研究与开发

文章编号: 1007-1423(2022)01-0001-09 DOI: 10.3969/j.issn.1007-1423.2022.01.001

基于知识图谱和层次聚类的水族文化主题演化研究

令狐秋萍，何世群，齐梦珂，罗子江，杨秀璋

（贵州财经大学信息学院，贵阳 550025）

摘要：水族文化的主题挖掘和主题演化分析有助于从不同角度了解水族发展状况、热点主题和研究趋势，为后续水族文化传承和文献挖掘提供相关参考价值。基于此，本文采集中国知网990篇水族文献，提取水族文化关键词及主题，利用共现分析和层次聚类挖掘水族关键词间的联系，构建水族主题共现知识图谱和主题演化网络。实验结果表明，本文的方法能有效分析水族文献的主题演化趋势，发现特征词共现关系和相似度，构建水族文化主题知识图谱，并聚类形成五大类水族文献主题，对少数民族文化研究和文献挖掘具有一定的应用价值和理论意义。

关键词：水族文化；主题演化；层次聚类；知识图谱；文本挖掘

基金项目：贵州省科学技术基金项目：基于大数据及图像识别的水族文献及濒危水书抢救性整理研究(黔科合基础[2020]1Y279)；贵州省科学技术基金项目：多源地理数据融合知识图谱构建方法在舆情分析中的应用——以贵州省为例(黔科合基础[2019]1041)；贵州省教育厅青年科技人才成长项目：基于大数据和知识图谱的公共卫生事件智能预警与分析研究(黔教合KY字[2021]135)；贵州财经大学校级科研基金：AI大数据赋能的贵州濒危水书和水族古籍识别与抢救研究(2021KYQN03)

0 引言

主题是文献的主旨和核心内容，通过研究某学科领域文献主题的发展变化，并沿着时间轴挖掘其经历新生、成长、分裂、融合、衰退和消亡等过程，能够有效揭示出该学科领域内不同主题的发展规律与趋势。主题演化研究有助于科研工作者深入了解学科知识，把握学科发展方向和研究趋势，及时调整学科战略布局，从而优化学科发展。

近年来，我国各领域学者致力于主题演化研究，研究方法主要包括共引分析法、词频分析法、共词分析法、LDA模型等。范少萍等[1]提出利用密度和热度开展核心主题识别的研究，从关键关联与核心主题两方面共同识别医学文献的主题演化路径。逯万辉与谭宗颖[2]构建基于知识基因游离与重组的主题演化研究模型，通过对领域知识基因进行识别和聚类建立主题演化网络。刘艳华与钱爱兵[3]通过共词分析和聚类分析，揭示国际图书情报领域五个研究方向上的七条知识演化路径。陈伟等[4]结合LDA主题模型与隐马尔可夫模型，分析专利技术主题的分布特征和演变规律并对未来技术趋势进行定量预测。吴江等[5]通过文献计量、聚类分析、纵向映射分析等方法，对在线医疗健康的国家合作、研究热点、主题演化和研究方法进行深入分析。安璐等[6]利用LDA模型比较分析MERS病毒爆发时新浪微博和微信平台上各利益相关者在不同阶段的话题关注点，并揭示其演化模式异同点。李杰与陈超美[7]基于共被引的方法分析学科主题变换趋势，设计开发了CiteSpace软件，通过可视化分析展示主题演化过程。

水族主要聚居在黔桂滇交界的龙江、都柳江上游地带，具有悠久的历史和古朴的文化。在中国少数民族文化遗产中，水族是引人注目的，它拥有独特的古文字“水书”，长期为世界各国学者所关注[8]。水族文献作为水族文化交

流的重要载体，了解其发展历程、挖掘其热点关键词并识别其演化趋势是重要的研究内容，将为后续水族文化传承和文献挖掘提供相关的参考价值，对于揭示水族经济及人文发展具有重大意义。当前水族领域的研究多采用传统的查阅资料、现场考察、问卷调查及文献归纳方法，关于水族文献的计量分析相对较少，这些分析手段不能有效地挖掘出水族文献的研究热点及核心主题，难以全面系统地把握水族主题演化趋势、学科现状，也没有形成较为完整的理论体系。基于此，本文采用共词分析和层次聚类方法，挖掘水族文献关键词间的联系，构建水族文献关键词共现知识图谱和主题演化网络，可视化地展现水族的发展趋势、研究热点等，其实验和结论对民族研究和文献挖掘具有一定应用价值和理论意义。

1 研究方法

1.1 算法总体流程

本文提出基于共词分析和层次聚类的水族文献主题演化算法，其总体框架如图1所示。

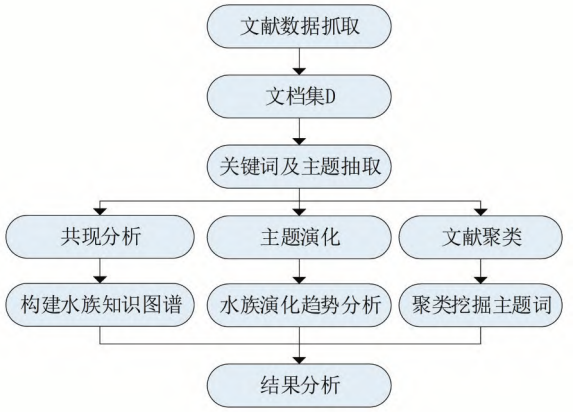

图 1 水族文献主题演化框架

（1）通过Python和Selenium技术自定义网络爬虫抓取中国知网水族文献，共爬取“水族+民族”关键词的文献990篇。

（2）对所爬取的中文语料进行预处理，包括中文分词、词性标注、数值提取、停用词过滤等，从而获取能够反映水族文献主题的关键词。

（3）抽取水族文献标题、摘要中的关键词，采用TF-IDF公式计算每个词的权重，并过滤掉TF-IDF值小于阈值的词，形成关键词集合。

（4）通过共现分析、文献聚类、主题演化分别构建水族关键词知识图谱，挖掘核心主题及分析水族文献演化趋势，并得出相关实验结果。

1.2 数据来源

本文共抓取了中国知网CNKI数据库1953年至2018年7月间水族相关文献990篇，检索关键词为“水族+民族”。其中，学术期刊662篇、会议论文106篇、博士论文7篇、硕士论文91篇、中国专利5篇、科技成果3篇、报纸116篇，详细信息如表1所示。

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	单篇最高引用	篇均下载数量	篇均引用数量
学术期刊	662	91478	1869	82	138.18	2.82
会议论文	106	4628	8	1	43.66	0.08
博士论文	7	7478	50	18	1068.29	7.14
硕士论文	91	28324	192	23	311.25	2.11
中国专利	5	0	0	0	0.00	0.00
科技成果	3	1	0	0	0.33	0.00
报纸	116	1589	5	2	13.70	0.04
总计	990	133498	2124	82	134.85	2.15

采用词频统计、共词分析、层次聚类、知识图谱和主题演化网络方法研究水族文献。包括：统计水族文献关键词出现频次，建立共现矩阵，利用Gephi软件生成对应的关键词共现知识图谱；将水族文献关键词的共现矩阵转换为相异矩阵，利用Ward方法计算簇间距离，再调用Python层次聚类算法挖掘水族文献主题；基于时间序列分析水族文献主题演化趋势。本研究将划分为2000年前、2001—2005年、2006—2010年、2011—2015年、2016—2018年5个子时期进行主题演化分析，再绘制相关的可视化图形，展现各个时期的核心主题。

2 水族文献热点主题挖掘

关键词作为论文的重要部分，可以反映论文的研究主题或热点话题等内容。水族文献关键词的共词分析和热点主题挖掘，可以把握该领域的研究方向和热点主题。

2.1　高频关键词分析

据1953—2018年中国知网来源的水族文献研究涉及的关键词及其频次统计显示，990篇文献共涉及关键词1316个，关键词总频数为2407次，其平均词频数值约为1.829。该领域1040个关键词仅出现1次，占总关键词的79.0%；158个关键词出现2次，占总关键词的12.0%；出现频数在5次及以上的高频关键词共46个，共出现810次，占关键词出现总频数的33.7%。从表2可以看出，“水族”词频居首，共计312次；“水书”“少数民族”“贵州”“马尾绣”“苗族”“文化”“研究”“非物质文化遗产”“传承”等均为高频关键词，在一定程度上反映出水族研究领域对文化、水书、民族和传承关注较多。

表2　水族文献高频关键词统计表

关键词	频次	关键词	频次	关键词	频次
水族	312	变迁	16	民歌	9
水书	30	端节	16	民族	9
少数民族	28	布依族	15	水语	9
贵州	27	基因频率	13	传统文化	8
马尾绣	25	三都	12	价值	7
苗族	22	水族马尾绣	12	三都水族	7
文化	22	传统体育	11	文化内涵	7
研究	20	民族文化	11	现状	7
非物质文化遗产	18	水族文化	10	铜鼓	6
传承	17	保护	9	发展	6

2.2　关键词共现分析

针对高频关键词无法反映词语和主题之间的内在关联，不能全面揭示出水族文献的研究热点及关键词动态。本文采用共词分析方法构建水族文献的关键词共现矩阵，如公式(1)所示，当两个关键词共同出现在一篇学术文章中，则认为共现并构建一条相关联的边，其边对应的权重加1；反之，两个关键词不存在共现关系，其权重为0。

$$y=\begin{cases}+1, & a\text{和}b\text{关键词同时出现在论文中}\\ 0, & a\text{和}b\text{关键词没有在论文中共现}\end{cases} \tag{1}$$

接着采用Ochiia系数法计算共现矩阵的相似度，计算公式如式(2)所示，O_{ij}为所求的共现系数，C_{ij}是关键词i和关键词j共现总次数，C_i是关键词i出现的总次数，C_j是关键词j出现的总次数。

$$O_{ij}=\frac{C_{ij}}{\sqrt{C_i}\times\sqrt{C_j}} \tag{2}$$

在共现分析中，两个关键词共同出现的次数越多，说明关键词联系越紧密，其相关系数越接近1，越能体现主题的研究内容相关联；如果相关系数值为0，说明两个关键词之间没有关系。本文根据水族文献关键词共现分析，得出了如表3所示的水族文献共现高频词及相关系数。

表3　水族文献共现高频词及Ochiia系数统计表

序号	关键词1	关键词2	共现频次	Ochiia系数
1	水族	文化	18	0.2172
2	水族	端节	15	0.2123
3	水族	苗族	15	0.1811
4	水族	马尾绣	13	0.1472
5	水族	布依族	12	0.1754
6	水族	传统体育	11	0.1878
7	水族	变迁	11	0.1557
8	布依族	苗族	11	0.6055
9	水族	传承	10	0.1373
10	水族	研究	8	0.1013
11	水族	贵州	7	0.0763
12	水族	非物质文化遗产	7	0.0934
13	贵州	基因频率	7	0.3736
14	水族	赛马	6	0.1387
15	水族	现状	6	0.1284
16	水族	传统文化	6	0.1201
17	水族	基因频率	6	0.0942
18	水族	民歌	5	0.1132
19	水族	吞口舞	5	0.1266
20	贵州	遗传性状	5	0.3928

其中排名前5位的分别是：“水族”和“文化”共现18次，相关系数为0.2172；“水族”和“端节”共现15次，相关系数为0.2123；“水族”和“苗族”共现15次，相关系数为0.1811；“水族”和“马尾绣”共现13次，相关系数为0.1472；“水族”和“布依族”共现12次，相关系数为0.1754。

2.3　主题共现知识图谱

采用Gephi软件构建水族文献关键词共现知识图谱，结果如图2所示，共构建1316个核心

关键词和3444条共现关系。图中圆圈表示关键词，连线表示共现关系，连线越粗其共现次数越多，反之越少。

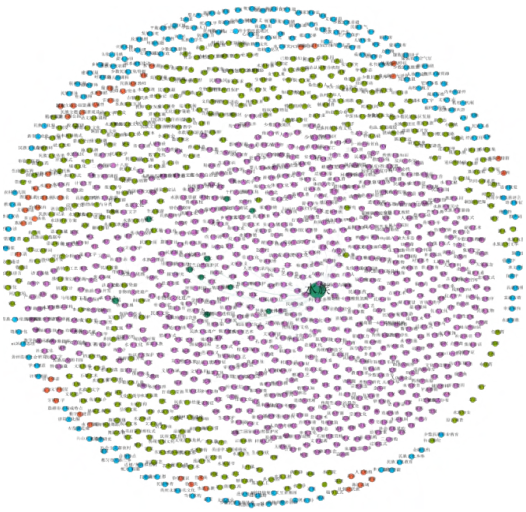

图2 水族文献关键词共现知识图谱

为了更精准地识别水族文献的关键词及主题知识图谱，本文过滤了较为单一的共现关系及关键词，得到如图3所示的水族核心关键词共现图谱。

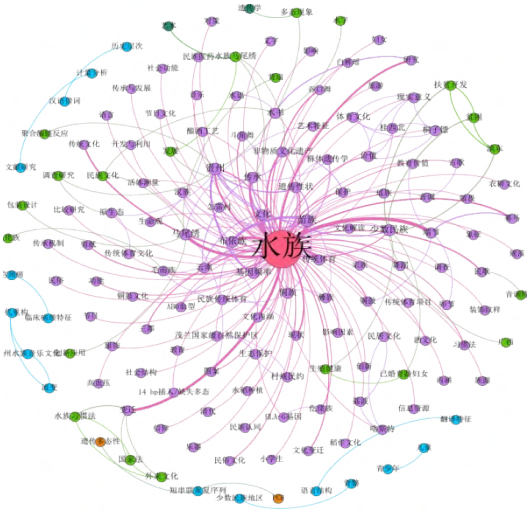

图3 水族文献核心关键词共现知识图谱

该图谱共包括153个节点关键词，居于中心位置的是“水族”，其他的主题词逐渐向边缘分布扩散。其中“水族”和“变迁”“端节”“水书”“马尾绣”“研究”“少数民族”“传统文化”“贵州”“传承”等关键词共现明显，其连线较粗。

3 水族文献主题演化分析

本研究尽可能地保证不同阶段文献数据的均衡性并避免数据的平滑性，将水族文献划分为5个时期：2000年之前、2001—2005年、2006—2010年、2011—2015年、2016—2018年。其中，2000年之前的水族文献总数为139篇，总下载量12568次，总被引用量376次，单篇最高下载量630次，单篇最高被引用量82次，其为李培春等[9]1994年在《人类学学报》发表的《水族的体质特征研究》；2001—2005年间的水族文献总数为100篇，总下载量14366次，总被引用量392次，单篇最高下载量1430次，单篇最高被引用量48次，其为苏和平[10]2004年在《贵州民族研究》发表的《水族审美意识探源》；2006–2010年间的水族文献总数为235篇，总下载量47043次，总被引用量828次，单篇最高下载量2040次，单篇最高被引用量38次，其为顾晓艳[11]2006年在《中国体育科技》发表的《传统体育文化在水族山寨中的生存状态——水族“端节”赛马活动的变迁》；2011—2015年间的水族文献总数为342篇，总下载量50896次，总被引用量490次，单篇最高下载量3664次，单篇最高被引用量38次，其为孙志国等[12]在《贵州民族学院学报(哲学社会科学版)》发表的《水族非物质文化遗产保护的探讨》；2016—2018年间的水族文献总数为174篇，总下载量8625次，总被引用量38次，单篇最高下载量541次，单篇最高被引用量5次，其为杨孝斌等[13]2016年在《数学通报》发表的《人类学视域下的水族数学文化研究》。

通过计算5个阶段各个主题的热度及出现频次，从而确定每个时间窗内的热点主题与潜在主题，构建各阶段主题演化的趋势及发展态势，得出如图4所示的水族文献主题演化趋势。

在图4中，时间轴上的蓝色方框表示各个主题，蓝色虚线方框为仅出现一次并消亡的主题，各种颜色的连线表示各主题的演化关系，每个阶段的12个主题，其热门程度从上往下依次排列。由图可知：

(1) 总体情况。第一阶段(2000年之前)热点主题为“水族”“民族文化”“少数民族”“遗

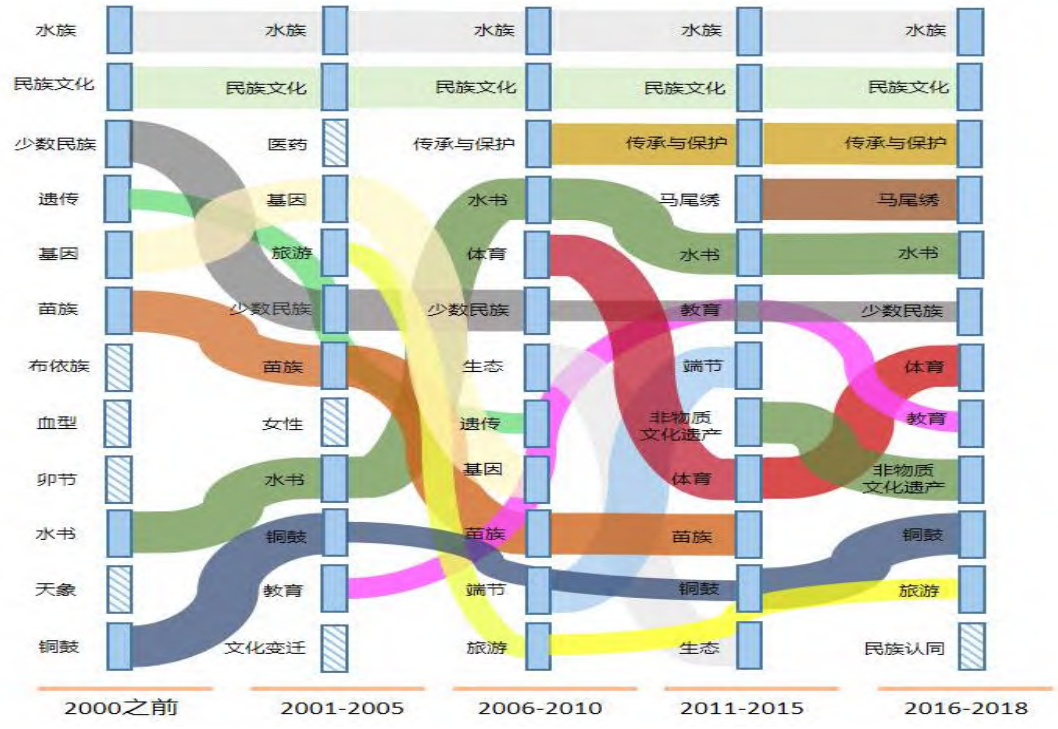

图 4　水族文献主题演化趋势

传”“基因”等；第二阶段(2001—2005年)热点主题为“水族”“民族文化”“医药”“基因”“旅游”等；第三阶段(2006—2010年)热点主题为“水族”“民族文化”“传承与保护”“水书”“体育”“少数民族”等；第四阶段(2011—2015年)热点主题为“水族”“民族文化”“传承与保护”“马尾绣”“水书”“教育”等；第五阶段(2016—2018年)热点主题为“水族”“民族文化”“传承与保护”“马尾绣”“水书”“少数民族”等。

(2) 热门主题及孤立主题。“水族”和“民族文化”为最热门的两个主题，“布依族”“血型”“卵节”“天象”“女性”“民族认同”“文化变迁”为仅出现一次就消亡的孤立主题。

(3) 各阶段的新生主题。第一阶段均为新生主题；第二阶段新生的主题包括“医药”“旅游”“女性”“教育”“文化变迁”；第三阶段新生的主题包括“传承与保护”“体育”“生态”“端节”；第四阶段新生的主题包括“马尾绣”“非物质文化遗产”；第五阶段新生的主题包括“民族认同”。

(4) 各阶段的消亡主题。除去孤立主题，第二阶段消亡的主题为“遗传”；第三阶段消亡的主题为“教育”和“铜鼓”；第四阶段消亡的主题为“少数民族”和“旅游”；第五阶段消亡的主题为“苗族”“端节”“生态”。

(5) 部分主题演化过程。“水书”主题从第一阶段新生，接着迅猛式增长，在第三阶段是第四热门的主题，后面两个阶段稳定第五，其增长率最快；“遗传”和“基因”主题第一阶段较为热门，后续逐渐回落并消失在演化趋势图中，其衰退率最快；“少数民族”主题第一阶段是第三热门的主题，接着逐渐降低并稳定在第六个位置；“苗族”主题第一阶段位于中间位置，之后逐渐下滑并消亡；“铜鼓”主题第一阶段位于最后，接着在各个阶段末尾徘徊。同时，随着民族文化保护意识的增强，中间出现了“生态”“马尾绣”“非物质文化遗产”等主题；随着边远地区少数民族教育的增强，“教育”“体育”等主题新生并发展；2016—2018年出现了“民族认同”主题，今后也会逐渐热门。

4 水族文献主题聚类分析

层次聚类分析是利用相似性算法发现高频关键词间亲疏程度并进行自动分类的技术。本文首先将水族文献关键词的共现矩阵转换为相异矩阵，接着使用Python层次聚类分析，簇间距离采用Ward方法计算，结果如图5所示。横向坐标轴表示各类别间的距离，纵向坐标轴表示各高频关键词。由图可知，我国水族文献分为5类：第一类为水族文化，包括“传统文化”“马尾绣”“非物质文化遗产”“文化变迁”“三都水族”“端节”“铜鼓”“艺术特征”等关键词；第二类为水族体育，包括“民族传统体育”“水族”“吞口舞”“传统体育”“传承”等关键词；第三类为水族遗传，包括“遗传形状”“基因频率”“布依族”“苗族”“侗族”“贵州”等关键词；第四类为水书水语，包括“贵州水族”“水族文化”“水语”“水书”“水书文化”“文化内涵”“民族文化”等关键词；第五类为民族区域发展，包括“水族地区”“民族地区”“民族”“发展”等关键词。

4.1 水族文化

水族文化具有悠久的历史、深刻的内涵及鲜明的特点，水族人民在发展过程中创造了丰富的民族文化，其物质文化、精神文化、观念文化是文化资源的基本表现形式。

（1）水族物质文化。包括“干栏”式建筑、水族民间工艺、水族饮食文化等。“干栏”也可称为麻栏或阁栏，其“干”和“栏”在水族语言中分别代表楼和家，干栏式房屋空间构造独特且功能分明，具有浓厚的水族特色；水族民间工艺最具代表性的是“马尾绣”[14]，成品中白色马尾是经过数道精密工序绘制，再配有龙凤、花鸟鱼虫等图案，代表吉祥如意等美好寓意；水族人民在饮食上偏爱鱼和九阡酒，他们将鱼用于进行民俗活动及重要日子的各个环节，并视为圣物。同时水族人民热爱的九阡酒以纯糯米为原料，配以巴岩香、地瓜香等100余种草药酝酿而成，具有舒筋活血的功能。

（2）水族精神文化。主要包括水族音乐舞蹈、水族神话传说等，水族人民能歌善舞，他们认为铜鼓具有幸福欢乐的象征[15]，水族著名的音乐舞蹈有《双歌调》《铜鼓舞曲调》等，水族著名的神话传说有《牙巫造人》《人类起源》等，它们既是符字，又代表着水族人民的精神世界。

（3）观念文化是民族意识、民族自尊心及民族信心的核心内容，社会上流传的《水书》古

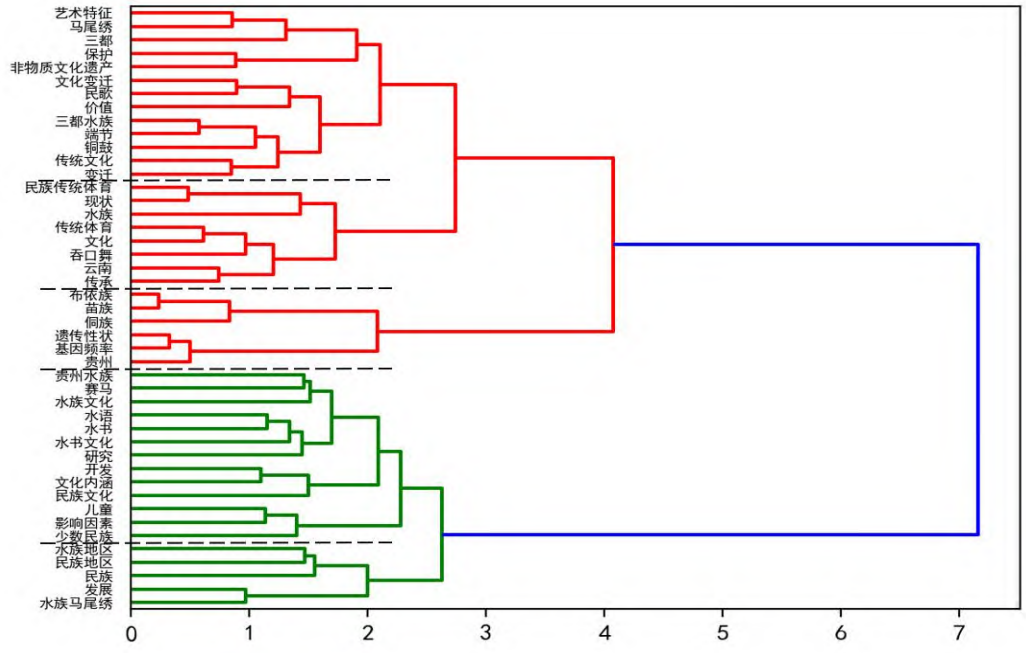

图5 水族文献主题聚类分析

文字与殷商时期的甲骨文极为相似，有着悠久的历史，在漫长的历史中，他们祖先就学习了周边民族的先进技术，后人将其传承改进，因此，水族一直被誉为勤劳勇敢、开放的民族。

4.2　水族体育

水族体育是水族人民在其漫长的历史进程中不断创造、发展和积淀起来的一种重要的文化生活方式[16]。它源于当地人们的农耕活动，具有典型的民族区域特征，满足了水族人民日常生活中的物质文化需求。水族体育内容丰富，种类繁多，大体上可分成5类，分别是武术类、竞技类、歌舞类、表演游戏类和棋类，这些体育项目主要体现在民俗及节日活动中，并随着人们的参与世代传承。例如“端节”赛马活动，水族人民乐在其中，不仅增强全族人际间的交往关系，还加强了民族凝聚力，使水族传统文化得到了延续。近年来由于水族社会经济和现代体育的发展，水族体育受到了较大冲击，一些水族传统体育项目逐渐被人们所遗忘，但随着国家对水族非物质文化遗产保护力度的加大，水族体育在与现代体育的碰撞中不断丰富和发展，既保留了自身民族特色，又融入了新的元素，展现出更强的生命力。

4.3　水族遗传

水族基因、遗传以及人群生长变化一直都是民族研究的热点问题，针对这一问题有学者从多种的角度在不同领域展开了许多工作。如张庆忠等人[17]站在民族差异的角度，运用群体遗传学人体测量方法对水族与苗族的遗传性状和血型进行了研究。余跃生等人[18]采用Kimura双参数模型估算单倍型遗传距离和邻结法构建单倍型无根系统树，实现对水族线粒体DNA序列的变异检测。朱江等人[19]从人群生长发育的角度，利用问卷法对水族婴儿和产妇进行为期1年的实地调查和体格检查，并对水族地区的人群发展影响因素进行深入研究与总结。水族遗传的研究对于水族的传承和民族多样性的发展至关重要，对水族人民的繁衍、生长、疾病预防和医疗发展具有一定的现实意义。

4.4　水书水语

水书是水族文字的通称，包括水族古文字和其相关典籍的汉译，水语称其为“勒睢”，用来记录水族民间各类知识和文化等重要信息，涉及生产生活、天文历法、民俗宗教等诸多方面的内容，目前主要在贵州水族聚集区被广泛流传及使用[20]。2006年，水书被列为首批国家级非物质文化遗产名录，是中华民族文化的重要组成部分，蕴藏着丰富的文化和历史价值，对相关学术研究具有借鉴意义[21]。当前有学者从不同角度对水书水语展开研究，其中王观玉等[22]从水书作为一种文献信息资源的视角对其进行详细分析，并从水书文献的收集、整理、保护及开发过程中出现的问题提出改进措施和方法，使水书资源得到了有效开发和利用。张漾艺[23]对三都水族自治县的水书习俗进行深入研究，发现水书习俗正在面临消亡的窘境和其影响因素，并提出采用新媒体传播等方式来促进水书文化的发展和传承。综上所述，文字是一个民族文化传承的重要工具和因素，水族文字对于水族文化的传承和发扬起到了举足轻重的作用。

4.5　民族区域发展

水族主要分布于黔桂交界的龙江、都柳江上游地带，自然条件差，经济基础薄弱，市场不够完善，科学教育不足，开发难度大，这些问题一直都是民族地区发展的绊脚石[24-25]。由于脱贫工作的不断深入以及党和国家对民族地区发展的高度重视，政府加大对水族地区的支持，积极发展教育，重视对人才的培养。以三都水族自治县为例，学前适龄儿童入园(班)率75.2%；适龄儿童入学率为99.6%；适龄少年入学率为98.4%；九年义务教育巩固率为91.01%；高中阶段毛入学率为81.06%，极大地改善了水族人民的教育水平，为水族地区的发展提供了人才支撑。同时，党和国家还积极推行产业化发展，大力扶持水族地区的文化旅游等产业，增加就业岗位，提高人均收入，加速贫困群众脱贫致富。此外，国家也培育了新的经济增长点，为水族地区营造良好的投资环境，促进经济结构的合理调整，进而实现经济的可持续发展[26]。作为脱贫攻坚的主战场，在新时代的大环境下，党和国家为水族人民的发展注入了新的动力，使得水族地区的交通、教育、经济以

及生活水平都得到了巨大的改善，也让更多的人了解到水族的特色与变化。

5 结语

本文提出了一种基于共词分析和层次聚类的水族文献主题演化算法，通过构建水族主题共现知识图谱和主题演化网络分析水族文献，弥补传统方法不能有效地挖掘出水族文献的研究热点及核心主题，难以全面系统地把握水族主题演化趋势、学科现状的不足。现得出结论如下：

（1）通过构建高频关键词共现矩阵和知识图谱，发现“水族”“变迁”“贵州”“传统文化”“少数民族”等关键词共现明显且频次较高，即为当前水族研究的热点主题。同时，利用ochiia计算矩阵相似度，有效揭示了各主题间联系及其关联程度。

（2）主题演化分析显示，国内水族文献主题可划分成5个阶段，包括第一阶段（2000年之前）、第二阶段（2001—2005年）、第三阶段（2006—2010年）、第四阶段（2011—2015年）、第五阶段（2016—2018年），借助共词分析、相似度计算和社交网络分析方法及相关软件对各阶段的特点详尽分析，涉及主题新生、消亡、演化、孤立等类型，有效梳理了水族领域各主题及其之间的发展脉络。

（3）层次聚类分析显示，我国水族研究主要包括水族文化、水族体育、水族遗传、水书水语和民族区域发展五大类主题，即水族领域的主要研究方向，各类主题间既相对独立但又互相关联，共同构成了独特的水族精粹，是中华民族文化的重要组成部分。

综上所述，本文的方法能有效分析水族文献的主题演化趋势和主题聚类详情，发现关键词共现关系和相似度，无论在实际应用还是学术理论研究，都具有一定意义，并对民族研究和文献挖掘提供相关的应用价值和理论意义。

参考文献：

[1] 范少萍，安新颖. 基于医学文献的主题演化类型与演化路径识别方法研究[J]. 情报理论与实践，2019，42(03)：114-119.

[2] 逯万辉，谭宗颖. 基于知识基因游离与重组的领域主题演化研究[J]. 情报理论与实践，2019，42(02)：101-107.

[3] 刘艳华，钱爱兵. 21世纪以来国际图书情报领域研究热点的演化路径探析[J]. 西南民族大学学报(人文社科版)，2018，39(05)：229-235.

[4] 陈伟，林超然，李金秋，等. 基于LDA-HMM的专利技术主题演化趋势分析：以船用柴油机技术为例[J]. 情报学报，2018，37(07)：732-741.

[5] 吴江，刘冠君，胡仙. 在线医疗健康研究的系统综述：研究热点、主题演化和研究方法[J]. 数据分析与知识发现，2019，3(04)：2-12.

[6] 安璐，杜廷尧，李纲，等. 突发公共卫生事件利益相关者在社交媒体中的关注点及演化模式[J]. 情报学报，2018，37(04)：394-405.

[7] 李杰，陈超美. CiteSpace：科技文本挖掘及可视化[M]. 2016版. 北京：首都经济贸易大学出版社，2016：110-135.

[8] 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，2019，01(04)：90-94.

[9] 李培春，梁明康，吴荣敏，等. 水族的体质特征研究[J]. 人类学学报，1994(01)：56-63.

[10] 苏和平. 水族审美意识探源[J]. 贵州民族研究，2004(03)：70-73.

[11] 顾晓艳，石国义，等. 传统体育文化在水族山寨中的生存状态：水族“端节”赛马活动的变迁[J]. 中国体育科技，2006(05)：38-40.

[12] 孙志国，黄莉敏，等. 水族非物质文化遗产保护的探讨[J]. 贵州民族学院学报(哲学社会科学版)，2011(06)：10-13.

[13] 杨孝斌，罗永，张和平. 人类学视域下的水族数学文化研究[J]. 数学通报，2016，55(08)：9-16.

[14] 邱晓敏. 水族文化的开发与保护初探：以贵州三都为例[J]. 群文天地，2012(05)：295-296.

[15] 钟华. 从“饭稻羹鱼”探析水族文化之源[J]. 农业考古，2014(03)：320-322.

[16] 顾晓艳，徐辉. 论水族传统体育的文化特征[J]. 体育学刊，2006(06)：60-62.

[17] 张庆忠，陆玉炯，等. 贵州苗族、水族7对性状的调查[J]. 现代预防医学，2009，36(06)：1130-1133.

[18] 余跃生，姚永刚，等. 贵州水族人群线粒体DNA序列多态分析[J]. 遗传学报，2001(08)：692-698.

[19] 朱江，陆卫群，等. 贵州省水族婴儿生长发育水平和影响因素调查[J]. 中国妇幼保健，2015，30(12)：

1899-1902.
[20] 杨秀璋,武帅,夏换,等.基于自适应图像增强技术的水族文字提取与识别研究[J].计算机科学,2020,48(1 Suppl.):74-79.
[21] 杨秀璋,夏换,于小民.一种基于水族濒危文字的图像增强及识别方法[J].计算机科学,2019,46(2 Suppl.):324-328.
[22] 王观玉,张娅妮.水书文献信息资源管理与开发利用探讨[J].图书情报工作,2009,53(9):74-77.
[23] 张潇艺.非物质文化遗产的新媒体传播策略研究:以三都水族自治县水书习俗为例[J].电影评介,2018(23):109-112.
[24] 潘朝霖.非公有制经济是水族地区经济发展的重头戏[J].贵州民族研究,2000(1 Suppl.):60-64.
[25] 杨秀璋.基于LDA模型和文本聚类的水族文献主题挖掘研究[J].现代计算机,2019(05):13-17.
[26] 杨建春,王佳联.民族地区旅游扶贫研究回顾与展望:基于文献计量分析[J].贵州民族研究,2019,40(08):118-124.

作者简介:

令狐秋萍(1995—),女,贵州遵义人,硕士研究生,研究方向为信息服务、图书情报

何世群(1997—),女,贵州遵义人,硕士研究生,研究方向为信息服务、图书情报

齐梦珂(1998—),女,河南漯河人,硕士研究生,研究方向为信息服务、图书情报

罗子江(1980—),男,贵州遵义人,教授,博士,研究方向为信息服务、模式识别

通信作者:杨秀璋(1991—),男,贵州凯里人,助教,硕士,研究方向为Web数据挖掘、知识图谱、人工智能,E-mail:1455136241@qq.com

收稿日期: 2021-10-11　**修稿日期:** 2021-12-24

Research on the Theme Evolution of Shui Literature Based on Knowledge Graph and Hierarchical Clustering

Linghu Qiuping, He Shiqun, Qi Mengke, Luo Zijiang, Yang Xiuzhang

(School of Information of Guizhou University of Finance and Economics, Guiyang 550025)

Abstract: The topic mining and topic evolution analysis of Shui culture can help understand Shui culture's development status, hot topics, and research trends from different perspectives. Also, this research provides a relevant reference value for the subsequent inheritance and literature mining of Shui culture. To this end, this paper collected 990 pieces of Shui literature from CNKI, extracted the keywords and themes of Shui culture. Then, it extracted the relationships among keywords by co-occurrence analysis and hierarchical clustering. Finally, it constructed the knowledge graph and topic evolution network of aquatic theme co-occurrence. The experimental results show that the method presented in this paper can effectively analyze the theme evolution trend of Shui literature. Moreover, it can find the co-occurrence relationship and similarity of feature words, construct the knowledge graph of Shui theme, and cluster to form five categories of Shui literature theme. In short, this paper has specific application value and theoretical significance for Shui culture research and literature mining.

Keywords: shui culture; thematic evolution; hierarchical clustering; knowledge graph; text mining