6	80	2.209
3	王亚琼	10	132	3.657	13	赵凌	14	25	2.169
4	黄胜	10	106	3.177	14	刘世彬	11	41	2.099
5	张东秀	8	116	3.118	15	曹显明	5	72	1.939
6	何燕	12	59	2.553	16	陆玉炯	6	65	1.932
7	吴昌学	11	57	2.394	17	黄世宁	2	86	1.832
8	任锡麟	10	61	2.346	18	戎聚全	6	58	1.803
9	单可人	10	61	2.346	19	张振江	11	22	1.748
10	潘朝霖	13	39	2.306	20	蒙爱军	8	36	1.641

3 基于知识图谱的水族核心作者群分析

针对水族文献核心作者群分析，本文提出了一种基于知识图谱和共现矩阵的识别方法，构建中国知网水族文献作者间的关系，从而挖掘出对水族文化做出重要贡献的科研群体。其分析流程如下：

首先计算出1953–2018年收录于中国知网的990篇水族学术成果的所有作者名单。

构建水族学术成果作者间的共现矩阵。当两名作者合作完成一篇学术文章时，则认为共现并构建一条相关联的边，其边所对应的权重加1；否则当两名作者没有合作关系时，其权重为0。

采用Gephi构建水族作者间合作关系的知识图谱，并得出如图2所示的实验结果。图2中圆圈代表发文作者，圆圈越大发文量越多，反之越少；连线代表

作者间的合作关系，连线越粗合作次数越多，反之越少。该知识图谱共构建了497个核心作者和1095条关系，并将经常合作的科研群体聚集在一起，形成了以余跃生、顾晓艳、何燕、吴昌学、刘世彬、单可人、戎聚全、潘朝霖等学者为核心的学术研究团体。

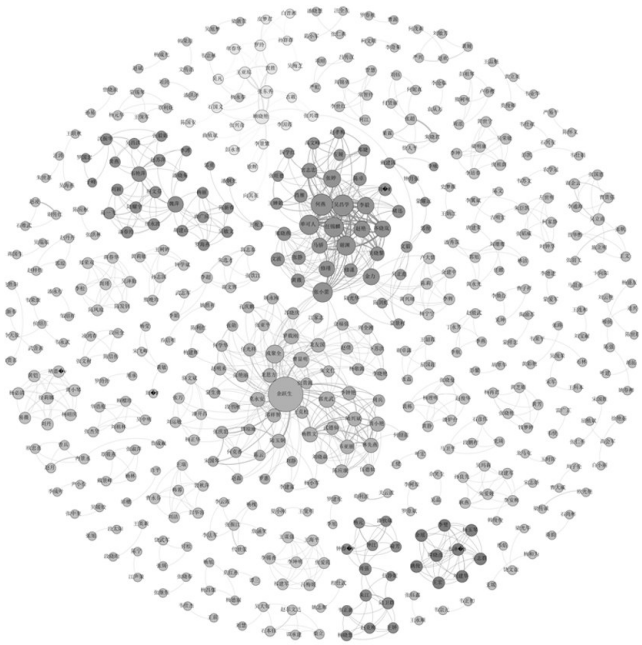
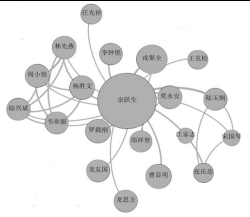
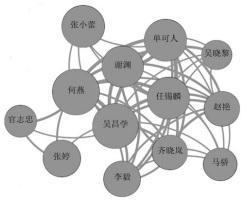
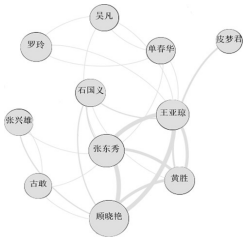
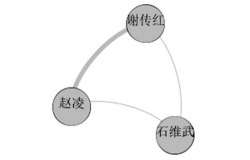
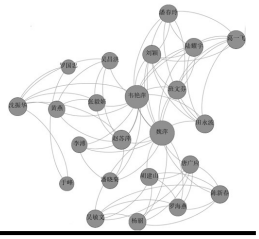


图2 水族文献作者合作关系的知识图谱

为了更好地挖掘出水族文献的核心科研团队，本文通过计算每个节点的度和每条边的权重，过滤掉合作较为单一的节点及关系，将水族领域的核心科研群体聚集在一起，得到如表3所示的五个水族文献核心科研团队，他们对水族领域的研究有着突出的贡献。其中以余跃生、戎聚全、杨胜文等为首的科研团队来自黔南民族医学高等专科学校，主要研究方向为水族医学和水族基因，代表著作有《贵州水族人群线粒体DNA序列多态分析》、《贵州南部6个民族5对遗传性状的基因频率》；以何燕、单可人、任锡麟等为首的科研团队来自贵阳医学院，主要研究水族医学及心血管疾病，代表著作有《贵州三都水族Y染色体单倍型频率分析》、《贵州三都水族β-地中海贫血筛查及基因分析》；以顾晓艳、张东秀、王亚琼等为首的团队研究方向为水族体育和水族传承，来自黔南民族师范学院，代表著作有《水族传统体育舞蹈的保护与传承》、《对水族山寨原生态传统体育文化的调查研究》；以赵凌、谢传红、石维武为首的科研团队主要研究水族音乐和水族乐器，来自黔南民族师范学院，代表著作有《贵州三都水族端节铜鼓音乐文化考察与分析》、《马联村水族端节铜鼓音乐文化初探》；以魏萍、韦艳萍、赵苏萍等为首的科研团队主要研究水族儿童体格发育，来自黔南州中医医院，代表著作有《贵州省黔南州农村布依、苗、水族儿童体格发育状况调查及其影响因素分析》、《黔南州农村水族和布依族7~12岁女性儿童骨骼发育差异性比较》。

表3 中国知网1953-2018年间的水族文献核心科研群体

水族核心科研群体	科研机构	研究方向	关系图谱
余跃生;戎聚全;杨胜文;罗载刚;莫永安;陆玉炯;张庆忠;陆兴斌;王克松;等	黔南民族医学高等专科学校	水族医学 水族基因 遗传学	
何 燕;单可人;任锡麟;谢 渊;张小蕾;吴昌学;张 婷;齐晓岚;吴晓黎等	贵阳医学院	水族医学 水族基因 心血管疾病	
顾晓艳;张东秀;王亚琼;黄 胜;张兴雄;石国义;罗 玲;单春华;皮梦君等	黔南民族师范学院	水族体育 水族传承 体育文化	
赵 凌;谢传红;石维武;	黔南民族师范学院	水族音乐 音乐舞蹈 水族乐器	
魏 萍;韦艳萍;赵苏萍;潘晓荣;胡建山;沈振华;唐广应;班文芬;吴敬文等	黔南州中医医院	水族儿童 体格发育 疾病分析	

4 结束语

本文采用基于综合指数和知识图谱的方法研究中国知网的水族文献，涉及1953-2018年共990篇水族领域的学术成果。实验结果表明，本文提出的基于普赖斯定律和综合指数的文献核心作者识别方法有效可行，从发文量和被引用量两方面评估核心作者，并挖掘出水族文献前20位核心作者，包括余跃生、顾晓艳、王亚琼等。本文基于知识图谱和共现矩阵的水族核心作者群识别方法，有效构建了水族作者间的合作图谱，挖掘出以余跃生、顾晓艳、何燕、吴昌学、

刘世彬、单可人、戎聚全、潘朝霖等学者为核心的水族科研团体，这些团队主要来自于黔南民族医学高等专科学校、贵阳医学院、黔南民族师范学院、黔南州中医医院等机构。

本文提出的方法精准地识别出水族研究的核心作者及科研团队，展示了研究我国水族文化、水族医学、水族体育、水族文字领域的专家人群及研究方向，有效地把握水族学科脉络，减轻了人力筛选和分析的负担，提高了研究效率和准确度，为大数据时代提高论文索引效率、分析研究群体、识别核心作者提供有效支持。同时，本文为下一步的水族文献挖掘、追踪水族源流、研究水族群体变迁、保护和传承水族文化提供有效支撑，对传承与弘扬民族传统文化具有重要意义，该研究成果具有一定的应用前景和实用价值。

参考文献(References):

[1] 廉清.《图书情报工作》核心作者群分析研究[J].现代情报,2004.11:55-59

[2] 钟文娟.基于普赖斯定律与综合指数法的核心作者测评——以《图书馆建设》为例[J].科技管理研究,2012.2:57-60

[3] 姜春林.基于文献计量学历时法引文的案例分析[J].现代情报,2005.10:140-145

[4] 梁永霞,杨中楷,刘则渊.基于CSSCI的中国引文分析的可视化研究[J].情报研究,2008:34-38

[5] 黄晓斌,张欢庆.我国情报学高被引论文分析[J].情报科学,2018.36(1):54-60

[6] 蔡文伯,马杰.我国民族教育研究文献态势的计量分析[J].民族教育研究,2014.25(2):138-144

[7] 王宗水,赵红,刘宇,秦续忠.社会网络研究范式的演化、发展与应用——基于1998~2014年中国社会科学引文数据分析[J].情报学报,2015.34(12):1235-1245

[8] 徐彦睿,章成志,卢超.利用引文内容进行主题级学科交叉类型分析[J].图书情报工作,2017.61(23):15-24

[9] 罗双玲,张文琪,夏昊翔.基于学积累引文网络社区发现的学科领域主题演化分析——以"合作演化"领域为例[J].情报学报,2017.36(1):100-110

[10] 马文博,陈占明.《经济研究》近十年载文的文献计量与知识图谱分析[J].现代情报,2018.38(2):148-156

[11] 任晓松,孙天美,赵国浩.中国碳排放研究热点演化知识图谱分析[J].科技管理研究,2018.10:235-243

[12] 饶文谊,梁光华.关于水族水字水书起源时代的学术思考[J].原生态民族文化学刊,2009.4:90-93

[13] 丁学东.文献计量学基础[M].北京大学出版社,1992:204-209,220-232

(上接第 12 页)

行挖掘研究，采用了将朴素贝叶斯与Hadoop相结合的处理数据的新方法，在预测日最高气温中具有较高的预测率和正确率。该方法具有以下特点：①可以充分利用海量数据，有效地避免了信息的丢失；②在大量样本下，用较为简单的算法达到了不为逊色的结果；③能够处理不完全、不精确的训练数据集；④对连续数据的离散化采用较为简单的PKI算法，对气象数据某些分布不是很均匀的属性来说，离散效果还有待提高。在数据量海量增加的今天，此方法提供了在海量数据中挖掘有用的信息的新思路，可在移动互联网、电子商务等诸多领域的应用中进一步去研究。

参考文献(References):

[1] 乔梁.数据挖掘在气象服务中的应用研究[J].信息通信,2016.2:96-97

[2] 张硕,张永宁.大数据时代气象数据新闻的探索与实践—以中国天气网为例[J].NEW MEDIA RESEARCH,2017.22:120-122

[3] 张晨阳,刘利民,马志强.云计算下基于贝叶斯分类的气象数据挖掘研究[D].内蒙古工业大学,2014.

[4] 苑立民,郝成亮,刘昶,徐峰,潘建宏,张凯.基于Hadoop生态环境的大数据平台在电网公司海量数据准实时处理中的应用[J].大众用电,2017增刊.1:38-41

[5] 李斌,张建平,刘学军.基于Hadoop的不确定异常时间序列检测[J].传感技术学报,2015.28(7):1066-1072

[6] 陈坚钊.MapReduce的工作机理及其应用研究[D].华侨大学,2013.

[7] 赵力.基于贝叶斯压缩感知的说话人识别方法[J].电子器件,2015.38(5):1135-1137

[8] 谢作将.面向朴素贝叶斯算法的离散化方法研究[D].北京交通大学,2008.

[9] Yang Y, Webb G I. Proportional k-Interval Discretization for Bavie-Bayes Classifiers[J]. Proc. of the Twelfth European Conf.on Machine Learning,2001.2167:564

[10] 刘君.基于Hadoop技术的气象数据采集及数据挖掘平台的研究[D].天津理工大学,2015.

[11] 闫永刚.基于Hadoop的KNN分类的气象数据预测研究[D].南京信息工程大学,2012.