



开放科学
(资源服务)
标识码
(OSID)

时间加权 z 指数探测热点关键词研究

邱小敏¹ 奉国和¹ 刘宇东²

1. 华南师范大学经济与管理学院 广州 510006;
2. 华南师范大学计算机学院 广州 510631

摘要: [目的/意义] 构建关键词热度 h 值模型, 探究近年图书情报领域热点与发展趋势。[方法/过程] 基于学者 Prathp 的 z 指数模型, 对关键词词频和文章被引值进行赋权, 引入前人时间加权思想, 多维角度归一化处理, 计算热度值与年度排名, 获得趋势。[结果/结论] 比较加权前后绝对词频、z 指数和关键词热度 h 值, 排名结果差异明显; 热度 h 值模型可提升热点关键词排名, 拉低非热点关键词排名, 排名情况验证 h 值的有效性, 效果优良。

关键词: z 指数; 绝对词频; 时间加权被引值; 研究热点; 发展趋势

中图分类号: G350

Research on Hot Keywords Detected by Time Weighted z-index

QIU Xiaomin¹ FENG Guohe¹ LIU Yudong²

1. School of Economics & Management, South China Normal University, Guangzhou 510006, China;
2. School of Computer Science, South China Normal University, Guangzhou 510631, China

Abstract: [Objective/ Significance] To construct the h-value model and explore the hot spots and development trends in the field of library and information in recent years. [Methods/Process] Based on the z-index model of Prathp, this paper weights the keyword frequency and the quoted value of the article, introduces the time weighting thought of predecessors, normalizes the data from multiple angles, calculates the heat value and annual ranking, and obtains the trend. [Results /Conclusions] The absolute word frequency, z-index and h-value of keyword popularity were compared before and after weighting, and the ranking results were significantly different. The h-value can improve the ranking of hot keywords, and pull down the ranking of non-hot keywords. The ranking situation verifies the validity of the h-value, and the effect is good.

Keywords: z-index; absolute word frequency; time weighted citation value; research hotspot; development trend

基金项目 广州市科技计划基础与应用基础研究专题面上项目“突发词探测理论、方法与应用研究题目”(202002030384)。

作者简介 邱小敏(1995-), 硕士研究生, 研究方向为数据挖掘、信息计量学; 奉国和(1971-), 博士, 教授, 研究方向为数据挖掘、信息计量学, E-mail: ghfeng@163.com; 刘宇东(1974-), 硕士, 讲师, 研究方向为自然语言处理、机器学习。

引言

从浩瀚的科技文献中准确把握研究热点是科研人员、科研管理者把握前沿和科技创新的关键任务之一,也是制定科技发展战略时面临的一大问题。我们国家非常重视科学前沿的科研部署,习近平主席在多种场合提及科技创新的重要性,鼓励广大科技工作者多创新,多进行前沿问题探索。2016年8月8日,国务院印发了《“十三五”国家科技创新规划》的通知,明确了“十三五”时期科技创新的总体思路、发展目标、主要任务和重大举措。其中,发展重点为把握世界科技前沿发展态势,超前规划基础前沿领域,全面增强自主创新能力。此外,国际科技论文被引次数的世界排名也作为“十三五”科技创新的12个主要指标之一^[1]。2019年11月19日,中国科学技术信息研究所发布中国科技论文统计结果,2009年至2019年10月,中国科技人员共发表国际论文260.64万篇,排在世界第2位,数量比2018年统计时增加了14.7%;论文共被引用2845.23万次,增加了25.2%,排在世界第2位^[2]。科技论文被引情况已成为领域创新与研究前沿的重要指标,被广泛用于学者、文献、期刊、专利等领域的评估与预测,对把握学科发展动态有积极作用^[3]。

研究热点一直都是众学者重点关注的问题,随着科技发展,知识输出体量呈指数增长,探究领域发展状况与当前研究热点能更好指明未来研究方向。每年大量研究成果通过期刊文献的形式发表,如何从海量文献中挖掘热点是各学科研究焦点之一,目前相关研究指标包括:

关键词词频、论文下载量、论文被引量 and 期刊均篇被引量等。基于这种启发,本研究拟依托上述基本指标,基于 z 指数,开展热点关键词探测研究。

1 相关研究

学术期刊作为研究人员创新性科研成果的重要载体,其评价方法的科学性、准确性、合理性受到广大研究人员的关注^[4]。2005年,Hirsh^[5]提出 h 指数,用于学术价值评价,即某学者至少有 h 篇论文的被引值至少为 h 次,但 h 指数不适用于跨学科领域比较与低质量论文评价^[6]。针对 h 指数的缺陷,Prathap 在 h 指数的基础上加入总发文量与均篇被引量,提出 p 指数^[7],但其无法体现高被引论文价值^[8]与文章被引时间因素^[9]。2014年,Prathap 在 p 指数的基础上提出 z 指数,增加了一致性指标,用于协调高被引文章与文章篇数的关系^[10]。z 指数的提出创新了科学评价体系,为计量领域提供了一个重要研究指标与研究方向。

国外对 z 指数的研究主要集中于 Prathap 对 z 指标的探索与完善,其进一步解释了 z 指数,分析了 z 指数与 h 指数之间的关系^[11]; Prathap 认为 z 指数是一种 3D 评价指标模型,属于高级评价指标^[12]。在实际应用中,Prathap 使用 z 指数评价印度科研状况^[13],太阳能电池领域 (polymer solar cells) 研究进展,从国家、机构、学者、期刊等角度论证了 z 指数科学评价的适用性^[13]。国内学者通过比较 h 指数、p 指数与 z 指数,探究 z 指数在学者^[14]、期刊^[5]、科研机构^[15]中的评价效果,研究结果显示 z 指数是一

个灵敏度高、区分度高、评价全面的综合性评价指标,具备一定的通用性^[16]。同时,z指数也存在缺陷:在期刊评价中,z指数无法对施引文献评价^[5];在科研机构评价中,长尾效应严重影响z指数的评价效果,以及忽视了专利等科研指标^[16];对于文章低被引集中度的缺陷,有学者对z指数进行了相关改进^[16]。此外,z指数不涉及时间因素,但文章被引量随时间不断增加,被引量的变化率也随时间发生改变。因此,z指数无法区分与评估被引量相同但发表时间不同,被引变化率上升或下降的文章影响力^[17]。

针对上述问题,本文在z指数的基础上引入俞立平^[18]教授被引峰值的时间思想与李长玲^[1]教授的时间因子公式,根据本文对关键词热度的定义,在单一被引值的基础上,加入关键词词频要素,通过模型计算词频与加权被引值在总年份与出版年份的占比值^[19],最终获取关键词热度。其中,关键词词频体现重要性与关注度,是常用的文献计量指标;被引次数体现文章之间的传播性与关联性,是追踪文章热点的分析指标之一。

2 时间加权 z 指数 (z_{tw})

2.1 z指数基本理论

z指数由 Prathap 提出,通过改进 h 指数和 p 指数得到,是一种综合评价指标。本质是将数量、质量、一致性程度三个指标等权重相乘,故也称 quantity(数量)-quality(质量)-consistence(一致性) 3D 效能评价指标^[11]。表达公式如下:

$$z = \left[c \times \frac{c}{n} \times \frac{\frac{c^2}{n}}{\sum_{k=1}^n c_k^2} \right]^{\frac{1}{3}} \quad (1)$$

其中, c 为总被引频次,表示研究对象被引规模; n 为文章总篇数; $\frac{c}{n}$ 为文章均篇被引量,反映研究对象在单篇论文层面的平均影响力,也称“质量”指标; $\frac{c^2}{\sum_{k=1}^n c_k^2}$ 为分布一致性指标, c_k 表示第 k 篇文章的被引频次 ($k=1,2,\dots,n$),是平衡引用极值及分布频次对结果影响的调节因子,从“数量与质量”一致性层面反映研究对象被引分布的均匀性水平。

2.2 影响因素的相关计算

基于 z 指数,引入总年份占比和同年占比的加权思想^[20],加入时间要素改进 z 指数的数量、质量指标,分别获得关键词词频与时间加权被引值的 z_{tw} 指数值,最终得到关键词热度排名与等级趋势变化情况。

(1) 时间加权被引函数

本文根据李长玲^[1]教授的时间因子公式,对不同年份文章的被引值进行时间加权。被引行为体现学者对发表内容的关注度,本文将时间加权被引值的 z_{tw} 值称为热度 h_c 。关键词词频从作者第一视角反映文章主体思想与研究内容,本文将关键词词频的 z_{tw} 值称为热度 h_f 。由于出版文章的关键词词频与被引值在时间维度上存在差异,出版当年文章篇数与关键词词频已为确定值,而文章的被引值在出版后动态递增。因此,本文引入时间加权被引思想,将出版年份与被引年份的差值和被引年份与当前分析年份的差值赋予适当

权重, 公式如下:

$$s = \sum_{y=0}^Y (Y-y) = \frac{(1+Y-y)(Y-y)}{2} \quad y \leq Y \quad (2)$$

$$v = \sum_{x=0}^Y \left(\frac{x+1}{s} \right) \times c_x \quad x \leq Y \quad (3)$$

其中, Y 为当前年份, y 为出版年份, x 为被引年份, 出版后的第一年被引年份 x 为 0, c_x 为第 x 年的被引频次。表 1 为 2014 年出版的文献在 2014—2019 年各年度发生的被引值。

表 1 2014-2019 年关键词 i 的被引情况

被引年份	2014	2015	2016	2017	2018	2019
被引值 (c_x)	196	207	180	315	193	66

根据公式 (2)、(3) 得:

$$s = \sum_{y=0}^Y (Y-y) = \sum_{y=0}^5 (5-y) = 5+4+3+2+1+0=15$$

$$v = \sum_{x=0}^Y \left(\frac{x+1}{s} \right) \times c_x = \frac{1}{15} \times 196 + \frac{2}{15} \times 207 + \frac{3}{15} \times 180 + \frac{4}{15} \times 315 + \frac{5}{15} \times 193 + \frac{6}{15} \times 66 = 251.4$$

其中, 2014 年出版的文章在 2014 年的时间被引权重为 $\frac{1}{15}$, 在 2019 年的时间被引权重为 $\frac{6}{15}$, 251.4 为 2014 年出版的文献在 6 年间的加权被引值。

(2) “时间加权被引—出版年份” 矩阵

通过时间加权被引函数的计算, 得到关键词不同出版年份的加权被引值, 进而获取“时间加权被引值—出版年份” 矩阵。如矩阵 (1) 所示, v 表示时间加权被引值, n 为关键词, m 为出版年份, 以 $v_{2,3}$ 为例, 表示第二个关键词在

第三个时间窗口(年份)出版的时间加权被引值。

$$M(v) = \begin{pmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,m} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n,1} & v_{n,1} & \cdots & v_{n,m} \end{pmatrix} \quad \text{矩阵 (1)}$$

(3) “关键词词频—出版年份” 矩阵

“关键词词频—出版年份” 矩阵如矩阵 (2) 所示, 矩阵 (2) 的数值由各年度出版文章的关键词绝对词频构成。被引值通过引用行为动态递增, 而当年出版文章的关键词词频不再发生改变, 所以, 两种数值产生的影响存在差异性, 但相同之处在于都以出版年份为时间基准进行计算。例如 2015 年出版的文献, 其关键词词频反映 2015 年的研究内容, 时间加权被引值则通过时间加权被引函数的计算, 将不同年份的被引值映射至出版年份, 获得 2015 年出版文章的综合被引强度。 f 表示绝对词频, n 为关键词, m 为出版年份, 以 $f_{2,3}$ 为例, 表示第二个关键词在第三个时间窗口(年份)出版的绝对词频。

$$M(f) = \begin{pmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,m} \\ f_{2,1} & f_{2,2} & \cdots & f_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n,1} & f_{n,1} & \cdots & f_{n,m} \end{pmatrix} \quad \text{矩阵 (2)}$$

2.3 z_{nw} 指数设计

本文关键词热度围绕关键词词频与被引值, 在 z 指数数量、质量、一致性程度三个指标的基础上加入时间因素, 计算 z_{nw} 指数值, 最终得到关键词热度。

(1) $z_{nw_quality}$ 质量指标

在 z 指数中, 总被引值表示数量指标。相比之下, 矩阵 (1)、(2) 表示被引要素 v 与词频要素 f 在 z_{nw} 指数下的数量指标。此外, z

指数中的质量指标表示均篇被引,在此基础上,本文加入时间因素,通过时间维度反映研究对象在不同时期的平均影响力,从总年份占比和同年占比两个角度进行加权,得到 z_{tw} 指数的质量指标。以时间加权被引值为例, z_{tw} 指数的质量指标如下所示:

$$z_{tw_quality}(v) = v_{i,y} \times \frac{1}{\sum_{i=1}^n v_{i,y}} \times v_{i,y} \times \frac{1}{\sum_{y=0}^m v_{i,y}} \quad (4)$$

其中, i 表示第 i 个关键词, n 为关键词总数, $i \leq n$; y 表示关键词对应的出版年份, m 为出版总年份, $y \leq m$ 。

(2) $z_{tw_consistence}$ 一致性指标

z_{tw} 指数遵循 z 指数中的一致性原则,在数量、质量指标的基础上,通过一致性指标平衡高低被引值与文章总篇数关系,被引 z_{tw} 指数的一致性指标如公式(5)所示:

$$z_{tw_consistence}(v) = \frac{v_{i,y} \times z_{tw_quality}(v)}{\sum_{i=1}^n v_{i,y}^2} \quad (5)$$

其中, $\sum_{i=1}^n v_{i,y}^2$ 为文章对应关键词在 y 年的被引平方和。

(3) 热度 h_v 与热度 h_f

根据本文对热度的定义和上述相关指标的描述,构建热度 h_v 与热度 h_f 。以热度 h_v 为例:

$$z(v) = \left[v_{i,y} \times z_{tw_quality}(v) \times z_{tw_consistence}(v) \right]^{\frac{1}{3}} \quad (6)$$

其中, $z(v)$ 为时间加权被引值的 z_{tw} 指数值,即热度 h_v ; $z(f)$ 为关键词绝对词频的 z_{tw} 指数值,即热度 h_f 。

2.4 关键词热度与趋势变化

(1) 关键词热度值

由 $z(v)$ 与 $z(f)$ 得到关键词热度 h 值,通过 h 值的变化情况,探究热点关键词变化趋势。由于关键词词频与被引值为不同分析角度,对 $z(v)$ 、 $z(f)$ 标准化处理,关键词热度公式如下所示:

$$h = \frac{z(v) - z_{\min}(v)}{z_{\max}(v) - z_{\min}(v)} + \frac{z(f) - z_{\min}(f)}{z_{\max}(f) - z_{\min}(f)} \quad (7)$$

(2) 关键词热度与趋势判断

对关键词每年热度 h 值降序排名并划分四个等级,排名前25%的关键词定义为热点词,25%~50%的为次热点词,50%~75%的为弱热点词,75%~100%的为非热点词。比较关键词各年度等级变化,得到关键词热度趋势。以2014、2015年部分关键词为例,如表2所示。

表2以2015年排名前50的热点词为分析对象。比较2015年与2014年的排名情况,创客空间、智库、互联网+与数字阅读从非热点词上升至热点词,上升趋势明显;公共文化服务、信息安全、资源建设、资源整合与信息传播从弱热点词上升至热点词,具有一定的上升趋势;全民阅读、知识图谱、虚拟社区、数据挖掘、信息生态与服务创新从次热点上升至热点词;此外,高校图书馆、图书馆、大数据、微信与微博等,排名变化小,且都位于热点词范围。上述实例,描述了热点关键词排名情况,对关键词分级能更好的比较趋势变化。

表 2 2014—2015 年关键词 h 值排名 (部分)

序	关键词	2014年	2015年	序	关键词	2014年	2015年	序	关键词	2014年	2015年
1	图书馆	1	1	18	社会网络分析	16	18	35	情报学	22	35
2	高校图书馆	2	2	19	知识服务	26	19	36	资源建设	114	36
3	大数据	4	3	20	云计算	23	20	37	资源整合	138	37
4	公共图书馆	3	4	21	数字资源	32	21	38	信息传播	147	38
5	阅读推广	7	5	22	全民阅读	81	22	39	互联网+	193	39
6	微信	6	6	23	学科服务	36	23	40	竞争情报	27	40
7	网络舆情	5	7	24	智库	163	24	41	服务模式	20	41
8	微博	13	8	25	知识图谱	53	25	42	本体	44	42
9	mooc	12	9	26	可视化	38	26	43	虚拟社区	76	43
10	图书馆服务	10	10	27	图书馆学	21	27	44	科学数据	30	44
11	数字图书馆	8	11	28	信息安全	127	28	45	数字阅读	168	45
12	信息素养	14	12	29	文献计量	15	29	46	文献综述	24	46
13	信息服务	11	13	30	开放获取	40	30	47	数据挖掘	62	47
14	创客空间	151	14	31	影响因素	25	31	48	突发事件	45	48
15	公共文化服务	108	15	32	机构知识库	33	32	49	信息生态	52	49
16	美国	18	16	33	关联数据	43	33	50	服务创新	57	50
17	移动图书馆	9	17	34	大学图书馆	41	34				

3 实证分析

利用 Python 编程工具, 分析知网引文检索数据库与文献检索数据库采集的相关数据。

3.1 数据来源与预处理

数据来源于 2014—2019 年图书情报领域 18 种核心期刊文献, 它们分别为: 中国图书馆学报、图书情报知识、图书与情报、大学图书馆学报、情报理论与实践、图书情报工作、情报科学、情报杂志、情报资料工作、图书馆论坛、图书馆工作与研究、图书馆建设、数据分析与

知识发现、情报学报、国家图书馆学刊、图书馆、图书馆学研究、图书馆杂志。人为去除会议、通知、征稿等信息, 共获取有效被引期刊文献 20091 篇, 有效出版期刊文献 24649 篇。以 24649 篇出版文献的关键词作为分析对象, 进行数据处理。

3.2 选定关键词

本文以 2014—2019 年总词频排名前 200 的关键词作为研究对象, 计算热度 h 值, 对各年度关键词排名划分等级并研究发展趋势。排名前 50 的关键词如表 3 所示。

表 3 Top50 的关键词

序	关键词	词频	序	关键词	词频	序	关键词	词频	序	关键词	词频
1	图书馆	1848	14	图书馆服务	280	27	科学数据	203	40	信息生态	160
2	高校图书馆	1670	15	微信	266	28	竞争情报	200	41	共词分析	155
3	公共图书馆	1167	16	知识图谱	265	29	数字资源	193	42	数据挖掘	152
4	大数据	765	17	图书馆学	261	30	mooc	193	43	全民阅读	150
5	阅读推广	616	18	情报学	259	31	关联数据	191	44	突发事件	148
6	美国	399	19	知识服务	240	32	本体	189	45	图书馆员	147
7	网络舆情	375	20	智库	222	33	开放获取	186	46	云计算	147
8	数字图书馆	361	21	社会网络分析	222	34	评价指标	173	47	情报分析	147
9	影响因素	335	22	学科服务	222	35	数字人文	172	48	知识管理	146
10	信息服务	306	23	移动图书馆	221	36	机构知识库	172	49	数据管理	143
11	文献计量	297	24	公共文化服务	213	37	知识共享	165	50	引文分析	143
12	微博	286	25	大学图书馆	212	38	信息行为	162			
13	信息素养	285	26	可视化	205	39	智慧图书馆	160			

3.3 关键词热度与排名分析

3.3.1 关键词热度与排名

根据公式(7)得到排名前200的关键词热度 h 值,如表4所示。

基于上述结果,将关键词热度 h 值进行年度排名并划分等级。分析2019年排名前50的热点关键词,得到2014—2019年排名变化情况,部分数据如表5所示。

表 4 关键词热度 h 值(保留三位小数)

序	关键词	2014年	2015年	2016年	2017年	2018年	2019年
1	图书馆	2.000	2.000	1.608	1.825	1.306	1.046
2	高校图书馆	1.534	1.434	1.969	2.000	2.000	2.000
3	公共图书馆	0.780	0.910	0.728	0.890	0.896	0.938
4	大数据	0.538	0.960	1.327	1.040	0.721	0.951
5	阅读推广	0.436	0.643	0.540	0.458	0.706	0.917
6	美国	0.196	0.149	0.185	0.285	0.185	0.389
7	网络舆情	0.495	0.269	0.397	0.348	0.238	0.228
8	数字图书馆	0.377	0.229	0.166	0.263	0.132	0.185
9	影响因素	0.152	0.100	0.211	0.461	0.273	0.252
10	信息服务	0.345	0.199	0.163	0.114	0.126	0.037
11	文献计量	0.216	0.104	0.211	0.149	0.083	0.064
12	微博	0.341	0.268	0.160	0.099	0.132	0.049
13	信息素养	0.260	0.203	0.294	0.162	0.066	0.100
14	图书馆服务	0.349	0.243	0.143	0.197	0.064	0.058
15	微信	0.454	0.406	0.410	0.232	0.153	0.034

表 5 关键词热度 h 值排名

序	关键词	2014年	2015年	2016年	2017年	2018年	2019年
1	高校图书馆	2	2	1	1	1	1
2	图书馆	1	1	2	2	2	2
3	大数据	4	3	3	3	4	3
4	公共图书馆	3	4	4	4	3	4
5	用户画像	198	199	198	186	9	5
6	阅读推广	7	5	5	6	5	6
7	人工智能	194	196	199	9	10	7
8	智慧图书馆	120	67	89	13	7	8
9	美国	18	16	13	8	16	9
10	数字人文	169	194	45	21	6	10
11	公共文化服务	108	15	33	39	21	11
12	智慧服务	174	130	195	69	14	12
13	知识图谱	53	25	29	17	19	13
14	影响因素	25	31	10	5	11	14
15	网络舆情	5	7	7	7	12	15

从表 5 可知, 高校图书馆、图书馆、大数据、公共图书馆等关键词排名位于前列且变化小, 属于热点关键词; 相比之下, 用户画像在 2014—2017 年间属于非热点词, 2018 年排名迅速上升, 成为热点词; 此外, 2017 年, 人工智能排名上升明显, 数字人文在 2016 年和 2017 年间排名变化也十分显著; 公共文化服务和智慧服务在国家大力发展数字基础建设的背景下成为研究热点, 以上为排名显著上升的关键词, 下文以折线图对

前 200 个关键词的变化趋势进行分析。

3.3.2 趋势分析

根据表 5 数据与分级标准, 本文将 2019 年热点关键词 (排名前 50) 作为研究对象, 探究其过去 6 年排名变化趋势, 对关键词在 2014 年所属四个等级分别进行讨论。横坐标为年份, 纵坐标为排名, 排名数值越小, 热度越大。变化情况如图 1 所示。

(1) 热点关键词上升为热点关键词

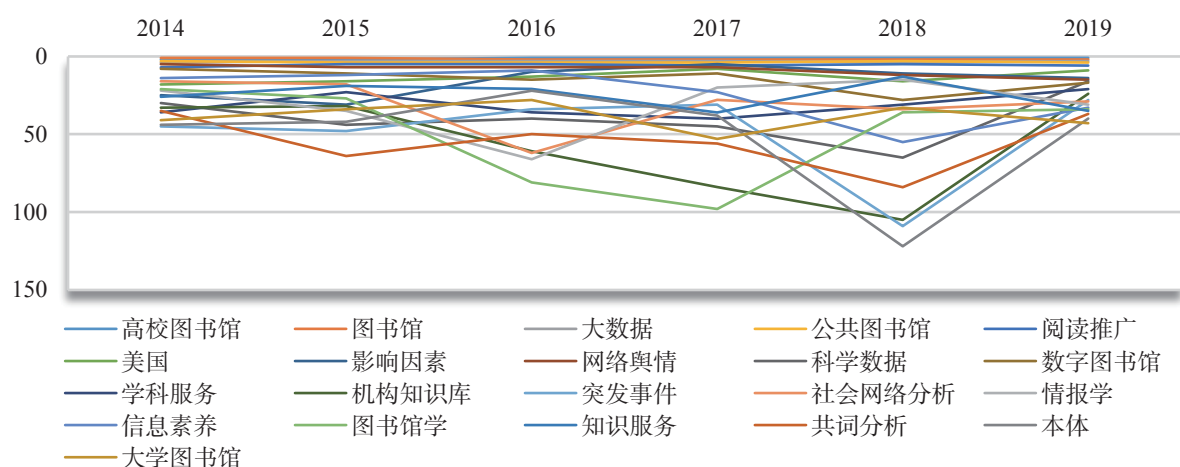


图 1 2019 年热点关键词趋势变化情况 (1)

图1所示,2014年热点关键词在2019年仍属于热点关键词,排名变化稳定。部分关键词有明显的学科特点,如:高校图书馆、图书馆、公共图书馆、图书馆学、情报学、学科服务等。除学科关键词外,与文献计量学相关的研究方法排名也较为稳定,如共词分析法、社会网络分析法等。此外,在互联网背景下,大数据、突发事件、网络舆情等成为该领域的研究热点词。

(2) 次热点关键词上升至热点关键词

图2所示,2014年次热点关键词在六年间上升为热点关键词。与图1相比,图2关键词涉及研究方法,除文本挖掘波动幅度较大,其余关键词稳中有进。近年,知识图谱、内容分析、文本挖掘、评价指标等研究方法广泛运用于不同的学术领域,大数据、自然语言处理、机器学习作为社会热门话题,直接或间接提高科学计量方法在不同领域的使用。

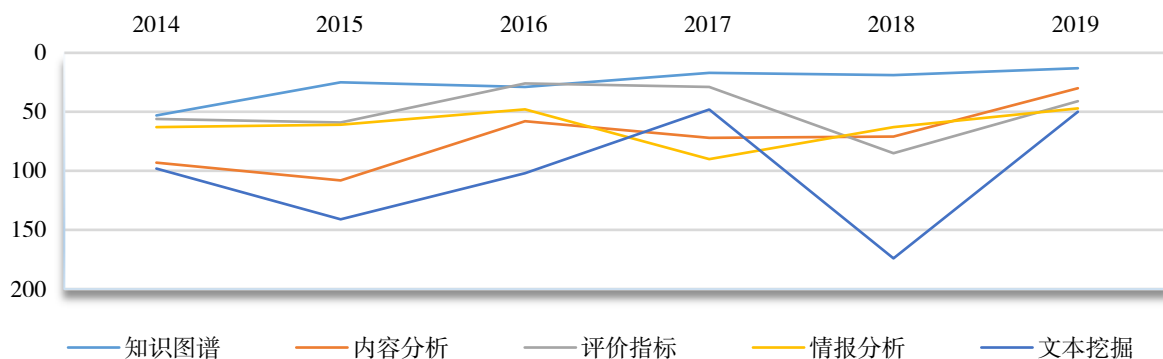


图2 2019年热点关键词趋势变化情况(2)

(3) 弱热点关键词上升至热点关键词

图3所示,2014年弱热点关键词在2019年上升为热点关键词,排名变化显著,部分关键词在2016年上升趋势明显,对领域热点的研究具有参考价值。2016—2017年,AlphaGo与围棋冠军的人机对战将机器学习拉入大众视野,并获得学术界的关注。2017年底,国家大数据战略第二次会议提出加快完善数字基础设施,推进数据资源整合和开放数据共享,各大城市、区域政府开启与互联网企业合作,推出了一系列惠民、便民、利民的数据资源开放共享平台,开启了大数据服务的探索。如广东省于2017年

联合腾讯、移动、电信、联通推出数字广东,推进珠三角地区信息化发展。

(4) 非热点关键词上升至热点关键词

图4所示,2014年非热点关键词在2019年上升为热点词,上升迅速,对该领域研究热点有积极参考价值。2015年,国务院印发《国务院关于积极推进“互联网+”行动的指导意见》,同年,也成为中国特色新型智库建设元年,智库不仅上升为国家战略,也走进了公众视线,其关注度延续至今。人工智能、深度学习同机器学习,受到2017年AlphaGo的影响,关注度上升。用户画像最早由Alan Cooper提出,

于 2017 年得到国内学者关注,作为研究用户特点的工具,广泛用于电子商务、图书馆、医疗健康等领域。在线评价作为用户对产品体验的一部分,伴随电商、社交媒体的发展和企业

对用户画像的关注,成为 2019 年图书情报领域热点关键词。2009 年,数字人文引入学界,近年随着相关技术的发展,数字人文逐渐得到关注。

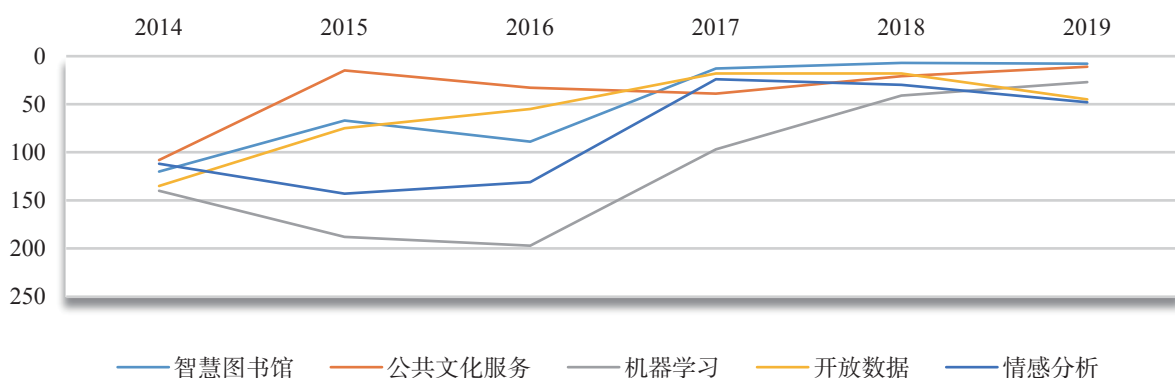


图 3 2019 年热点关键词趋势变化情况 (3)

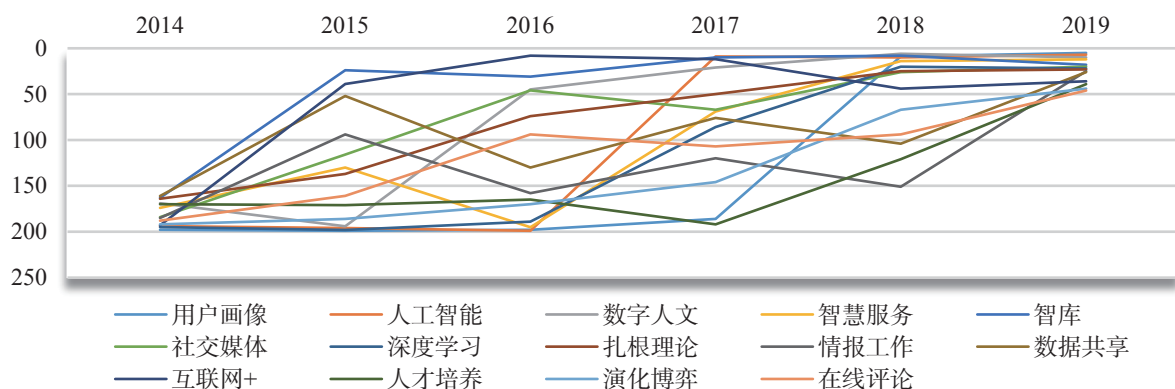


图 4 2019 年热点关键词趋势变化情况 (4)

上述四类关键词排名特点如下：学科类关键词排名高、排名变化小、对热点判断参考价值小；经典文献计量方法排名稳中有升；基于计算机与互联网技术高速发展的背景，计算方法、研究理论快速成为热门焦点,受到广泛关注,具有较高的参考意义。

3.4 z_{tw} 指数效果对比分析

将 2017—2019 年作为一个时间单位,以热

度 h 值为基准降序排列,比较原始 z 指数值与绝对词频排名,检验 z_{tw} 指数评价效果。图 5 中,横坐标为关键词,纵坐标为排名,排名数值越小,越为热点关键词。

(1) 热点关键词比较

图 5 为排名前 50 的热点关键词,可以看出,热度 h 值排名前列的关键词三种排名差距小。其中,高校图书馆等学科热点词位于前列,互联网、大数据、智库、网络舆情、知识图谱、可视

化、微信、微博、社交媒体、人工智能等关键词都为近年研究热点词。此外,三种方法排名差异

较大的有开放获取、健康信息与深度学习等,深度学习在 z 指数计算下,排名仅为171。

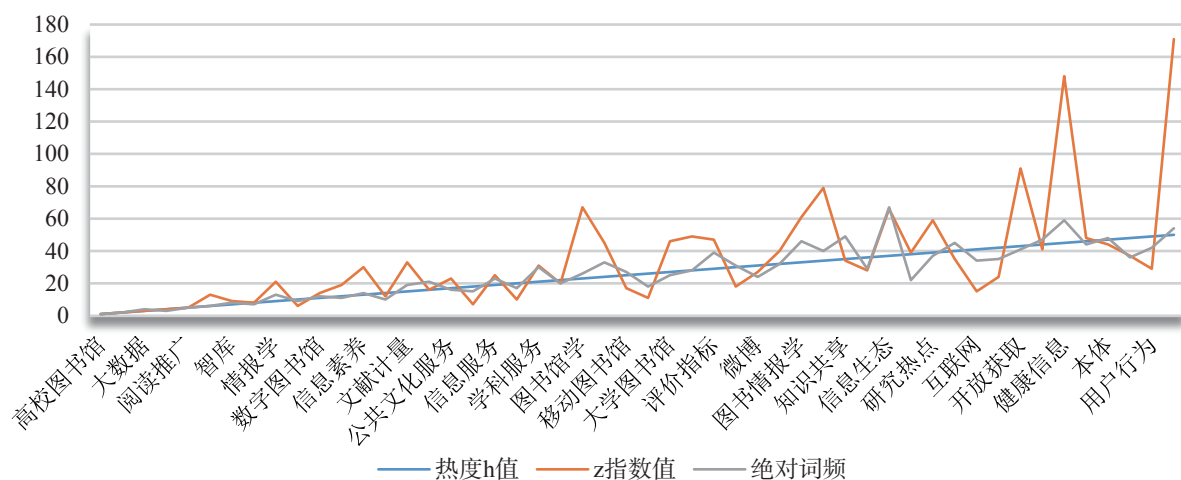


图5 热点关键词比较(部分显示)

(2) 次热点关键词比较

图6为排名50~100的次热点关键词,比较三种排名结果,差异特征明显。热度 h 值排名显著高于原始 z 指数值与绝对词频,排名前列的有:机器学习、开放政府数据、数据共享、

开发存取、用户画像、区块链、主题模型等,都为近年热点词。通过比较不同方法的排名结果,可以看出热度 h 值能够提升部分关键词排名,且结果较为符合当前领域状况,也符合上述趋势分析的相关背景。

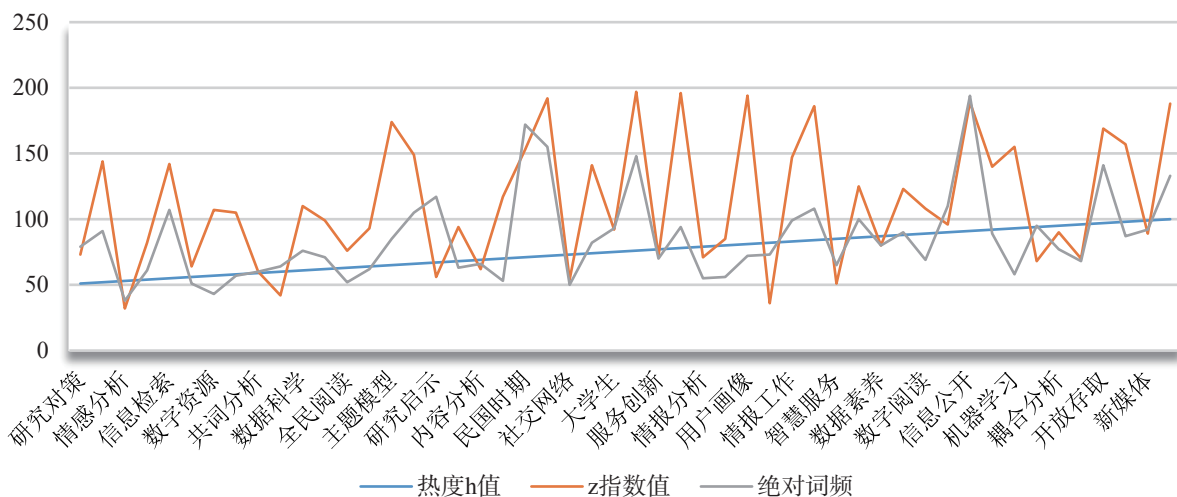


图6 次热点关键词比较(部分显示)

(3) 弱热点关键词比较

图 7 为排名 100~150 的弱热点关键词，可以看出，热度 h 值与 z 指数值和绝对词频排名存在正负排名差，热度 h 值排名高于其他两种方法的关键词有主题识别、信息质量、文本挖掘、政府数据、知识挖掘等；热度 h 值排名低于其他两个方法的关键词有主题演化、信息安全、在线评论、知识管理、CiteSpace、期刊评价等。弱热点关键词排名特点在于词频与被引相互影响，三种方法的排名高低差值分布均匀。

(4) 非热点关键词比较

图 8 为排名 150~200 的非热点关键词，热度 h 值排名普遍低于 z 指数值与绝对词频。其中，热度 h 值排名高于其他两种方法的关键词有 ischool、人才培养、政府数据开放等；热度 h 值排名低于其他两个方法的关键词有馆藏建设、情报研究、总分馆制、知识元、电子书、知识组织等。热度 h 值排名较高的关键词与大环境下的研究热点相关。综上，热度 h 值排名能拉低非热点词，也较为符合传统情报领域的研究现状。

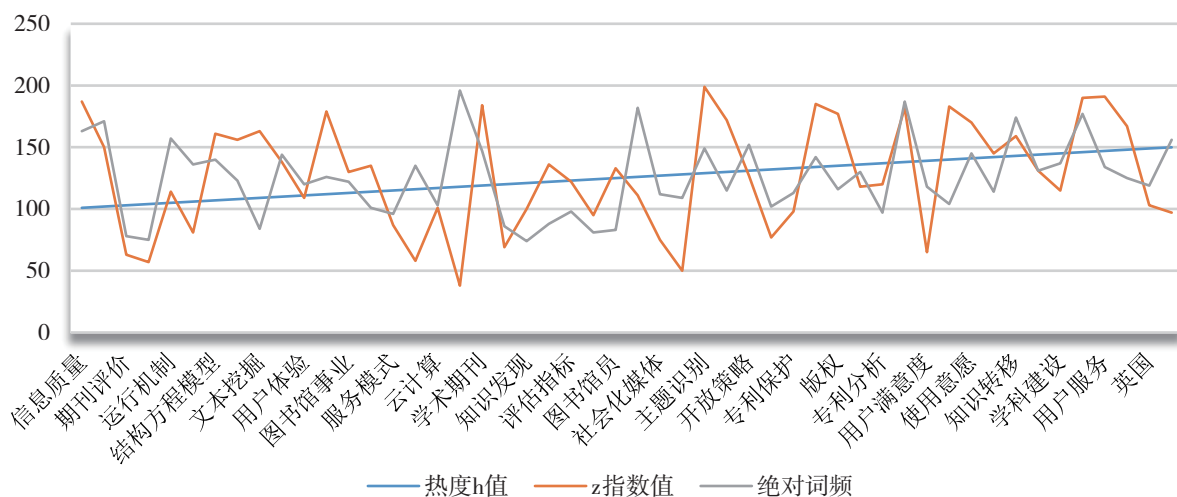


图 7 弱热点关键词比较（部分显示）

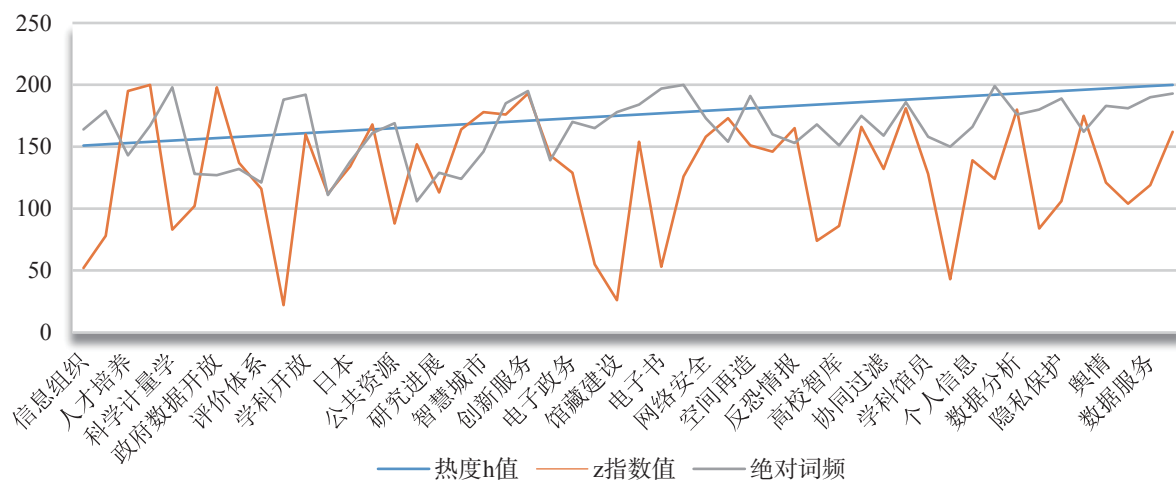


图 8 非热点关键词比较（部分显示）

通过三种不同方法的比较,热度 h 值排名较高的关键词为近年学者重点关注方向,除学科关键词外,次热点关键词能很好的体现当前互联网背景下热门研究方法与话题。在分析的过程中,次热点关键词能很好的提升排名,对于非热点关键词能较好的拉低排名,凸显分析结果,使关键词热点评价更加合理、更加有效。

4 结论

本文设计 z_{rw} 指数探究图书情报领域研究热点与变化趋势。根据原始 z 指数在评价研究对象数量、质量与一致性指标即综合性优势,基于俞立平、李长玲教授的时间加权被引公式与奉国和教授的加权时间思想,对已出版文章的被引动态性赋予权重。根据新模型计算关键词热度值,通过排名划分等级的方式得到热度变化趋势,排名变化大且位于前列的关键词有大数据、用户画像、人工智能、深度学习、机器学习等。通过比较热度 h 值与原始 z 指数和绝对词频排名差异,检验热度 h 值评价效果,结果显示:热点关键词排名差异小,次热点与非热点关键词差异特征明显;在次热点部分,热度 h 值排名高于其他两种方法;在非热点部分,热度 h 值排名低于其他两种方法。可以得出, z_{rw} 指数能够提升次热点关键词排名,拉低非热点关键词排名,得到的关键词也较为符合当前社会发展与领域研究状况,在热点探究中结果更加有效与合理。

本文将关键词词频与被引值作为热度分析要素,对图书情报领域关键词评价,比较三种排名结果检验 z_{rw} 指数效果,不同等级排名差异

特征显著。对于排名差异,可深入探究,指标深度即数量、质量与一致性,指标广度即关键词词频与被引值,分析各指标对热度影响程度,从而挖掘不同等级排名差异因素。由于不同领域不同学科的文章存在差异,可将本研究方法扩充至其他领域进行数据分析,比较结果,进而判断模型适用性。

参考文献

- [1] 中华人民共和国国务院. 国务院关于印发“十三五”国家科技创新规划的通知[R]. 中华人民共和国国务院公报, 2016(24): 6-53.
- [2] 光明日报. 2019 中国科技论文统计结果发布: 从求数量到重质量评价指标变化显著[EB/OL]. (2019-11-19) [2019-11-20]. http://www.gov.cn/xinwen/2019-11/20/content_5453698.htm.
- [3] 鲍玉芳, 马建霞. 科学论文被引频次预测的现状分析与研究[J]. 情报杂志, 2015, 34(5): 66-71.
- [4] 何晓庆, 王圣洁, 胡琳. z 指数在期刊评价实践中的应用研究[J]. 中国科技期刊研究, 2018, 29(5):509-514.
- [5] Hirsch J E. An index to quantify an individual's scientific research output[J]. PNAS, 2005, 102(46):16569-16572.
- [6] Bornmann L, Daniel H D. Does the h-index for ranking of scientists really work?[J]. Scientometrics, 2005, 65(3):391-392.
- [7] Prathap G. The iCE approach for journal evaluation[J]. Scientometrics, 2010, 85(2):561-565.
- [8] Egghe L. Theory and practise of the g-index[J]. Scientometrics, 2006, 69(1):131-152.
- [9] 金碧辉, Rousseau R. R 指数、AR 指数: h 指数功能扩展的补充指标[J]. 科学观察, 2007, 2(3):1-8.
- [10] Prathap G. A three-class, three-dimensional bibliometric performance indicator[J]. Journal of the Association for Information Science and Technology, 2014, 65(7):1506-1508.
- [11] Prathap G. Letter to the editor: Measures for bibliometric size, impact, and concentration[J]. Journal of

- the Association for Information Science and Technology, 2015, 66(8):1740-1741.
- [12] Prathap G. A three-dimensional bibliometric evaluation of recent research in India[J]. Scientometrics, 2017, 110(3):1085-1097.
- [13] Prathap G. A three-dimensional bibliometric evaluation of research in polymer solar cells[J]. Scientometrics, 2014, 101(1): 889-898.
- [14] 何晓庆, 王圣洁, 胡琳. z 指数在学者学术影响力评价中的应用 [J]. 情报理论与实践, 2018, 41(5):50-54.
- [15] 何晓庆, 王圣洁, 胡琳. 基于 z 指数的科研机构评价的有效性实证研究 [J]. 现代情报, 2018, 38(5):82-86.
- [16] 俞立平, 王作功. z 指数评价学术期刊的适用性及其改进研究 [J]. 情报学报, 2018, 37(11):1132-1139.
- [17] 许新军. p 指数在期刊评价中的应用 [J]. 情报学报, 2015, 34(12):1246-1251.
- [18] 俞立平, 郭强华. 被引峰值悖论及影响因子的修正研究: 时间影响因子 [J]. 情报理论与实践, 2019, 42(7):54-58.
- [19] 奉国和, 孔泳欣, 肖洁琼. 基于加权关键词的领域热点与趋势分析新方法 [J]. 图书情报工作, 2018, 62(18):102-109.