

优化关键词利用策略的共词分析研究^{*}

马宇驰¹ 牟冬梅² 杨鑫禹³

(1. 吉林财经大学图书馆, 长春 130117; 2. 吉林大学第一医院, 长春 130021;

3. 吉林大学公共卫生学院, 长春 130021)

摘要: 本文提出关键词利用策略的优化方案, 解决小数量级概念失焦、关键词组概念缺失等问题, 优化共词分析结果, 以发现潜在研究热点, 拓展研究热点主题识别的深度。关键词利用策略优化方案在高频词共词分析的基础上, 引入关键词与关键词组相结合的处理方案, 通过调整数据集范围, 实现共词分析结果优化。实证部分以“主动健康”主题为例, 使用DDA软件, 完成基于关键词利用策略优化方案的共词分析, 探测主动健康的学科主题热点。在初始发现的5类研究领域、12个热点之外, 基于关键词利用策略优化方案的共词分析扩展识别了7个潜在热点话题, 补充发现5个复合词组表达的研究概念。关键词利用策略优化方案令小数量级概念聚焦形成类团, 在聚类过程中得到表达, 令关键词组代表的概念得到完整呈现。

关键词: 共词分析; 关键词; 关键词组; 主动健康; 优化方案; DDA

中图分类号: G250.252 **DOI:** 10.3772/j.issn.1673-2286.2021.12.006

引文格式: 马宇驰, 牟冬梅, 杨鑫禹. 优化关键词利用策略的共词分析研究[J]. 数字图书馆论坛, 2021 (12): 34-40.

学科知识结构揭示模型将学科知识结构模块分为低中高三个层级, 其中对关键词、主题词、标题词、特征词等节点的共词分析是揭示中层级知识结构的主要技术方法, 在数据整合与智慧服务领域起到发现学科研究热点、探究学科发展进程的作用^[1]。结合时间轴属性, 共词分析能在纵向上反映一段时间内专业领域的动态发展演化历程, 在横向上反映某个时间节点静态知识单元分布结构^[2], 揭示了领域的基本特征, 对该领域的研究人员有重要的指导作用^[3]。共词分析方法基本分为6个环节, 即确定分析问题、术语词源选择、高频词选定、术语相关计算、多元统计分析 & 统计结果分析^[4]。共词分析结果的展示主要有聚类树图、战略坐标及社会网络图谱等方式, 聚类树图展现学科领域的主题结构, 战略坐标展现各主题在整个学科结构上的重要性或特性, 社会网络图谱展现各主题内部关系^[5]。

对于共词分析的改良优化, 学者从多个环节进行探索, 积累了丰富的研究成果。李纲等^[6]通过在关键词

统计和计数过程中使用改进加权算法, 实现了为关键词重要性加权, 而后采取混合关键词策略, 选取低频关键词和突发词, 优化了共词聚类效果以及关键词网络共现效果^[7]。傅柱等^[8]就共词分析词源选择述评了元数据取词与全文自动标引取词的研究现状, 就术语规范化问题述评了基于受控词典与基于人工方式的方案。杨丽等^[9]应用自然语言的处理方法, 考察了除关键词以外的分析单元, 提取了题名、摘要及全文的高频词, 识别动物学领域的研究热点。李锋^[10]提出了在选词个数范围内的高频区全部采纳、中低频区按2:1实行配额采纳的兼顾中低频关键词的选词方案。徐坤等^[11]提出次高频词概念, 利用次高频词进行了领域研究内容的聚类分析, 在揭示领域研究热点上具有较好效果。余本功等^[12]针对文献层面和词层面“同量不同质”、高频孤立词等问题, 提出基于文献多属性加权的共词分析方法。唐晓波等^[13]抽取并融合文本的词语特征和词权重特征, 对文本聚类, 提取关键词识别主题并进行分析。

^{*} 本研究得到国家自然科学基金项目“信息链视域下电子病历数据驱动健康服务供给侧决策的路径与模式研究”(编号: 71974074)资助。

滕广青等^[14]对科技信息多维复合分析演进过程进行归纳,总结通过多数据源更全面地识别领域知识或技术前沿。

但现有共词分析研究仍然存在一定的局限。胡昌平等^[15]指出,传统的词源选定、关键词提取方法忽略了关键词本身的特征,未充分发挥共词分析的优势,大量的强共现关系由中频词与高频词或中频词与中频词构成,截取高频词进行共词分析对重要关系的保留情况不能令人满意。词语可以是关键词、主题词或自然语言处理下的全文标引等不同元数据,不论来自何种分析单元,关键词词频及共现强度的幂律分布都带来难以调和的局限,导致共词分析在发现热点与全面概括两项目标难以达到平衡^[16]。牺牲低频关键词,基于高频词进行共词分析可以发现主题热点,但不能很好地探查学科全貌;针对低频词分析有助于探查隐含主题或前沿热点,然而人工降低截取关键词的词频阈值受人为影响大,也无法在发现重点和认识全貌中求得合理的平衡。

针对上述问题,笔者将关口前移,在聚类前提出关键词利用策略的优化方案,从而拓展研究热点主题识别的广度与深度,以期实现对共词分析理论的进一步完善和方法的进一步优化。

1 关键词利用策略优化方案

1.1 关键词分析问题梳理

高频词阈值选取方法主要有自定义选取法(频次选取法、前N位选取法、中心度选取法)、高低频词界定公式选取法、普赖斯公式选取法及混合选取法^[17]。然而无论采取何种方法,现有的共词分析都位于截取关键词步骤与聚类分析步骤之间,囿于聚类分析中类团形成的客观过程以及关键词表义能力的局限,关键词分析往往存在如下缺陷。①高频关键词聚类效果不足,在类团划分过程中出现“马太效应”。概念(Concept)是学科领域内科学共同体共享的基本知识单元,通过术语的形式被语言符号限定或表达出来,这些术语之间的相关强度决定了共词分析聚类的结果与形态。观察聚类过程可知,起初不存在明显确定的聚类中心,各组分在逐步形成类团时倾向发生强者愈强的吸引效应^[4]。致使与某个个体相关密切的其他成员会被吞噬纳入“超级类团”,强大的类团变得愈发强大,弱小类团地位愈

加无法突出,导致小数量级的类团无法脱出,失去表达信息的机会,最终使得以这些术语为载体的概念得不到有效的呈现。②关键词表达不完整,关键词组表达的复杂概念缺失。概念被术语表达,具有语义信息的术语按一定的逻辑结构组成文献,这些逻辑不仅包含物理顺序的先后关系,还存在句法篇章上的支配从属关系与隐含的语义联系^[18]。关键词分析中以词组形式凝练的概念,被拆分成单独的单词归档统计,导致词组切割,语义呈现不完善,重要信息不完整,概念的完整性受限,使得最终的分析结果不全面。

1.2 关键词利用策略优化方案构建

针对关键词分析的主要问题,提出关键词利用策略优化方案,实现对关键词的深入开发和对其携带信息的充分挖掘。关键词利用策略优化方案是指在获取数据集步骤之后、执行聚类分析步骤之前,在挖掘“超级类团”核心概念的一次分析基础上,通过多粒度、多角度的透视,调整数据集范围,进行二次聚类分析,聚焦小数量级关键词所代表的概念;随后引入关键词组,进行三次分析,增加由词组单元呈现的科学概念。

(1)调整数据集范围,聚焦小数量级关键词所代表的概念。关键词、热点都来自论文作者的实际工作,并非在科技论文写作时预先固定,因此“关键词-文献数据记录-热点”组成的网络往往是立体的,具有高自由度、高复杂度的特性,并不严格遵循一对一关系或一对多关系。在考虑去掉高频关键词时,提出“最大化去除”原则,以文献数据记录为最小单位,在文献数据集层面进行调整。其优势在于去掉引起“超级类团”的高频关键词的同时,同样移除了其来源文献数据记录中包含的其他关键词,实现了对与“超级类团”高度相关关键词的限制,使“超级类团”的清除更加彻底。进而更多地让位给前期被吸入“超级类团”中、未得到体现的关键词及其代表的研究热点,使整个发掘过程更加充分深入。

将关口前移,在聚类开始之前进行优化操作,将位于头部“超级类团”的核心概念所对应的文献数据集删除,使其余“弱势类团”得以免于吸入“超级类团”,小数量级关键词所代表的概念有机会在聚类过程中被充分表达,弱化了聚类过程中量级较大关键词引起的“马太效应”,使聚类结果能够更加丰富和全面地展示学科研究潜在热点。

(2) 引入关键词组, 增加由词组单元呈现的科学概念的表达。关键词组以多角度、更全面反映领域主题为目的, 可以是作者与关键词的组合, 也可以是关键词与其他词的组合; 可由数据库商提供、计算机程序自动标引, 也可由作者及领域专家人工组配。引入关键词组的共词分析能够利用关键词组类似于主题词组配的特性, 通过词组的形态保留研究的核心主题, 从不同视角补充主题信息, 以最贴近研究者本意的形式多维度地全面描述研究主题。

1.3 基于关键词利用策略优化方案的学科热点探测

关键词利用策略优化方案在标准的共词分析的基础上, 遵循“数据导入→优化关键词数据处理→聚类分析、识别热点→整合热点、形成最终探测结果”的过程, 其最终结果由3个识别结果汇总而成: ①对检索获得的原始数据集执行普遍采用的高频词共词分析, 所获得的初步发现的热点; ②调整数据集范围, 在高频关键词共词分析之外, 剥离原始数据集中词频数排行靠前概念对应的文献数据集, 对剩余文献数据集内的关键词执行第二次分析, 聚焦相对小数量级概念所代表的细节热点; ③引入关键词组视角, 对来自原始数据集的关键词组执行第三次聚类分析, 通过不破坏概念的方案, 透视作者在选题、构思、实验、分析等各环节的研究思维, 发现主题脉络, 获得补充热点信息。将三步的探测结果相结合, 使全部文献数据集蕴含的科学信息得到更充分体现, 最终填补聚类过程中“超级类团”导致的潜在信息遗漏, 提供还原作者科研意图的信息发现与学科热点探测服务。

2 基于关键词利用策略优化方案的研究热点探测实证研究——以“主动健康”主题为例

2.1 数据收集

我国自2016年至今陆续发布《“健康中国2030”规划纲要》《“十三五”卫生与健康科技创新专项规划》《健康中国行动(2019—2030年)》等文件, 以增进健康为导向, 推动以疾病治疗为中心向健康提升为中心的转变, 研发健康管理解决方案, 从注重“治已病”向

“治未病”过渡, 制定实施主动健康计划^[19-21]。主动健康的研究是实施“健康中国”行动的重要组成部分, 着眼当前是为人民谋幸福、谋健康, 展望长远是为民族谋复兴、谋发展。通过实证研究评测关键词利用策略优化方案对共词分析的提升效果, 同时分析国外主动健康相关研究进展, 以期为我国健康服务与管理领域相关研究提供参考借鉴。

选取数据来自Web of Science (WoS) 核心合集, 以“健康自我管理”(health self-management)、“主动医疗”(proactive medicine)、“主动干预”(proactive intervention)、“主动健康”(proactive health、positive health)、“疾病预防性治疗”(preventive treatment of disease)、“健康促进行为”(health promotion behavior)、“自我护理”(self-care)为检索词, 布尔逻辑检索词为“或”(OR), 限制语种为“英语”, 时间跨度为“1900—2020”, 索引包含SCI-EXPANDED、SSCI、A&HCI、CPCI-S、CPCI-SSH、ESCI。共检索到文献4 493篇, 构成本研究的原始数据。

2.2 截取高频词

将上述原始数据以文本形式导入, 将每条文献记录内容按作者、摘要、作者地址、作者关键词等提取, 写入对应字段, 生成DDA数据集。在分析之前对原始数据进行清洗, 剔除可能导致分析结果失准的“DEAN”4类数据: 数据库内重复或多个数据库检索结果合并之后的重复记录(Duplicates), 不符合检索需求的文献类型记录(Errors), 不同数据分类和标引规范下的相同概念关键词(Alias), 对研究热点造成干扰的低频关键词或子网络(Noises)^[22]。

本次实验应用的工具Derwent Data Analyzer (DDA)是由科睿唯安和乔治亚理工学院共同研发的信息情报分析和挖掘软件。统计“WoS入藏号”, 确保该字段内的每个对象出现次数为且仅为1, 结果中不存在重复记录。统计“文献类型”字段, 确保记录的文献均为学术信息本身, 不包含书目、传记、案例报告或数据集等, 符合本次实验需求。处理近义关键词和同义关键词, 清洗“关键词”字段, 最终得到原始数据集关键词列表(见表1), 共有关键词6 786个, 其中频次为1的关键词有4 701个。

最终采取混合选取法^[17]: 按高低频词分界公式^[23]确定高频词范围, 计算截取关键词降序排列的前96个,

表1 原始数据集关键词列表 (部分展示)

序 号	记录数量/条	关键词	序 号	记录数量/条	关键词	序 号	记录数量/条	关键词
1	4 076	self-care	38	45	randomized controlled trial	87	27	spinal cord injuries
2	692	diabetes	39	43	symptom management	88	26	Iran
3	466	heart failure	40	42	empowerment	89	26	Mobile health
4	269	nursing	41	42	qualitative	90	25	behavior
5	206	quality of life	42	42	well-being	91	25	hospitalization
6	191	chronic disease	43	41	primary care	92	25	oral self-care
7	180	self-efficacy	44	41	resilience	93	25	review
8	170	education	45	40	glycemic control	94	24	decision making
9	165	adherence	46	39	compassion fatigue	95	24	focus groups
10	165	patient education	47	39	health behavior	96	24	Humans
.....								

考虑研究需要及结果易读性,经多次预实验,应用前N位选取法截取排名前40的关键词进行分析。

2.3 关键词利用策略优化处理

观察预处理数据,首位关键词“自我护理/自我医疗”(self-care)频次4 076,数量和时间维度优势突出,其余高频关键词所代表的概念难以在其覆盖之下的分析中得到充分体现,因此考虑删除包含“self-care”关键词的部分文献数据集。

(1) 调整数据集范围。利用DDA调整数据集范围的功能,在原始数据集(包含关键词“self-care”的文献记录)基础之上调整生成子数据集(不包含关键词“self-care”的文献记录),为调整数据集范围的共词分析提供数据准备。

(2) 引入关键词组。关键词组提取自WoS数据集中的“Keywords Plus”字段。该字段是由计算机程序对每篇文章参考文献的标题内容进行自动标引而生成,数量多、含义广,可以有效探讨学科领域的知识结构^[24],能够从方法、技术等视角补充领域主题信息^[25]。将其写入DDA特定字段,并移除其中“rights reserved”“current study”等无实义类型词组以及“control group”“descriptive statistics”等描述实验操作的语义类型的词组,进行引入关键词组的共词分析。

2.4 结果与讨论

关键词利用策略优化方案下的实证研究共执行了

三次分析,最终的呈现结果来自三次共词分析结果的整合汇总。第一次是对于完整原始数据集执行的普遍采用的高频词共词分析,确保了“自我护理/自我医疗”(self-care)核心概念所代表的热点得到充分的呈现,保证核心重要概念热点不丢失。第二次共词分析将关口前移,通过聚类前的优化操作,以预实验结合人工经验的模式移除了“超级类团”(超级关键词)“self-care”,使小数量级概念类团得到体现,成功识别出潜在研究热点。第三次共词分析引入提取自WoS数据集中“Keywords Plus”字段的关键词组,凭借不同于第一、第二次共词分析关键词且同样具备探查学科领域知识结构作用的分析素材,明晰了热点侧重方向,拓宽了所发现研究热点的广度。将第一次共词分析发现的研究热点同第二、第三次经过关键词利用策略优化方案处理的共词分析新增研究热点进行对比,形成表2,并在下文具体分析。

2.4.1 原始数据集未经优化的共词分析结果

第一次对原始数据集执行未经优化的高频词共词分析方法,确定了5类基本研究领域,对应识别了12个研究热点(见表2):①护理学研究领域识别出“自主锻炼活动”(self-care maintenance)、“症状和体征监测”(self-care monitoring)、“症状和体征应对”(self-care management)3个热点;②临床医学研究领域识别出“糖尿病症状管理”(diabetes)、“心力衰竭症状管理”(heart failure)2个热点;③医学信息学研究领域识别出“远程医疗”(telemedicine)、“移动医疗”

表2 国外主动健康领域研究热点探测结果对比

单一高频关键词共词分析发现研究热点		经过关键词利用策略优化方案处理的共词分析新增研究热点	
研究领域	热点主题	调整数据集范围 识别潜在热点	引入关键词组 明晰热点侧重方向
护理学	自主锻炼活动 症状和体征监测 症状和体征应对	戒烟主动干预 感知健康状况	健康自主管理的“认知障碍”侧重
临床医学	糖尿病症状管理 心力衰竭症状管理	口腔保健干预	糖尿病研究的“血糖控制与主动干预”
医学信息学	远程医疗 移动医疗	个人健康记录	医学信息学的“心理计量学应用”
心理学	抑郁症状 焦虑 职业倦怠	同情疲劳	抑郁症状的“健康相关生活质量”方面
公共卫生学	医护人员自身属性 社区患者/家属健康教育	护理人员素养 营养与运动康复	“专业护理人员”的工作策略方面

(Internet) 两个热点; ④心理学研究领域识别出“抑郁症状”(depression)、“焦虑”(anxiety)、“职业倦怠”(burnout) 3个热点; ⑤公共卫生学研究领域识别出“医护人员自身属性”(empowerment)、“社区患者/家属健康教育”(health promotion) 2个热点。

2.4.2 调整数据集范围的共词分析结果

第二次分析是基于关键词利用策略优化方案的调整数据集范围的实验。对比第一次共词分析的热点识别效果, 它移除了“self-care”关键词所代表“超级类团”所对应的文献数据集, 弱化了聚类过程中“self-care”“超级类团”对诸如“戒烟主动干预”(smoking cessation、proactive intervention)、“感知健康状况”(perceived health status)、“口腔保健干预”(oral health)、“个人健康记录”(personal health record)、“同情疲劳”(compassion fatigue)、“护理人员素养”(health education) 和“营养与运动康复”(nutrition、physical activity) 等小数量级概念的吸引效应, 聚焦了7个小数量级关键词所代表的概念, 令其成功地在聚类过程形成类团, 在结果中得到表达。在原始数据固定的前提下, 关键词利用策略优化方案发挥了聚焦小数量级概念, 拓展学科潜在热点探测广度的优势。

2.4.3 引入关键词组的共词分析结果

第三次的分析是基于关键词利用策略优化方案的引入关键词组的实验。对比第一次共词分析的热点识别效果, 优势体现在: 引入了“认知障碍”(cognitive impairment)、“血糖控制”(glycemic control)、“心理计量学”(psychometric properties)、“健康相关生活质量”(health-related quality)、“专业护理人员”(health care providers) 等关键词组的方案, 不分解、不破坏作者科研概念的完整度, 避免以词组形式凝练表现出的概念被拆分成单独的单词归档统计, 以致语义信息不完善、重要信息不完整。在识别出研究对象之外, 捕获了更多关于热点的侧重方向、应用技术、发生场景等起补充作用的信息, 发现了健康自主管理的“认知障碍”侧重、糖尿病研究的“血糖控制与主动干预”、医学信息学的“心理计量学应用”等5个热点补充。在原始数据固定的前提下, 关键词利用策略优化方案发挥了保留关键词组概念, 明晰学科热点侧重方向, 扩展学科热点探测深度的优势。

本次实证分析在第一次共词分析初始发现的5类研究领域、12个热点之外, 经关键词利用策略优化方案处理后第二次及第三次共词分析扩展识别了7个潜在热点话题, 补充发现了5个复合词组表达的研究概念。对领域主题热点的拓展, 扩充了概念数目, 使之达到原有的2倍。

3 结论

针对共词分析研究中高频词分析聚类不足, 类团划分过程中“马太效应”突出, 以及独立关键词分裂了词组表达信息, 使主题不完整的问题, 提出了关键词利用策略优化方案。通过调整数据集范围削弱了聚类过程中类团间的马太效应。在原始数据集高频词共词分析的基础上, 基于调整之后的数据集成功聚焦小数量级概念, 发现多个潜在研究热点, 拓展了研究热点的识别广度。此外, 引入关键词组具有热点补充作用。通过词组间携带的组配特性, 从语义的层面保留了作者的科研意图, 在研究对象范围不变的情况下补充得出更多研究热点信息, 扩展了识别的深度。综上所述, 关键词利用策略优化方案令小数量级关键词代表的概念聚焦, 关键词组代表的概念呈现完整, 从多维度、多角度丰富了共词分析结果。通过调整数据集范围, 借助软件DDA的创建数据子集功能, 分析人员能够自由选择过滤某些“干扰”强的关键词, 放大知识单元分析的细节, 进一步增益了对知识全貌的还原。引入关键词组有助于获得隐含的主题信息, 在相当数量的高频关键词处理的基础上, 融入表意更丰富的关键词组, 是对共词分析的有益增补。不过, 本文仍然存在一定局限, 在剥离靠前概念对应的文献集时, 仍需以预实验结合人工经验的模式判断剥离分界程度, 而且共词分析在得到研究热点主题后, 需要返回原始论文中进行学科领域认知与内容分析, 对内容的概括与认知程度取决于研究执行者的学术经验。

参考文献

- [1] 牟冬梅, 郑晓月, 琚沅红, 等. 学科知识结构揭示模型构建[J]. 图书情报工作, 2017, 61(12): 6-13.
- [2] 钟伟金, 李佳. 共词分析法研究(一)——共词分析的过程与方式[J]. 情报杂志, 2008(5): 70-72.
- [3] 黄月, 王鑫. 基于高维稀疏聚类的知识结构识别研究[J]. 现代情报, 2019, 39(12): 72-80.
- [4] 李纲, 巴志超. 共词分析过程中的若干问题研究[J]. 中国图书馆学报, 2017, 43(4): 93-113.
- [5] 杨颖. 基于共词分析的学科结构可视化研究[D]. 沈阳: 中国医科大学, 2010.
- [6] 李纲, 李轶. 一种基于关键词加权的共词分析方法[J]. 情报科学, 2011, 29(3): 321-324, 332.
- [7] 李纲, 李昱瑶, 谢子霖, 等. 混合关键词选择策略对共词分析效果的影响研究[J]. 情报理论与实践, 2017, 40(11): 110-116.
- [8] 傅柱, 王曰芬. 共词分析中术语收集阶段的若干问题研究[J]. 情报学报, 2016, 35(7): 704-713.
- [9] 杨丽, 张彤彤, 周文杰. 共词分析识别研究热点的效标关联度研究: 基于自然语言处理[J]. 图书与情报, 2018(1): 15-19.
- [10] 李锋. 兼顾中低频关键词的共词分析实践——以图书情报学高被引文章为例[J]. 图书馆杂志, 2018, 37(4): 34-42.
- [11] 徐坤, 毕强. 次高频关键词的选择及在共词分析中的应用[J]. 情报理论与实践, 2019, 42(5): 148-152.
- [12] 余本功, 王龙飞, 陈杨楠, 等. 基于文献多属性加权的共词分析方法研究[J]. 情报科学, 2019, 37(1): 122-128.
- [13] 唐晓波, 李津. 在线健康社区信息需求主题分析[J]. 数字图书馆论坛, 2019(2): 12-17.
- [14] 滕广青, 叶心, 郭思月, 等. 科技信息分析从单一维度到多维复合的演进[J]. 数字图书馆论坛, 2019(12): 2-8.
- [15] 胡昌平, 陈果. 科技论文关键词特征及其对共词分析的影响[J]. 情报学报, 2014, 33(1): 23-32.
- [16] 杨建林. 关键词选择策略及其对共词分析的影响[J]. 情报学报, 2014, 33(10): 1083-1090.
- [17] 刘奕杉, 王玉琳, 李明鑫. 词频分析法中高频词阈值界定方法适用性的实证分析[J]. 数字图书馆论坛, 2017(9): 42-49.
- [18] 郭红梅, 张智雄. 基于图挖掘的文本主题识别方法研究综述[J]. 中国图书馆学报, 2015, 41(6): 97-108.
- [19] 中共中央 国务院印发《“健康中国2030”规划纲要》[EB/OL]. [2021-12-01]. http://www.gov.cn/zhengce/2016-10/25/content_5124174.htm.
- [20] 关于印发《“十三五”卫生与健康科技创新专项规划》的通知[EB/OL]. [2021-12-01]. http://www.most.gov.cn/tztg/201706/t20170613_133484.html.
- [21] 新华社. 聚焦“治未病”健康指标纳入政府考核——权威解读健康中国行动有关文件[EB/OL]. [2021-12-01]. <https://baijiahao.baidu.com/s?id=1639124386954279536&wfr=spider&for=pc>.
- [22] 潘玮, 郑鹏, 黄锦泉, 等. 基于数据清洗“DEAN”流程的健康信息领域研究热点探测[J]. 现代情报, 2018, 38(10): 73-77.
- [23] 孙清兰. 高频词与低频词的界分及词频估算法[J]. 中国图书馆学报, 1992(2): 78-81, 95-96.
- [24] 章娟, 卢祖洵, 段志光. Web of Science数据库中Keywords Plus的准确性研究——以患者依从性研究论文为例[C]. 第一届两岸三地科学计量学与信息计量学研讨会论文集, 西安: 中国科学学与科技政策研究会, 中国科学技术信息研究所, 2013.

[25] ZHANG J, YU Q, ZHENG F, et al. Comparing keywords plus of WOS and author keywords: a case study of patient adherence

research [J]. Journal of the Association for Information Science and Technology, 2016, 67 (4) : 967-972.

作者简介

马宇驰, 男, 1989年生, 硕士, 助理馆员, 研究方向: 数据整合与智慧服务。

牟冬梅, 女, 1970年生, 博士, 教授, 通信作者, 研究方向: 医学数据整合与智慧服务, E-mail: moudm@jlu.edu.cn。

杨鑫禹, 女, 1996年生, 博士研究生, 研究方向: 医学数据整合与智慧服务。

Research on Co-word Analysis Based on Keyword Optimization

MA YuChi¹ MU DongMei² YANG XinYu³

(1. Library of Jilin University of Finance and Economics, Changchun 130117, P. R. China; 2. The First Hospital of Jilin University, Changchun 130021, P. R. China; 3. School of Public Health, Jilin University, Changchun 130021, P. R. China)

Abstract: This paper propose a keyword optimization plan to solve the problems of small-scale concept defocus and keyword group concept missing, optimize the results of co-word analysis, and expand the breadth and depth of research hotspots recognition. The keyword optimization mode is based on the high-frequency word co-word analysis, introduces the processing method of combining keywords and keyword groups, and obtains the optimization of the co-word analysis results by adjusting the range of the data set. The empirical part takes the topic of "proactive health" as an example, uses DDA software to complete a co-word analysis based on the keyword optimization plan, and detects the research hotspots of proactive health. A keyword optimization plan is proposed, which optimizes the co-word analysis by adjusting the scope of the data set and introducing keyword group recognition. In the empirical research part, in addition to the 5 types of research fields and 12 hotspots initially discovered, the co-word analysis based on the keyword optimization plan expanded to identify 7 potential hotspots and supplemented the discovery of 5 research concepts expressed by compound phrases. The keyword optimization mode enables small-scale concepts to focus on forming clusters, which can be expressed in the clustering process, and also enables the concepts represented by keyword groups to be fully presented.

Keywords: Co-word Analysis; Keywords; Keyword Groups; Proactive Health; Optimization Plan; DDA

(收稿日期: 2021-12-01)