

# 一种新兴的学科领域历史根源探究方法: RPYS\*

■ 李信<sup>1,2</sup> 陆伟<sup>1,2</sup> 李旭晖<sup>1,2</sup>

<sup>1</sup> 武汉大学信息管理学院 武汉 430072 <sup>2</sup> 信息检索与知识挖掘研究所 武汉 430072

**摘要:** [目的/意义]探索一个学科或研究领域的历史根源问题,找出对该学科领域起源具有重要作用的根源文献,对于学科建设和研究具有重要意义。[方法/过程]引入一种被称为“参考文献出版年图谱”(RPYS)的新方法,对其概念、特点和原理进行详细阐述,并将RPYS分析过程归纳为4个步骤。在此基础上,以健康信息素养研究领域为例,应用RPYS从参考文献的角度对该领域的历史根源文献进行识别和分析。[结果/结论]结果表明,运用RPYS能够较为准确地揭示健康信息素养研究领域的历史根源文献,同时发现RPYS也可以用于学科或研究领域的研究热点、研究趋势的探究。

**关键词:** RPYS 学科根源 参考文献 出版年图谱 健康信息素养

**分类号:** G250

**DOI:** 10.13266/j.issn.0252-3116.2016.20.009

## 1 引言

一个新的学科或研究领域的出现,往往是其他多个特定学科或研究领域中已有研究成果的深化或拓展,这体现了科研工作的继承性,表现在科研论文写作中即为对参考文献的引用<sup>[1-2]</sup>。然而,由于一些学科领域具有多学科交叉性或一定的技术难度,刚进入该学科的科研工作者很难全面透彻地了解前人对于学科起源和发展的贡献,继而可能不知道该追溯多早的文献,导致一些重要的历史文献被忽略甚至成为无人问津的“睡美人”文献<sup>[3]</sup>。因此,探索一个学科或研究领域的历史根源问题,找出对该学科起源具有重要作用的根源文献,对于学科建设和研究具有重要意义。通过文献调研发现,科学计量学、情报学等领域的先驱们对该研究问题已有一定的探讨,并取得了一些成果,比如E. Garfield等<sup>[4]</sup>于2003年基于引文分析开发的引文图谱分析软件HistCite可以快速描绘一个学科领域的发展历史,定位领域的重要文献和最新的重要研究成果;再如张卫民等<sup>[5]</sup>和C. Machado等<sup>[6]</sup>从历史学的角度出发,通过人工统计和分析的方法分别对中国林

业经济管理学科和脑死亡研究领域的历史起源问题进行了探索分析。上述成果虽然已经可以解决一些问题,但仍然存在不足之处:①HistCite主要是对文献的“本地引用次数”(location citation score, LCS)和“全局引用次数”(global citation score, GCS)进行统计和可视化呈现,无法对所研究领域之外的文献进行探讨,而往往遗漏的这少量文献可能对学科领域(尤其是多学科交叉形成的新学科)的起源起到重要作用;②在探测时间范围上也限制在学科领域产生之后,而学科领域的根源文献大多是在学科形成之前或萌芽期;③从历史学角度和人工统计的方法来探讨学科领域起源,已经不适用于大数据时代的海量领域学术资源。

在第十四届国际科学计量学和信息计量学大会(14th International Conference on Scientometrics & Informetrics, ISSI 2013)上,W. Marx等<sup>[7]</sup>首次提出了“参考文献出版年图谱”(RPYS)这一学科领域历史根源探究的新方法,该方法从参考文献的角度出发,较好地弥补了HistCite的不足,可以一定程度地揭示一个学科或研究领域可能的历史根源文献。基于此,本研究在

\* 本文系国家自然科学基金面上项目“基于多语义信息融合的学术文献引文推荐研究”(项目编号:71673211)和国家自然科学基金面上项目“本体导向的大规模语义信息声明式抽取方法”(项目编号:61272110)的研究成果之一。

作者简介:李信(ORCID:0000-0002-8169-6059),博士研究生;陆伟(ORCID:0000-0002-0929-7416),教授,博士生导师,通讯作者,E-mail:weilu@whu.edu.cn;李旭晖(ORCID:0000-0002-1155-3597),副教授,博士。

收稿日期:2016-08-22 修回日期:2016-10-04 本文起止页码:70-76 本文责任编辑:刘远颖

介绍 RPYS 的概念和原理的基础上, 将 RPYS 分析过程归纳为 4 个步骤, 以健康信息素养研究领域为例, 应用 RPYS 从参考文献的角度来对该领域的历史根源文献进行识别和分析。

## 2 RPYS 的概念及其原理

### 2.1 RPYS 的概念

RPY, 即 reference publication year, 指参考文献的出版年份。需要指出的是, 从 RPY 角度出发来解决科学计量学或文献计量学领域的问题并不是一个全新的视角, 早在 1974 年 D. Price<sup>[8]</sup> 就对学术文献的参考文献出版年份 (RPY) 进行了探讨, V. Raan 等<sup>[9]</sup> 还将 RPY 应用于测量科学的增长和对科学领域重大突破的探测等。在 RPY 的基础上, 2013 年 W. Marx 等<sup>[7]</sup> 类比物理学等自然科学领域的“图谱”(spectroscopy) 提出了 RPYS 的概念, 即 reference publication year spectroscopy, 本文将其译为“参考文献出版年图谱”, 指以某一个学科领域的全部相关文献所引用的全部参考文献的“出版年份”为横轴, 以每年全部被引参考文献的“总被引频次”为纵轴而形成的分布图<sup>[7]</sup>。在不同的学科领域中, 该图谱具有以下共同特点: ①参考文献的出版年份跨度包含“在学科领域产生之前”和“学科领域产生之后”; ②学科领域产生之前, 每年的参考文献数量和总被引频次远小于学科领域产生之后; ③学科领域产生前后, RPYS 一般会出现一个或多个峰值。W. Marx 等通过实验发现, 利用 RPYS 对某一学科领域文献集进行分析可以较为准确地揭示学科领域的历史根源<sup>[7]</sup>, 具体的 RPYS 分析原理将在下一节详细叙述。

RPYS 由 W. Marx 等<sup>[7]</sup> 于 2013 年提出后, 短时间内在国际上得到了较多应用, 从 Web of Science 的检索结果来看, RPYS 的相关应用文献已经超过 10 篇(检索时间: 2016 年 5 月 2 日), 这体现了该方法具有较好的可行性和可靠性。具体的应用实例有: A. Barth 等<sup>[10]</sup> 利用 RPYS 探讨了粒子物理学中希格斯玻色子(Higgs boson)的起源文献, L. Leydesdorff 等<sup>[11]</sup> 借用 RPYS 对石墨烯和太阳能电池领域的历史根源文献进行了探索, 还有 W. Marx 等<sup>[12]</sup> 对生物学中“达尔文雀传奇”(Darwin's Finches)<sup>[13]</sup> 进行了探索, 2015 年 J. Comins 等<sup>[14]</sup> 对“全球定位系统”起源的研究以及 2016 年 B. Elango 等<sup>[15]</sup> 对“摩擦学”起源的探索, 均使用了 RPYS 分析的方法, 并取得了较好的研究成果。通过以上研究可以发现, 利用 RPYS 进行学科领域历史根源文献探测的方法具有较好的普适性。此外, 目前还没有发

现国内有与 RPYS 相关的研究。

### 2.2 RPYS 分析原理

利用参考文献出版年图谱 (RPYS) 对学科领域的历史根源进行分析、探索的过程称为参考文献出版年图谱分析。其原理是: 在一个学科领域的相关文献所引用的所有参考文献中, 只有很小比例的参考文献是在该学科领域产生之前发表的; 且在这小部分参考文献中, 通常存在几篇文献的被引用频次远高于同年或前后几年内发表的其他参考文献, 那么这几篇文献很可能是对该学科领域的产生具有重要作用的根源文献。同时, 满足上述条件的参考文献在图谱上出现的位置一定是图谱上的某个峰值。因此, 通过对参考文献出版年图谱在学科领域产生之前的峰值进行分析, 就有可能找到该学科领域的历史根源文献<sup>[7]</sup>。本文总结了 11 篇 RPYS 分析应用的相关文献, 将参考文献出版年图谱 (RPYS) 分析归纳为以下几个步骤: ①从 Web of Science 检索和收集该学科领域的所有相关文献的题录信息(选择“全记录与引用的参考文献”); ②从题录信息中抽取所有的参考文献及其发表年, 制作参考文献出版年图谱, 找到在学科领域产生之前的峰值, 即重要的参考文献出版年份; ③对峰值年份进行单独分析, 找出该年被引用频次最高的参考文献, 即寻找可能的候选根源文献; ④检验: 由学科或研究领域专家鉴定, 确定候选历史根源文献是否符合事实。需要注意的是, 因为 RPYS 分析事实上只能识别出某一领域可能的根源文献, 因此步骤④是必须的。

## 3 研究设计

### 3.1 数据来源

本研究的数据来源于 Web of Science<sup>TM</sup> (以下简称“WoS”)核心集合, 检索策略为: 主题 = (health literacy) OR 主题 = (health information literacy); 时间跨度 = 1900 - 2015, 引文索引 = SCI-EXPANDED(1979 -), SSCI(1900 -), A&HCI(1975 -), CPCI-S(1996 -), CPCI-SSH(1996 -), 其中括号内的年份为本单位(武汉大学)购买引文索引数据的起始年; 检索时间: 2016 年 4 月 10 日。最终得到与健康信息素养相关的文献共 8 116 篇(具体见表 1), 选择“全记录 and 引用的参考文献”导出全部文献题录信息。

表 1 实验数据概览

名称(单位)	数量
总文献数(篇)	8 116
文献出版年(年)	1975 - 2015
引用的总参考文献数(篇)	166 820
参考文献出版年(年)	1800 - 2015

健康信息素养是由医学、心理学、教育学、信息科学和图书情报科学等形成的交叉科学,这一术语第一次出现在1974年K. Simonds<sup>[16]</sup>的一篇关于健康教育作为政策问题对卫生保健系统、教育系统和大众传播影响的文章中,20多年后有关健康信息素养的研究才逐渐展开。1995年,美国《国家健康教育标准》中第一次对健康信息素养给出定义,即“个体具有获得、解释和理解基本健康相关信息与服务,并运用这些信息或服务来促进个体健康的能力”<sup>[17]</sup>。随后,世界卫生组织、美国国家医学会、美国国家医学图书馆才分别在1998年、1999年和2000年对健康信息素养给出各自的定义<sup>[18-20]</sup>。因此,本文将1995年作为现代健康信息素养领域的研究起点。

### 3.2 实现过程

本研究采用了2014年荷兰阿姆斯特丹大学L. Leydesdorff等开发的RPYS软件包(L. Leydesdorff的个人网页 <http://www.leydesdorff.net/software/rpys/>上可免费获取),包含3个子程序:rpys.exe、yearcr.exe和RefMatchCluster.jar。将8116篇健康信息素养相关文献的题录信息整合到一个单独的文件并重命名为data.txt,执行rpys.exe对data.txt进行处理可得到rpys.dbf、rpys.dbf的内容包含两列数据,分别对应参考文献出版年图谱的横轴和纵轴;此外,rpys.exe还返回了一个名为median.dbf的文件,该文件包含了该RPY的参考文献的总被引频次相对于该RPY的前一年、前两年、该RPY、后一年、后两年(即 $[t-2;t-1;t;t+1;t+2]$ ,其中t为该RPY)的总被引频次的平均数的偏差:若该RPY的参考文献总被引频次高于5年平均值,则偏差大于0;若该RPY的参考文献总被引频次等于5年平均值,则偏差等于0;否则,偏差小于0。使用EXCEL打开rpys.dbf和median.dbf,选择数据、调整合适的RPY跨度,便可以制作得到健康信息素养研究领域的参考文献出版年图谱(见图1)和Quantile值热度图(见图2)。为了比较不同年份的参考文献对该学科领域的重要程度,本研究将RPY按照参考文献的总被引频次降序排列,分别计算每一年的Quantile值,计算公式如下:

$$Quantile = ((i - 0.5) / n * 100) \quad (1)$$

其中i指某一特定的RPY的序号,n为参考文献出版年的总年数。Quantile的值可以在median.dbf文件中得到,某一出版年的Quantile的值越大,表示该年的总被引频次最高。

执行yearcr.exe,选择需要分析的起始年份和结束

年份,对cr.dbf(rpys.exe返回)进行处理,就可以得到包含所选时间段内每年所有参考文献的篇名和单独对应的被引用次数的yearcr.dbf文件;执行RefMatchCluster.jar可以对yearcr.dbf内可能存在的重复参考文献进行合并处理,减小误差。

## 4 结果和分析

从1900-2015年健康信息素养领域参考文献出版年图谱(见图1)可以发现参考文献被引总频次最高的是2006年,表明2006年是全球健康信息素养领域研究的高产期,很多重要的、高质量的领域相关文献在这一年发表。从健康信息素养领域在不同RPY(1900-2015年)的年被引总频次的Quantile值热度图(见图2)可以看出,被引用最频繁的参考文献的出版年份是2002-2011年;但也同时发现,一些年份较早的RPY也很重要,比如1908年、1948年、1953年等。

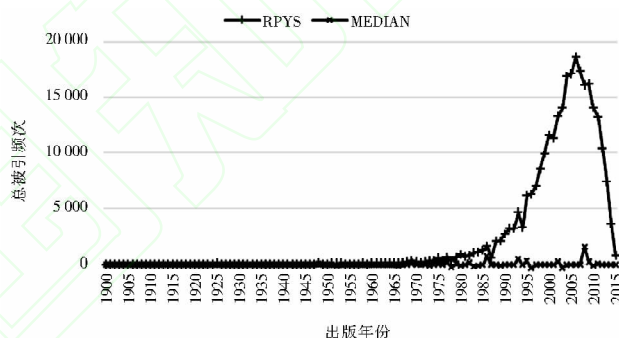


图1 1900-2015年健康信息素养领域参考文献出版年图谱

1900	19.94	1924	47.58	1948	77.52	1972	80.77	1996	92.11
1901	24.00	1925	65.98	1949	72.78	1973	78.82	1997	89.16
1902	24.55	1926	45.44	1950	76.63	1974	84.97	1998	93.26
1903	19.89	1927	35.89	1951	67.12	1975	87.02	1999	94.31
1904	43.95	1928	39.69	1952	76.93	1976	82.92	2000	95.45
1905	29.84	1929	24.35	1953	87.32	1977	89.12	2001	95.41
1906	29.79	1930	34.79	1954	76.73	1978	82.07	2002	97.55
1907	34.94	1931	43.16	1955	76.53	1979	81.17	2003	97.60
1908	59.93	1932	55.64	1956	76.83	1980	83.36	2004	97.80
1909	34.89	1933	45.19	1957	76.68	1981	82.22	2005	98.85
1910	4.50	1934	55.39	1958	77.03	1982	81.31	2006	98.95
1911	9.74	1935	15.59	1959	76.88	1983	68.41	2007	99.90
1912	0.10	1936	25.54	1960	67.22	1984	78.46	2008	97.70
1913	0.05	1937	45.49	1961	63.17	1985	83.51	2009	98.75
1914	0.00	1938	35.98	1962	58.08	1986	89.61	2010	96.85
1915	23.90	1939	38.84	1963	54.37	1987	84.56	2011	97.50
1916	44.84	1940	66.13	1964	77.27	1988	78.66	2012	97.36
1917	34.45	1941	58.79	1965	69.42	1989	88.71	2013	97.21
1918	45.29	1942	64.74	1966	23.47	1990	84.76	2014	97.91
1919	63.85	1943	77.43	1967	68.57	1991	85.81	2015	96.26
1920	53.21	1944	55.74	1968	78.67	1992	87.86		
1921	63.80	1945	46.23	1969	90.87	1993	91.01		
1922	44.40	1946	36.38	1970	83.72	1994	86.96		
1923	55.24	1947	64.18	1971	81.62	1995	93.06		

图2 健康信息素养领域在不同RPY(1900-2015年)的年被引总频次的Quantile值

注:Quantile的值越大,颜色越黑;反之,颜色越白

因此,本研究根据现代健康信息素养研究的起点将RPY的时间跨度设定为1900-1994年,以全面、系统地梳理健康信息素养领域的历史起源情况。同时,

将 1900-1994 年分为 1900-1950 年和 1951-1994 年两个阶段, 原因如下: ①1900-1950 年间相继爆发两次世界大战, 美国等国家政府大幅度加大对科研活动的投入, 极大地推动了医学研究和医学教育的发展, 但是在该时期医学领域大多只注重患者的生理学属性, 忽视环境、行为方式、心理因素和社会等对疾病和健康的影响, 只有极少数学者意识到这些问题<sup>[21]</sup>, 因此, 对 1900-1950 年间的 RPYS 进行分析就是希望找到少数这些对健康信息素养领域形成起到萌芽作用的重要历史根源文献; ②1951-1994 年间, 随着计算机网络技术的逐渐兴起和信息时代的来临, 患者获取健康信息和服务的途径发生了巨大的转变, 公共卫生和临床领域也开始关注患者获取、理解和运用健康信息和服务的能力对其健康结果的影响<sup>[22]</sup>, 对这段时间进行 RPYS 分析可以找到该时期对健康信息素养研究领域形成过程具有重要影响的历史文献。

#### 4.1 1900-1950 年健康信息素养领域参考文献出版年图谱分析

从图 3 可以看出, 1900-1950 年间, 健康信息素养领域参考文献出版年图谱出现了 5 个明显的峰值, 分

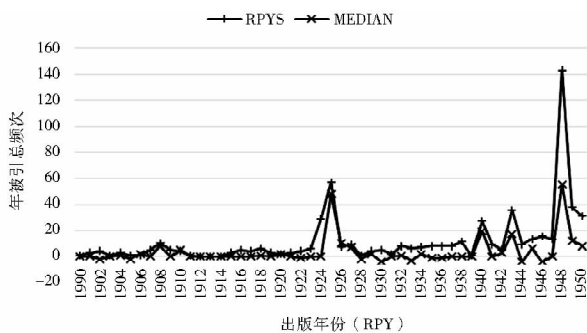


图 3 1900-1950 年健康信息素养领域参考文献出版年图谱

别是 1908 年、1925 年、1940 年、1943 年和 1948 年, 这表明在 1900-1950 年存在一些对健康信息素养研究的起源和发展比较重要的历史文献。W. Marx 等<sup>[12]</sup>在 2014 年的研究指出, 在对参考文献出版年图谱的峰值进行分析的时候, 往往只需对单篇被引次数最高的参考文献进行分析, 即单篇被引频次最高的文献很可能是对该学科或领域的起源和发展起到重要作用的历史文献。在此基础上, 本文对 1900-1950 年被引次数最多的参考文献进行了整理, 发现被引次数最多的文献恰好出现在 5 个峰值, 具体情况如表 2 所示:

表 2 1900-1950 年间被引频次最多的参考文献

参考文献出版年	年被引总频次	被引频次最高的文献 / 被引频次
1908 年	10	W. Henry. Aristotle's politics. Clarendon Press(1908) / 7 次
1925 年	57	E. Bogardus. "Social distance and its origins." Journal of Applied Sociology (1925): 216-226. / 36 次
1940 年	27	C. Walter. "A self-administering scale for measuring intellectual impairment and deterioration." The Journal of Psychology (1940): 371-377. / 22 次
1943 年	36	CO. Houle, R. Flesch. "Marks of readable style; a study in adult education." Teachers College Contributions to Education (1943). / 28 次
1948 年	143	R. Flesch. "A new readability yardstick." Journal of applied psychology (1948): 221. / 100 次

由图 3 和表 2 可知, 第一个峰值指向 1908 年牛津大学出版社出版、W. Henry 编著的 Aristotle's Politics (《亚里士多德政治学》)<sup>[23]</sup>, 该书是古希腊第一部全面、系统地论述政治问题的著作, 内容涉及政体的分类、变革和维持, 理想城邦及其构建, 公民教育等方面。著作中亚里士多德主张通过公民教育和培养来实现城邦生活的完善与和谐, 并关注了公民教育的方方面面, 比如婚配、生育、儿童等。阅读文献可以发现, 健康教育是公民教育的一个重要组成部分, 通过健康教育可以使人们自觉地采纳有益于健康的行为和生活方式, 从而促进身体健康和维持社会稳定。同时, 健康信息素养与健康教育之间的关系非常密切, 前者是手段方法, 后者是成果体现。

第二个峰值指向 1925 年 E. Bogardus 等<sup>[24]</sup>发表的

一篇题为 *Social distance and its origins* (《社会距离和它的起源》) 的文献, 该文讨论了社会距离的内涵及其起源和演化情况, 并在此基础上将社会距离从概念变成了具体的测量刻度, 即通过考察行动主体是否愿意与其他人或其他社会群体交往来测量他们之间的社会距离。笔者认为该文为健康信息素养研究的热点问题——健康信息素养的评估研究及健康信息素养评估工具的开发等提供了一定的参考依据。第三个峰值指向 1940 年 W. Shipley 等<sup>[25]</sup>发表的一篇题为 *A self-administering scale for measuring intellectual impairment and deterioration* 的文献, 该文主要介绍了一个用于对智力和精神疾病进行客观、快速的自我评估工具, 工具的开发和试用是基于临床实验环境的, 这对临床视角下健康信息素养内涵的发展及临床健康信息素养测评工具

的开发具有很好的借鉴意义。

第四、五个峰值分别指向1943年、1948年 C. O. Houle 和 R. Flesch<sup>[26-27]</sup>发表的两篇文章,两者均与可读性相关,前者研究了影响材料可读性的几个指标,比如文章摘要中句子的长度、词频等,并在此基础上提出了著名的可读性公式“Flesch-Kincaid Formulae”;后者则对前者提出的公式进行了修订,提出了测量文本材料可读性的新标准体系。

整体来看,5篇文献中有3篇发表在心理学相关的杂志上,1篇发表在教育相关的杂志上,1篇涉及公民教育方面,反映了健康信息素养研究的起源与心理学、教育学领域有非常紧密的联系;两篇文献与可读性相关,这反映了早期健康信息素养研究中对患者读写能力的强调;同时,4篇文献与测评相关,表明健康信息素养评估及评估工具的开发是健康信息素养研究领域的热点问题。

#### 4.2 1951-1994年健康信息素养参考文献出版年图谱分析

从图4可以得到,1951-1994年间,健康信息素养

参考文献出版年图谱也出现了5个显著的峰值,结合 median. dbf 文件,可以知道5个峰值分别对应1953年、1969年、1977年、1986年和1993年,反映1951-1994年中有对健康信息素养研究的发展比较重要的研究成果发表。对1900-1950年被引次数最多的参考文献进整理,发现被引次数最多的文献也恰好出现在5个峰值,具体情况见表3。

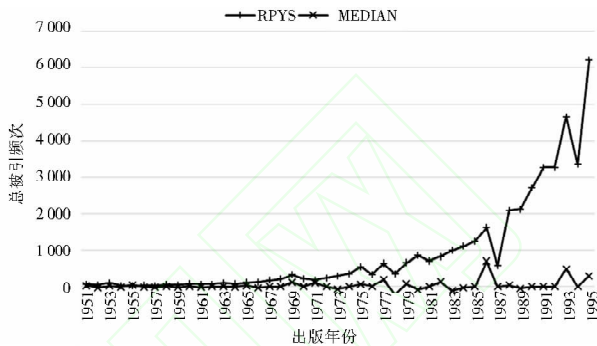


图4 1951-1994年健康信息素养领域参考文献出版年图谱

表3 1951-1994年间被引频次最多的参考文献

参考文献出版年	年被引总频次	被引频次最高的文献 / 被引频次
1953年	108	W. Taylor. "Cloze procedure: a new tool for measuring readability." <i>Journalism and Mass Communication Quarterly</i> (1953): 415. / 89次
1969年	333	M. Laughlin, G. Harry. "SMOG grading-a new readability formula." <i>Journal of Reading</i> (1969): 639-646. / 132次
1977年	646	A. Bandura. "Self-efficacy: toward a unifying theory of behavioral change." <i>Psychological Review</i> (1977): 191-193. / 374次
1986年	1 625	A. Bandura. <i>Social foundations of thought and action: A social cognitive theory</i> . Prentice-Hall, Inc (1986)/ 533次
1993年	3 552	T. C. Davis, S. Long, R. Jackson. "Rapid estimate of adult literacy in medicine: a shortened screening instrument." <i>Family Medicine</i> (1993): 391-395 / 871次

第一、二个峰值分别指向1953年 W. Taylor 等<sup>[28]</sup>和1969年 G. McLaughlin 等<sup>[29]</sup>发表的与可读性相关的两篇文章——*Cloze procedure: a new tool for measuring readability* 和 *SMOG grading-a new readability formula*,前者提出了一个测量文本材料可读性的新工具,后者给出了一个可读性评估的新公式。

第三、四个峰值指向 A. Bandura 等在1977年发表的 *Self-efficacy: toward a unifying theory of behavioral change*<sup>[30]</sup>和在1986年出版的心理学领域的里程碑式著作 *Social Foundations of Thought and Action: A Social Cognitive Theory*<sup>[31]</sup>。前面一篇文章提出了一个可以解释和预测在不同的治疗模式下患者心理变化情况的理论框架;后者则从社会认知的角度提出了经典的人类思想、动机和行为的理论,对他之前的社会学习理论 (social learning theory) 进行了扩充,并形成了相互决定

论 (reciprocal determinism) 的基础,阐释了人类认知结构、外显行为和所处环境之间的关系。

第五个峰值指向1993年 T. C. Davis 等<sup>[32]</sup>发表的一篇题为 *Rapid estimate of adult literacy in medicine: a shortened screening instrument* 的文章,介绍了一个用于公共卫生和基础医疗环境下快速筛查阅读能力较低患者的筛查量表,它将患者的阅读等级对应到学校教育的不同年级。虽然 REALM 量表只对患者的医学语言阅读能力进行了考察,但是目前它仍被作为其他健康信息素养测评工具开发的重要依据和“金标准”<sup>[33]</sup>。

总结 RPYS 分析和验证的结果,可以得到以下结论:健康信息素养虽然于1995年才在美国《国家健康教育标准》中被首次定义,但健康信息素养的历史根源可以追溯到1908年;作为健康教育的结果体现,健康信息素养最开始萌芽于亚里士多德《政治学》中对公

民教育意义的阐释,并与教育学和心理学领域有着非常紧密的联系。社会距离和社会认知理论、相互决定论等为健康信息素养领域的产生提供了理论依据。20世纪中后期,文本可读性研究的深入和可读性评估公式的提出、改善以及相应评估工具的开发,对早期临床视角健康信息素养概念的形成及其评估研究发挥了重要作用。

## 5 讨论

本文利用 RPYS 对健康信息素养研究领域的历史根源进行了探索,实验结果得到 10 篇高被引文献,经验证,它们在健康信息素养研究领域的形成过程中均发挥了举足轻重的作用,这表明 RPYS 能够较准确地识别健康信息素养研究领域的重要历史文献。同时,10 篇参考文献中有 6 篇是有关评估理论、方法或评估工具开发的,这与现代健康信息素养研究的热点和难点——“健康信息素养的评估问题”<sup>[34]</sup>相吻合,说明 RPYS 还可以用于揭示某个学科或研究领域的研究热点和研究趋势;此外,有 5 篇文献和文本可读性有关,这与英国健康信息素养专家 D. Nutbeam<sup>[35]</sup>所提出的功能性健康信息素养中对患者的读写能力的要求相对应,与传统临床视角的健康信息素养概念的形成有紧密联系。

与 HistCite 相比,RPYS 分析从领域相关文献的参考文献集的出版年份出发,而不是仅仅关注领域相关文献集,因此,一方面,将根源文献的探测时间范围扩展到学科领域产生之前,另一方面也使根源文献的探测领域范围不再限制在该学科和领域,可有效地解决多学科交叉形成的学科领域的历史根源探测问题。然而,虽然本实验的结果较为理想,但仍需要注意的是,通过 RPYS 对数据分析得到的结果只是可能的候选根源文献集,要确定候选文献是否为真正学科领域的根源文献,必须要经过仔细的分析和特定专业领域的专业人员的判断和鉴别。

## 6 结语

本文介绍了一种新兴的学科领域历史根源探究方法——“参考文献出版年图谱”(RPYS),在对其概念、特点和原理进行详细阐述的基础上,以健康信息素养研究领域为例进行了实践,得到了较为理想的结果。但也仍然存在一些问题需要进一步探索,比如目前 RPYS 分析主要针对的是 WoS 等少数数据库中学术文献的,而学科领域的起源并不一定局限在正式的学术

出版物中,因此应该进一步探索 RPYS 在融合多数据源、多样化数据格式场景下应用的可能性;此外,RPYS 分析的软件包操作繁琐、可视化效果差,应对其进一步集成和优化,以使其被更广泛地使用。

### 参考文献:

- [1] BORNMAN L, FÉLIX M A, LEYDESDORFF L. Do scientific advancements lean on the shoulders of giants? A bibliometric investigation of the ortega hypothesis[J]. PLOS ONE, 2010, 5(10): 4649-4672.
- [2] COSTAS R, LEEUWEN T, Bordons M. Referencing patterns of individual researchers: do top scientists rely on more extensive information sources? [J]. Journal of the American Society for Information Science & Technology, 2012, 63(12):2433-2450.
- [3] RAAN V. Sleeping beauties in science[J]. Scientometrics, 2004, 59(3):467-472.
- [4] GARFIELD E, PUDOVKIN A, ISTOMIN V. Why do we need algorithmic historiography? [J]. Journal of the American Society for Information Science & Technology, 2003, 54(5):400-412.
- [5] 张卫民, 王磊, 夏国雯. 关于中国林业经济管理学科起源的探讨[J]. 北京林业大学学报:社会科学版, 2013, 12(1):75-80.
- [6] MACHADO C. Historical evolution of the brain death concept: additional remarks [J]. Journal of critical care, 2014, 29(5):867-868.
- [7] MARX W, BORNMAN L, BARTH A. Detecting the historical roots of research fields by reference publication year spectroscopy (RPYS) [C]// VERLAGS F. Proceedings of 14th International Society of Scientometrics and Informetrics Conference. Stolbergasse:Facultasv, 2013:493-506.
- [8] PRICE D. Little science, big science[J]. The quarterly review of biology, 1963, 7(4):443-458.
- [9] RAAN V. On growth, ageing, and fractal differentiation of science [J]. Scientometrics, 2000, 47(2):347-362.
- [10] BARTH A, MARX W, BORNMAN L, et al. On the origins and the historical roots of the Higgs boson research from a bibliometric perspective[J]. European physical journal plus, 2014, 129(6):1-13.
- [11] LEYDESDORFF L, BORNMAN L, MARX W, et al. Referenced publication years spectroscopy applied to iMetrics: scientometrics, journal of informetrics, and a relevant subset of JASIST[J]. Journal of informetrics, 2013, 8(1):162-174.
- [12] MARX W, BORNMAN L. Tracing the origin of a scientific legend by reference publication year spectroscopy (RPYS): the legend of the Darwin finches[J]. Scientometrics, 2014, 99(3):839-844.
- [13] PIDOS J. Correlated evolution of morphology and vocal signal structure in Darwin's finches[J]. Nature, 2001, 409(6817):185-8.

- [14] COMINS J, HUSSEY T. Compressing multiple scales of impact detection by reference publication year spectroscopy[J]. Journal of informetrics, 2015, 9(3):449-454.
- [15] ELANGO B, BORNMAN L, KANNAN G. Detecting the historical roots of tribology research: a bibliometric analysis[J]. Scientometrics, 2016, 107(1):1-9
- [16] SIMONDS K. Health education as social policy[J]. Health education monographs, 1974, 2(1):1-10.
- [17] LAWRENCE L. Book Reviews:principles of health education and health promotion [J]. Health education & behavior, 1995, 22(1):146-148.
- [18] World Health Organization. Health promotion glossary[R]. Geneva: WHO,1998:10.
- [19] US American Medical Association. Health literacy - report of the council on scientific affairs[J]. the Journal of the American Medical Association, 1999, 281(6):552-557.
- [20] US National Library of Medicine. Current bibliographies in medicine [EB/OL]. [2016-05-01]. <http://www.nlm.nih.gov/pubs/cbm/hliteracy.html>.
- [21] 张艳荣. 第二次世界大战对美国高等医学教育的影响[J]. 中华医史杂志, 2005, 35(1):32-36.
- [22] 周志超, 张士靖. 国外信息素养领域研究热点分析——从信息素养到健康信息素养[J]. 情报杂志, 2012(9):147-151.
- [23] DAVIS H W. Aristotle's politics[M]. Oxford: Clarendon Press, 1908.
- [24] BOGARDUS E S. Social distance and its origins[J]. Journal of applied sociology, 1925, 9(216):216-226.
- [25] SHIPLEY W. A Self-Administering Scale for Measuring Intellectual Impairment and Deterioration[J]. Journal of psychology interdisciplinary & Applied, 2010, 9(2):371-377.
- [26] HOULE C O, FLESCHE R. Marks of readable style: a study in adult education[J]. Library Quarterly, 1945(1):85-87.
- [27] FLESCHE R. A new readability yardstick[J]. Journal of applied psychology, 1948, 32(3):221-33.
- [28] TAYLOR W. "Cloze procedure": a new tool for measuring readability[J]. Journalism quarterly, 1953, 30(4):415-433.
- [29] McLAUGHLIN G. SMOG grading - a new readability formula[J]. Journal of reading, 1969, 12(8):639-646.
- [30] BANDURA A. Self-efficacy: toward a unifying theory of behavior change[J]. Psychological review, 1977, 84(2):191-215.
- [31] BANDURA A. The social foundations of thought and action: a social cognitive theory[M]. Englewood Cliffs: Prentice Hall, 1986.
- [32] DAVIS T C, LONG S W, JACKSON R H, et al. Rapid estimate of adult literacy in medicine: a shortened screening instrument[J]. Family medicine, 1993, 25(6):391-395.
- [33] 张士靖, 李信, 刘海通. 国际健康信息素养测评工具概述[J]. 中国健康教育, 2014, 30(10):920-924.
- [34] 周志超, 张悦, 张士靖. 健康信息素养研究领域的演进路径与热点预测[J]. 中国健康教育, 2014, 30(10):904-908.
- [35] NUTBEAM D. The evolving concept of health literacy [J]. Social science & medicine, 2008, 67(12):2072-2078.

作者贡献说明:

李信:撰写和修改论文;  
 陆伟:进行理论指导和思路设计;  
 李旭晖:进行写作指导,参与论文修改。

A Novel Approach for Detecting the Historical Root of a Research Area:  
 Reference Publication Years Spectroscopy (RPYS)

Li Xin<sup>1,2</sup> Lu Wei<sup>1,2</sup> Li Xuhui<sup>1,2</sup>

<sup>1</sup> School of Information Management, Wuhan University, Wuhan 430072

<sup>2</sup> Information Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan 430072

**Abstract:** [Purpose/significance] The study of the historical root of a research area, and finding out the important origin papers is of great significance for the construction and research in this area. [Method/process] For this purpose, we introduced the connotation and philosophy of reference publication years spectroscopy (RPYS), and inducted RPYS analysis to four steps. On that basis, we take health literacy research as an example and use RPYS to explore the historical root of health literacy. [Result/conclusion] The result shows that RPYS can identify the historical root accurately in health literacy. At the meantime, we find that RPYS can be used to explore the hot research topic and research trend of a research area.

**Keywords:** RPYS subject historical root reference publication year spectroscopy health literacy