

测量学术贡献的关键词分析法探析

黄 晨 赵 星 卞杨奕 张家榕 张 慧 叶 鹰

摘 要 语言作为逻辑思维和推理工具,其基本要素是语词。重要语词映射成概念,核心概念建构知识,而知识演进促成学术发展。本文以学术文献中的关键词作为概念演化基础,尝试超越传统引文分析法探讨测量学术贡献的关键词分析法。研究发现以一个学科内的关键词向量和关键词流量为基准,可分别定义关键词向量通量谱和累积通量谱的 h 截断和 g 截断,构成 h 核和 g 核内主流关键词,进而可定义主流率和主流指数,用作测量对主流学术研究之学术贡献的新型学术测评参数。合并 CNKI、万方、维普三大中文数据库元数据和 WoS、Ei 里的中文记录元数据建立了 2008—2017 年数据支持平台,具体探讨了针对学术团体和学术期刊的中观层面实证案例。本文给出的关键词分析法基础框架重点是挖掘学术主体或客体对主流研究的学术贡献,建议在试用和应用中进一步改良和完善。图 5。表 5。参考文献 23。

关键词 关键词分析法 关键词向量 关键词流量 关键词通量 主流率 主流指数 学术评价

分类号 G250.252

Measuring Academic Contributions via Keyword Analytics

Charles C. HUANG, Star X. ZHAO, BIAN Yangyi, Ronda J. ZHANG, Helena H. ZHANG & Fred Y. YE

ABSTRACT

As a tool of logical thinking and derivation, the basic element of language is the word. Important words reflect concepts, and core concepts construct knowledge, while knowledge evolution contributes to academic development. Based on the theoretical foundation that keywords in the academic literature characterize concepts, this work attempts to get beyond traditional citation analysis and introduce a framework of academic contribution analysis based on keywords.

The framework of keyword analytics defines keyword vector flux spectrum and cumulative flux spectrum based on the keyword vector and keyword flow in a discipline, and then applies h -index and g -index to measure h -cutoff and g -cutoff, which constitute the mainstream keyword set in h -core and g -core. We then define the mainstream ratio and mainstream index, which can be used to measure the contribution of academic subjects (e.g. institutions) or academic objects (e.g. journals) to mainstream academic research.

The method of keyword analytics for measuring academic contributions inherits the theoretical characteristics of h -type metrics, and provides a new measurement in addition to citation analytics. In the field of humanities, the traditional quantitative evaluation methods have many limitations, and the keyword analysis method is worthy to explore. This method framework still has limitations and needs further

通信作者:叶鹰,Email: yye@nju.edu.cn, ORCID: 0000-0001-9426-934X (Correspondence should be addressed to Fred Y. Ye, Email: yye@nju.edu.cn, ORCID: 0000-0001-9426-934X)

innovation and development. For example, a series of conceptual clues such as keyword, theme-word, concept-word and ontology, are currently confused and need to be clarified. While using keyword analytics, the techniques of front-end word processing and back-end text mining still need to be further explored. On the basis of semantics, how to measure synonyms, near-synonyms, substitute words, and evolution of keyword itself accurately, is also a problem and needs further exploration.

The empirical data in this article is derived from the data platform built by the authors. The platform consists of three Chinese databases, CNKI, Wanfang and Weipu, and Chinese records in WoS and Ei. The data is cleaned and processed by machine plus manual work. The platform includes 87 million Chinese achievement data and 17 million foreign language achievement data, totaling 104 million, with a time span of 2008 to 2017. In this study, two data subsets on library and information science and philosophy were respectively selected as empirical cases. 5 figs. 5 tabs. 23 refs.

KEY WORDS

Keyword analytics. Keyword vector. Keyword flow. Keyword flux. Mainstream ratio. Mainstream index. Academic evaluation.

0 引言

引文分析自 Garfield 创立^[1]以来就一直被用作主流的学术测评方法,由此衍生的各种量化指标逐渐成为广泛应用的学术测度^[1]。尽管基于各类引文索引(Citation Index)及其伴生的诸如影响因子(IF)之类的量化测度曾经引导学术发展进步,但短暂的辉煌过后留下的问题也很多,不仅 SCI(Science Citation Index)被异化成 Stupid Chinese Indicator,而且出现有意增加或规避引文的不端行为、迎合测度指标的刻意造假等非学术干扰,以及各种计量指标皆有争议^[2]。在同行评议作为定性评价因复杂人际关系也经常失效的环境氛围里,如何保证学术测评的客观公正又定性定量一致,一直是图书情报学关注的问题^[3]。

在图书情报领域,以共词^[4]和关键词分析学科热点或前沿已成为一种通用方法并融入多种软件工具^[5-6]中,也有学者在关键词挖掘议题上做了有益的探索,如 Chen 等曾将“活性指数(Activity Index, AI)”指标用于关键词以探讨某一机构是否会强调某一特定的研究主题,进而提出“关键词活性指数(Keyword Activity Index,

KAI)”^[7],并用于对中国图书情报研究领域的 8 个机构的研究重点进行分析。Chen 和 Xiao 更进一步比较了三种基于文献关键词挖掘领域研究的方法^[8]:TF、TF-IDF 和 TF-KAI,并对它们的测量结果与专家定性判断的结果进行了对比,结果 TF-KAI 表现最佳。同时,有些学者对关键词属性与论文引用之间的关系进行过研究,如 Didegah 和 Thelwall 在调查作者高影响力论文的影响因素中发现关键词数目对论文影响力没有显著影响^[9]。Uddin 和 Khan 则从关键词的四个属性包括关键词长度、关键词多样性、关键词数量和新关键词所占比例探究了作者选择的关键词对引用次数的影响,研究表明除新关键词所占比例这个属性外,其他属性与引用次数表现为显著正相关^[10]。此外,共关键词网络则提供了关键词研究的另一走向^[11-12]。但关键词作为底层的知识单元表征,其潜力并未得到完全发挥,其质性优势有待量化表征,特别是在利用关键词构建新型学术测度方面有待突破。

近年国际上对新颖性(novelty)和创造性(creativity)的研究开启了另一热潮,已经有学者依据参考文献对新颖性和创造性指标进行了某些探索^[13-15]。Tahamtan 和 Bornmann^[16]针对参考文献的异常配对是否反映创造性或新颖性这

一问题进行了研究。他们挑选了科学计量学领域的一些具有里程碑意义的论文,并据此对相关作者进行了调查,研究表明在该领域中具有创造性的想法是在基于实际问题寻求解决方法、或与同事的讨论、或基于跨学科的交流中产生。随后有学者基于关键词而不是参考文献对衡量新颖性或创造性的指标进行了进一步探索^[17],但这些研究中都存在一个潜在的问题,即把新颖性和创造性混为一谈,最终难以深化。

在此基础上,本文基于关键词探寻可能存在的新方法,以期实现对创造性和新颖性的区分以及探索对学术的新型客观测评,期望找到定性内容与定量测度相融通的途径。

1 方法论

西方学术起源于古希腊人的理性逻辑思维,而语言是思维及推理工具。语言的基本要素是语词,重要语词映射成概念,故概念构成学术的基石。古希腊人正是在从概念到判断、从判断到推理中逐步建立起坚实的形式逻辑,构成哲学与科学的根基。如今,学术蓬勃发展,科学兴旺发达,离不开概念基础,每门学科都有基本概念即是明证。因此,通过学术文献中存在的概念和新产生的概念来判断学术发展及现状成为本文方法论思考的逻辑起点。

然而,学科中公认概念的形成需要长时间积累才能达成共识。基本概念一旦形成就凝聚成为学术思想的基石,同时也被固化并导致一定程度的坚守。这些固化的概念有利于学科知识体系稳固,但不利于学术开新。这种概念在图书情报学中映射为主题词表等,具有标引价值但失去创新活力。而那些以关键词形式出现的动态新概念则可能携带创新基因,应成研究重点。同时,在现有数据库记录中,关键词也是通用的数据字段,便于测量。因此,本文以关键词反映动态概念作为认识论起点并据以探讨方法论创新^[3]。

一般地,学术主体创造并使用关键词,并把

关键词记录在学术客体中。学术主体是关键词的创造者和使用者,包括学术个体(学者、个人)和学术团体(机构、大学)等;学术客体则是关键词的承载体和记录体,包括论著(论文、专著)和刊物(期刊、数据库)等。因此,学术主体创造知识,具有创造性;学术客体记录知识,具有新颖性。明确学术主体和学术客体概念有利于区分创造性和新颖性,而本文侧重考虑同时适用于学术主客体的主流创新测度。

1.1 关键词向量及其相似度

设某一学术主体使用过的关键词包括 K_{wi} , $i=1,2,\dots,i,\dots$,则其所用关键词构成主体关键词向量 V_s :

$$V_s = (K_{w1}, K_{w2}, \dots, K_{wi}, \dots) \quad (1)$$

同理,设某一学术客体中记录的关键词包括 K_{wj} , $j=1,2,\dots,j,\dots$,则其所有关键词构成客体关键词向量 V_o :

$$V_o = (K_{w1}, K_{w2}, \dots, K_{wj}, \dots) \quad (2)$$

与学科中所用关键词向量集 V 比较:

$$V = (K_{w1}, K_{w2}, \dots, K_{wk}, \dots) \quad (3)$$

可定义学术主体和学术客体的关键词向量及其与学科关键词向量的相似度如下:

定义 1. 学术主体关键词向量相似度:

$$S_s = \frac{V_s \cdot V}{|V_s| \cdot |V|} \quad (4)$$

定义 2. 学术客体关键词向量相似度:

$$S_o = \frac{V_o \cdot V}{|V_o| \cdot |V|} \quad (5)$$

关键词向量相似度越高的学术主体或学术客体,与存在的学科主流关键词向量集越近似,接近主流学术的概率越大。

1.2 关键词流量和通量

被文献数据库收录的学术论文大多标注有较为完善的关键词,这为通过关键词研究概念变迁提供了可能。一个被学界接受的关键词会被稳定地使用,因而学科主流关键词的主要特征如图 1 所示。

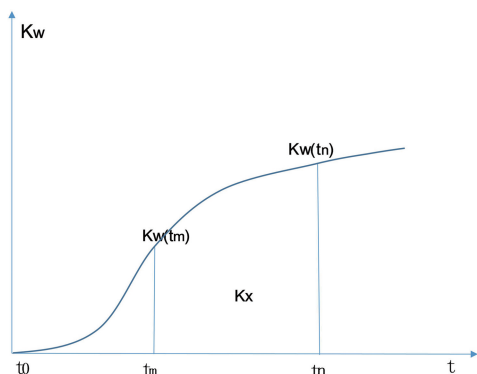


图1 学科主流关键词流量与通量示意图

图1横坐标为时间 t ,纵坐标是关键词 Kw 在相应时间的使用频次。

定义3. 关键词流量:关键词 Kw 在时间 t 的使用频次称为关键词流量 $Kw(t)$ 。

关键词流量 $Kw(t)$ 表征关键词在时间 t 被使用的频次。当关键词 Kw 在 t_0 时刻出现(创生)后,每年使用频次是连续时间中时刻发生着的使用量之叠加,最终以年度数值呈现,因而可把年度数值用平滑曲线连接形成关键词流量的分布曲线,简称关键词曲线。

流量积累则成通量,进而可定义关键词通量:

定义4. 关键词通量:关键词 Kw 在时间窗口从 t_m 到 t_n 的通量是

$$K_x = \int_{t_m}^{t_n} K_w(t) dt \quad (6)$$

关键词通量 K_x 表征关键词累积使用频次,通量 K_x 越大表明关键词累积使用越多。图1中可见 K_x 的几何意义是流量曲线 $Kw(t)$ 从 t_m 到 t_n 区间围成的面积。

一个学科中可能存在 N 个关键词,可产生 i 个($i=1,2,\dots,N$)关键词通量,于是可以有学科关键词累积通量概念,进而产生学科关键词平均通量。 t_m 至 t_n 时间窗口的学科关键词平均通量是

$$\bar{K}_x = \frac{1}{N} \sum_{i=1}^N \int_{t_m}^{t_n} K_{wi}(t) dt \quad (7)$$

尽管(6)和(7)以连续函数形式定义,具体计算可取离散数值加和形式。虽然通常使用的关键词计数是取离散值,但理论分析可采用连续函数。

1.3 关键词通量谱及其 h 截断和 g 截断

每个学科、学术主体或学术客体所用关键词构成的关键词向量各不相同,依各个关键词的通量降序排列后构成关键词向量通量谱:

$$S_v = \{K_{xw1}, K_{xw2}, \dots, K_{xwm}, \dots, K_{xwn}\} \quad (8)$$

设这些关键词的顺序累积通量为 K_{xk} ,则汇成关键词累积通量谱:

$$S_x = \{K_{x1}, K_{x2}, \dots, K_{xm}, \dots, K_{xn}\},$$

$$K_{xk} = \sum_{j=1}^k K_{xwj}, K_{x1} = K_{xw1} \quad (9)$$

依 h 指数^[18]概念,学科关键词向量通量谱可以有一个 h 截断;而把特定学术主体(学者个人、学术团体等)或特定学术客体(论著、论域等)中所用的 n 个关键词按通量降序排列,也都可以各有一个 h 截断,由此可以构成 h 型测度。这一顶端数据截断的 h 型测度方法在关键词分析中的理论依据是齐夫定律和信息学中广泛存在的幂律分布^[19]。其机理是:依据频次的排序分布都会存在高频顶端,而顶端关键词就是领域主流热词。

定义5. 关键词向量通量谱的 h 截断:设 r 是按通量降序排列的关键词序次, K_{xwr} 是关键词 r 的通量,则有

$$K_{xw1} \geq K_{xw2} \geq \dots K_{xwr} \geq \dots K_{xwn} \quad (10)$$

h 截断类比 h 指数定义为:

$$h_{cutoff} = \max\{r: r \leq K_{xwr}\} \quad (11)$$

一个学科关键词通量谱的 h 截断把该学科关键词区分为 h 核和 h 尾^[20]关键词。

同理,依 g 指数^[21]概念,关键词累积通量谱可以有一个 g 截断。

定义6. 关键词累积通量谱的 g 截断:设 r 是按通量降序排列的关键词序次, K_{xr} 是从1到 r 的累积通量,则有

$$K_{x1} \leq K_{x2} \leq \dots K_{xr} \leq \dots K_{xn} \quad (12)$$

类比 g 指数定义 g 截断

$$(g_{cutoff})^2 = \max\{r^2 : r^2 \leq K_x\} \quad (13)$$

一个学科关键词累积通量谱的 g 截断可以刻画该学科关键词序分布的 g 型分布结构。

关键词向量通量谱进行 h 截断、累积通量谱进行 g 截断后, h 核和 h 尾、 g 核和 g 尾的关键词分别构成关键词核向量和关键词尾向量, 如图 2, 于是理论上可定义一个学科的内核关键词

为主流关键词。当用学术主体或学术客体所用关键词与学科主流关键词进行比较时, 就能获得与主流学术关联的测度。 h 截断和 g 截断在方法上各有特点, h 截断主要与排序相关, 对 h 核内的关键词通量数值不敏感。 g 截断则将头部关键词通量纳入计算, 隐含假设: 头部关键词通量越大, 学术主体与客体的核心关键词规模也应随之扩大。应用中, 可根据测度目的选用。

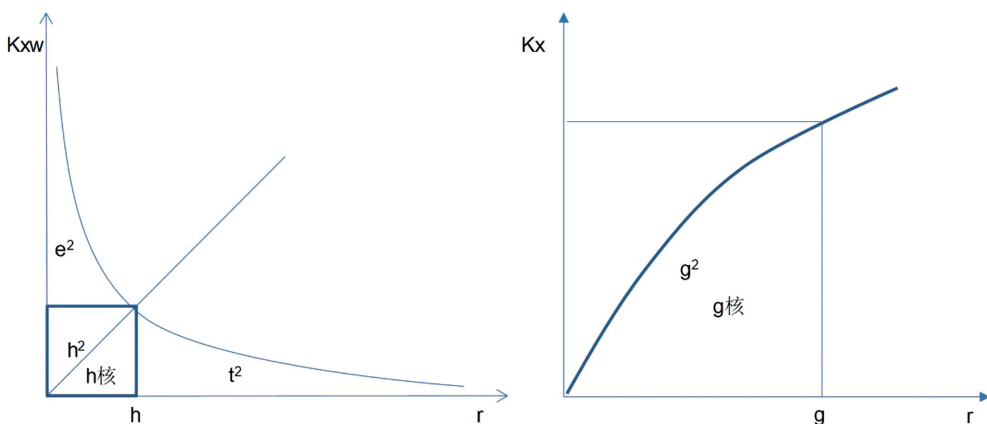


图 2 h 截断和 g 截断后的关键词核—尾分布

图 2 中 K_{xw} 表示按单个关键词 K_w 计算的通量, K_x 则表示其排序后的累积通量。

1.4 主流率和主流指数

学术主体和学术客体均可定义主流率和主流指数, 并分别有 h 型和 g 型。

定义 7. 主流率 MR : 分别以 h 截断和 g 截断定义 h 型主流率 MR_h 和 g 型主流率 MR_g 。以 h 截断后学科 h 核内主流关键词向量为 V , 把学术主体或学术客体所用关键词与 V 相比, 分别去除未用主流关键词而构成 V_s 或 V_o , 进而可按式 (4) 或式 (5) 计算学术主体或学术客体之 h 型主流率:

$$MR_h = S_s \text{ or } S_o \quad (14)$$

以 g 截断后学科 g 核主流关键词为基准, 分别按式 (4) 或式 (5) 可计算学术主体或学术客体之 g 型主流率:

$$MR_g = S_s \text{ or } S_o \quad (15)$$

MR 揭示学术主(客)体所用主流关键词与学科主流关键词相比的一致性, 即与主流研究的相似性, MR 越大相似性越高。

定义 8. 主流指数 MI : 学术主(客)体所用 h 或 g 核关键词通量与学科 h 或 g 核关键词通量之比分别构成其 h 型主流指数和 g 型主流指数:

$$MI_h = \frac{K_{ss-h-core}}{K_{x-h-core}} \quad (16)$$

$$MI_g = \frac{K_{ss-g-core}}{K_{x-g-core}} \quad (17)$$

MI 测量学术主(客)体所用核心关键词通量与学科主流关键词通量相比的比率, 表征从热点高频关键词角度测度一个学术主(客)体对学科的贡献。 MI 值越高, 就说明学术主(客)体在学科中重要学术议题上的贡献和参与度越高。根据本文关键词分析框架的定义, 理论上,

要取得较高的 MI 值,需要学术主(客)体的研究涉及领域和议题有一定广度、成果产出有一定数量的同时,对学科前沿热点有较宽的覆盖面。因此,MI 可以作为一个较为综合的贡献测度参数。

然而,MR 和 MI 不是价值导向参数,亦即并非 MR 或 MI 越大越好。最高 MR 和 MI 可能是追逐热点所致,中高 MR 和 MI 有望体现稳健态势。这会在后续实证案例中表现出来。

1.5 数据采集及支持平台

本文数据来源于作者团队构建的数据平台。该平台由 CNKI、万方、维普三大中文数据库元数据和 WoS、Ei 里中文记录元数据合并,采用机器清洗加人工介入的方式进行前端处理(特别是按词干归并同类词),目前完成中文成果数据总数量 8 700 万条,外文成果数据总数量 1 700 万条,合计数据 1.04 亿条。其中,核心关键词 505 万条,年度关键词跨度从 2008 至 2017 年。该平台集成了中文发表和中国学者发表的主要数据,对文科评价特有应用价值。本

研究中选用图书情报档案学和哲学两个数据子集作为实证案例。

2 实证结果

上述关键词分析框架适用于中观层面的学术测评,从数据平台提取关键词作为学术主体的学术团体和作为学术客体的学术期刊所进行的实证结果如下。

2.1 学术团体测算

学术主体包括学术个体(学者)和学术团体(院系、研究所、大学等),此处以 2008—2017 十年时间窗口对国内文科院系中的信息管理学院/系/所/图书馆和哲学系/所/学院进行测算。

(1)国内信息管理学院系/图书情报档案研究机构数据

以大学为单位,汇总每所大学内信息管理学院系、图书馆等图书情报档案研究机构的研究成果数据,测算其主流率和主流指数如表 1 所示。

表 1 国内图情档类学术团体的主流率和主流指数

学术机构	主流率		主流指数	
	MR _h	MR _g	MI _h	MI _g
武汉大学	0.607 7	0.628 2	0.204 8	0.245 4
南京大学	0.579 4	0.605 9	0.135 5	0.176 7
吉林大学	0.541 2	0.555 5	0.123 4	0.156 6
北京大学	0.529 8	0.548 3	0.115 6	0.136 6
中国科学院	0.526 9	0.551 3	0.132 6	0.172 8
中山大学	0.519 2	0.526 5	0.104 3	0.122 7
华中师范大学	0.496 1	0.515 1	0.081 9	0.113 4
华东师范大学	0.488 1	0.502 6	0.061 9	0.076 4
南开大学	0.484 8	0.483 2	0.098 5	0.099 5
四川大学	0.477 1	0.482 8	0.065 8	0.074 8
浙江大学	0.469 0	0.477 4	0.064 8	0.072 5
上海交通大学	0.444 7	0.440 3	0.061 9	0.073 3

续表

学术机构	主流率		主流指数	
	MR _h	MR _g	MI _h	MI _g
清华大学	0.434 3	0.440 8	0.063 9	0.074 1
安徽大学	0.416 9	0.441 9	0.058 5	0.073 3
东南大学	0.406 4	0.414 3	0.040 5	0.043 2
南昌大学	0.405 2	0.428 7	0.043 4	0.057 1
河海大学	0.399 0	0.400 3	0.037 1	0.045 5
宁波大学	0.394 3	0.404 7	0.048 8	0.064 0
中国人民大学	0.393 4	0.428 1	0.085 3	0.107 3
哈尔滨工业大学	0.387 8	0.392 3	0.046 3	0.044 8
南京农业大学	0.379 5	0.393 8	0.051 7	0.071 8
复旦大学	0.372 4	0.379 6	0.037 5	0.054 0
苏州大学	0.368 8	0.384 8	0.052 7	0.066 4
北京邮电大学	0.361 5	0.360 2	0.037 1	0.055 6
上海图书馆	0.358 2	0.361 0	0.063 9	0.064 0
暨南大学	0.351 6	0.350 9	0.052 7	0.038 6
兰州大学	0.344 3	0.352 6	0.046 8	0.054 0
山东大学	0.340 8	0.346 2	0.047 3	0.048 6
西北工业大学	0.340 6	0.347 2	0.041 9	0.052 5
重庆大学	0.340 2	0.344 6	0.055 1	0.055 6
江苏大学	0.338 8	0.349 2	0.048 3	0.071 0
集美大学	0.338 5	0.324 3	0.037 1	0.030 9
大连理工大学	0.334 0	0.352 7	0.055 6	0.078 7
同济大学	0.327 9	0.335 2	0.034 6	0.048 6
华侨大学	0.327 3	0.322 6	0.032 7	0.047 1
浙江理工大学	0.322 1	0.327 7	0.023 9	0.037 8
中国社会科学院	0.321 1	0.334 0	0.031 7	0.047 8
华东理工大学	0.319 5	0.347 2	0.024 9	0.041 7
南京邮电大学	0.310 2	0.314 6	0.022 4	0.037 8
北京理工大学	0.306 4	0.314 9	0.034 6	0.051 7
哈尔滨工程大学	0.303 8	0.322 1	0.049 2	0.047 1
西北大学	0.302 9	0.324 1	0.030 2	0.041 7
扬州大学	0.297 7	0.316 7	0.040 0	0.045 5
国家图书馆	0.292 3	0.301 4	0.072 6	0.084 1

续表

学术机构	主流率		主流指数	
	MR _h	MR _g	MI _h	MI _g
南京图书馆	0.284 4	0.299 2	0.046 8	0.039 4
南京师范大学	0.283 4	0.290 4	0.024 4	0.044 8
上海财经大学	0.279 6	0.276 0	0.015 1	0.023 9
首都师范大学	0.265 0	0.264 2	0.041 4	0.030 1
浙江图书馆	0.247 8	0.242 8	0.065 8	0.037 8
江苏科技大学	0.246 4	0.242 1	0.020 5	0.032 4
中国海洋大学	0.238 2	0.240 3	0.014 1	0.023 9
上海外国语大学	0.220 6	0.212 5	0.011 2	0.017 7
华东政法大学	0.218 3	0.220 8	0.012 2	0.019 3
中国美术学院	0.179 6	0.185 0	0.005 9	0.009 3
南京信息工程大学	0.134 1	0.136 6	0.016 6	0.026 2
国家档案局	0.119 3	0.136 8	0.018 0	0.032 4
杭州图书馆	0.117 4	0.131 3	0.028 8	0.026 2

表1数据中,武汉大学、南京大学所用关键词体现国内图书情报档案学主流和热点,其主流率和主流指数皆领先。而北京大学、中国科学院等则体现在图书情报档案主流关键词使用上的相对稳健。应指出的是,因研究数据采用图情档一级学科的合集,这一数据合集中,图书馆学和情报学期刊占主要部分,档案学期刊只有两

种,故以档案学见长的机构在主流率和主流指数上都没有突出表现。这一结果也再次说明,在学术测评中,细分到底层学科研究领域很有必要。

(2) 国内哲学系/所/院机构数据

同样以大学为单位,汇总每所大学内哲学系、学院和研究所的研究成果数据,测算其主流率和主流指数如表2所示。

表2 国内哲学类学术团体的主流率和主流指数

学术机构	主流率		主流指数	
	MR _h	MR _g	MI _h	MI _g
中山大学	0.404 4	0.366 8	0.145 1	0.159 0
中国人民大学	0.403 2	0.404 7	0.128 2	0.154 4
复旦大学	0.366 6	0.351 8	0.105 2	0.126 9
北京师范大学	0.353 6	0.342 6	0.099 2	0.119 3
南京大学	0.349 7	0.340 9	0.106 4	0.104 0
清华大学	0.348 3	0.338 3	0.104 0	0.131 5
武汉大学	0.338 4	0.320 8	0.100 4	0.110 1
中国社会科学院	0.336 7	0.337 9	0.111 2	0.134 6

续表

学术机构	主流率		主流指数	
	MR _h	MR _g	MI _h	MI _g
北京大学	0.334 2	0.318 1	0.093 1	0.111 6
南京师范大学	0.303 0	0.261 9	0.091 9	0.099 4
山东大学	0.301 2	0.283 7	0.166 9	0.139 1
南开大学	0.294 0	0.285 5	0.070 1	0.096 3
吉林大学	0.292 8	0.305 0	0.058 0	0.073 4
浙江大学	0.270 4	0.276 6	0.093 1	0.108 6
湖南师范大学	0.262 3	0.217 9	0.123 3	0.105 5
曲阜师范大学	0.261 7	0.227 8	0.072 6	0.078 0
湖北大学	0.260 4	0.225 1	0.049 6	0.067 3
湖南大学	0.258 7	0.207 4	0.056 8	0.071 9
黑龙江大学	0.255 4	0.216 1	0.067 7	0.091 7
中国政法大学	0.255 1	0.226 8	0.047 2	0.064 2
山西大学	0.249 5	0.248 2	0.140 3	0.110 1
东南大学	0.248 7	0.252 8	0.091 9	0.099 4
华南师范大学	0.236 2	0.229 5	0.079 8	0.100 9
华中科技大学	0.231 8	0.242 5	0.058 0	0.073 4
华中师范大学	0.227 0	0.226 9	0.049 6	0.067 3
厦门大学	0.225 4	0.188 3	0.090 7	0.096 3
苏州大学	0.219 1	0.188 4	0.027 8	0.035 2
同济大学	0.217 0	0.209 9	0.043 5	0.064 2
大连理工大学	0.216 9	0.195 0	0.062 9	0.084 1
陕西师范大学	0.197 1	0.200 6	0.061 7	0.078 0
华侨大学	0.195 3	0.178 6	0.030 2	0.041 3
安徽师范大学	0.194 2	0.189 6	0.044 7	0.061 2
中南大学	0.186 6	0.186 3	0.058 0	0.073 4
华东师范大学	0.185 0	0.211 9	0.056 8	0.071 9
中国科学院	0.183 6	0.1705	0.101 6	0.102 4
安徽大学	0.177 7	0.170 8	0.026 6	0.033 6
中南财经政法大学	0.172 9	0.151 5	0.045 9	0.062 7
西南大学	0.166 4	0.164 3	0.038 7	0.048 9
上海社会科学院	0.165 3	0.155 1	0.019 3	0.024 5
河南大学	0.160 9	0.149 0	0.027 8	0.035 2

续表

学术机构	主 流 率		主 流 指 数	
	MR _h	MR _g	MI _h	MI _g
上海财经大学	0.151 1	0.136 7	0.029 0	0.036 7
首都师范大学	0.147 3	0.175 8	0.042 3	0.053 5
西安交通大学	0.141 0	0.145 3	0.030 2	0.038 2
辽宁大学	0.130 1	0.139 2	0.021 8	0.030 6
内蒙古大学	0.126 6	0.136 8	0.029 0	0.036 7
武汉理工大学	0.116 3	0.132 2	0.033 9	0.047 4
中国自然辩证法研究会	0.097 5	0.090 7	0.1245	0.105 5
上海交通大学	0.089 9	0.126 6	0.047 2	0.064 2
燕山大学	0.069 4	0.089 0	0.023 0	0.029 1
东华大学	0.055 2	0.058 5	0.018 1	0.022 9
国防科技大学	0.044 8	0.043 7	0.024 2	0.033 6

从表2数据可见中山大学、中国人民大学所用关键词体现国内哲学主流和热点,其主流率和主流指数皆领先。而北京师范大学、南京大学等则体现在哲学主流关键词使用上的相对稳健。

2.2 学术期刊测算

分别以国内图书情报档案学和哲学的学术期刊为实例测算其主流率和主流指数。

(1) 国内图书情报档案学期刊

以国内20种图书情报档案学期刊为样本,测算其主流率和主流指数如表3所示。

表3数据表明《情报科学》《情报杂志》所用关键词位列国内图书情报档案学主流关键词前列,说明其具有发表流行热点的趋势。而《情报学报》《中国图书馆学报》则体现在主流关键词标引发表上的相对稳健。

表3 国内图书情报档案学期刊的主流率和主流指数

学术期刊	主 流 率		主 流 指 数	
	MR _h	MR _g	MI _h	MI _g
情报科学	0.634 9	0.650 3	0.267 2	0.298 6
情报杂志	0.573 1	0.594 6	0.305 7	0.298 6
图书情报工作	0.554 9	0.572 4	0.314 0	0.375 0
图书情报知识	0.553 9	0.561 3	0.126 3	0.149 7
情报理论与实践	0.523 3	0.538 4	0.211 1	0.251 5
情报学报	0.481 6	0.503 4	0.154 6	0.194 4
中国图书馆学报	0.480 1	0.492 9	0.130 7	0.150 5
图书与情报	0.479 7	0.483 0	0.118 5	0.133 5
图书馆杂志	0.456 7	0.461 7	0.149 7	0.169 0

续表

学术期刊	主流率		主流指数	
	MR _h	MR _g	MI _h	MI _g
情报资料工作	0.453 3	0.448 8	0.170 6	0.174 4
图书馆学研究	0.452 7	0.447 8	0.244 3	0.263 1
现代图书情报技术	0.450 3	0.461 7	0.182 4	0.226 9
大学图书馆学报	0.435 2	0.434 9	0.133 1	0.145 1
图书馆论坛	0.410 8	0.407 5	0.199 4	0.204 5
图书馆	0.384 8	0.381 7	0.168 7	0.183 6
图书馆工作与研究	0.376 7	0.370 0	0.200 9	0.196 0
图书馆建设	0.369 9	0.364 1	0.180 4	0.177 5
国家图书馆学刊	0.347 0	0.350 8	0.095 1	0.108 8
档案学研究	0.236 3	0.267 3	0.120 4	0.148 1
档案学通讯	0.216 1	0.251 3	0.121 4	0.165 9

(2) 国内哲学期刊

以国内 12 种哲学期刊为样本,测算其主流率和主流指数如表 4 所示。

从表 4 可见,《道德与文明》主流率位列第

一但主流指数列第二,表明其所用关键词多为学科主流关键词,但贡献的核心关键词却并非学科首要核心关键词。《自然辩证法研究》主流率和主流指数皆高,表明其研究论文与学科主

表 4 国内哲学期刊的主流率和主流指数

学术期刊	主流率		主流指数	
	MR _h	MR _g	MI _h	MI _g
道德与文明	0.420 8	0.379 6	0.303 5	0.319 6
自然辩证法研究	0.400 1	0.377 1	0.463 1	0.455 7
现代哲学	0.388 5	0.355 4	0.240 6	0.255 4
哲学研究	0.371 6	0.395 4	0.228 5	0.261 5
伦理学研究	0.364 4	0.323 8	0.274 5	0.276 8
世界哲学	0.364 3	0.375 8	0.227 3	0.235 5
哲学动态	0.346 3	0.411 1	0.251 5	0.279 8
中国哲学史	0.327 6	0.313 7	0.218 9	0.220 2
自然辩证法通讯	0.280 1	0.296 2	0.206 8	0.233 9
孔子研究	0.273 0	0.2414	0.266 0	0.2095
科学技术哲学研究	0.271 0	0.287 5	0.156 0	0.169 7
周易研究	0.220 0	0.196 9	0.2527	0.204 9

流热点密切相关。而《哲学研究》则体现学科关键词使用的稳健呈现。

以上学术团体和学术期刊的主流率或主流指数皆高度相关,其相关系数见表 5。

表 5 主流率和主流指数之相关系数

项目		Pearson	Spearman	N
图情档_学术团体	主流率	0.996 4	0.991 7	57
	主流指数	0.967 4	0.881 8	57
图情档_学术期刊	主流率	0.992 8	0.991 0	20
	主流指数	0.961 6	0.960 5	20
哲学_学术团体	主流率	0.977 3	0.970 7	51
	主流指数	0.946 3	0.982 5	51
哲学_学术期刊	主流率	0.870 2	0.776 2	12
	主流指数	0.926 9	0.629 4	12

其中主流率 MR_h 和 MR_g 、主流指数 MI_h 和 MI_g 之间的相关系数指向一致,表明 h 型主流率或主流指数与 g 型主流率或主流指数具有强正相关性,必要时只需选择其中某个指标作为代表。

g 核性能提供了一个分析依据:g 核参数与引文测度大体成正比,g 核与 h 核的强相关则提供了 h 型和 g 型主流率与主流指数的一致指向。至于关键词与主题词之间的关联,则有待深入研究。

如果关键词分布函数是逻辑曲线:

$$K_w(t) = \frac{k}{1 + ae^{-bt}} \quad (21)$$

则其分布函数 $F(t)$ 呈现为积分曲线:

$$F(t) = \int_{-\infty}^t K_w(x) dx = k \int_{-\infty}^t \frac{dx}{1 + ae^{-bx}} \quad (22)$$

由此可望进行特定条件下的理论分析。

同时,时序积分通量 Kx 是一个递增函数,同一数据平台上相同时间窗口的任何关键词通量应是一个可比数值,其数值大小表明关键词的累积影响大小。如果某学术主体提出的独特关键词获得学界认可,就会产生较大通量,因而独特关键词通量越大,其对主流学术的贡献也越大。

3 分析与讨论

现有 h 型测度和 g 型测度的研究成果可为本研究提供丰富的理论分析基础。

3.1 理论分析

在 Egghe 的 Lotkaian 信息计量学框架^[19]下,g 指数和 h 指数的数学关系为

$$g = \left(\frac{\alpha - 1}{\alpha - 2} \right)^{\frac{\alpha-1}{\alpha}} h \geq h \quad (18)$$

其中 α 是 Lotka 指数。赵星等分析过 g 指数和主题词之间的关联^[22-23],并在引文分析法框架下得到 g 指数与词频 W 之间存在幂律关系,g 核内论文总被引 C 接近 g^2 的经验规律:

$$g = cW^d \quad (19)$$

$$g^2 \approx \sum_{i=1}^g C_i \quad (20)$$

于是 g 核内篇均被引近似等于 g 。这为判断

3.2 局限与改进设想

任一关键词皆有生命周期,考虑关键词生命周期后细化后的关键词演化曲线如图 3 所示。

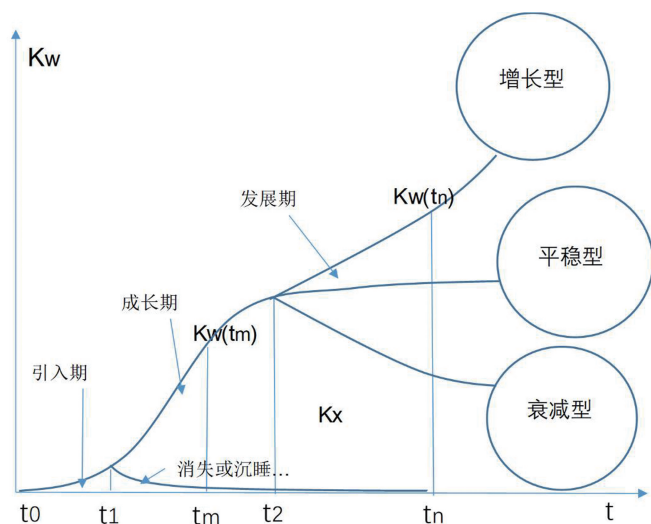


图3 关键词时序分布及演化模型

其中“冷词”将消失或衰减,只有“热词”能平稳应用甚至增长,所以本文讨论的“h 截断”和“g 截断”皆针对“热词”,如果每个“热词”的生命周期都能获得数据描述,则可细化关键词分析法,把每个热词关键词生命周期皆测算出来作为进一步研究的基础。图4和图5分别给出了图书情报档案学科中“数字图书馆”和“人工智能”两个热点关键词的时序分布及演化实

例。其中,“数字图书馆”近年属于衰减型,而“人工智能”属于增长型(2019年数据不完全)。

但由于关键词本身存在的固有局限,基于关键词发展的测评方法可能难以突破当前引文主导格局。虽然关键词分析法提出了一个可能的测评途径,但仍有明显局限,后续需要解决的问题如:

(1) 关键词—主题词—概念词—本体线索



图4 图书情报档案学科热点关键词“数字图书馆”时序分布及演化(1999—2019)

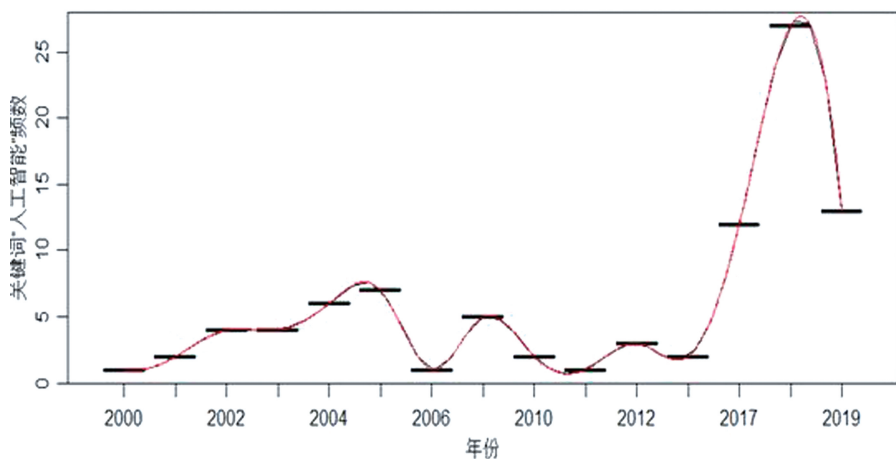


图5 图书情报档案学科热点关键词“人工智能”时序分布及演化(2000—2019)

的严格逻辑及定义;

(2)前端技术处理和后端文本挖掘的技术辅助及改良;

(3)同义词、近义词的语义处理。

此外,当把中观层面的实证结果向微观层面(比如学术个体)和宏观层面(比如国家)扩展时,需要评估其适用性和有效性。

3.3 扩展可能

关键词是学术主体使用或学术客体中标记的一种特征信息,具有简单易用性,然而,既然简单,就必然导致信息缺失,从方法论角度可能补救或扩展的可能有:

(1)将关键词范围扩展到标题、摘要甚至全文而不仅限于关键词字段,建立集成量,但这样将导致计算量大幅增加,需要考虑成本和收益的平衡。

(2)以关键词分析法为起点,今后在自然语言理解支持下采用计算机技术方法归纳提取概念,真正实现概念判识,应是一个发展方向。

(3)以关键词分析联动引文分析,以关键词而非全文为基础单元重构引文线索,从而考虑知识概念的传播与流动,体现学术内容的发展与演变。

(4)把学术主体指向 Creativity,学术客体指向 Novelty 测度,应是努力方向。

3.4 问题讨论

上述关键词分析法提供的是与主流学术研究的相近程度测量,MR 和 MI 越高意味着与主流吻合越大。MR 和 MI 过低在中观层面未必是好事,但 MR 和 MI 并非总是越大越好,建议不以 MR 和 MI 为价值判断意向。从发展角度看,与主流近似是好事,但从创新角度看则不一定。关键词分析法或许能启发问题逐渐暴露的引文分析法考虑一些学术测评新思想和新方法。

关键词法仍受制于数据库和时间窗口选择,该方法的理论基础及技术支持系统皆有待完善。从关键词生命周期看,新关键词提出后总会经历引入期,部分新关键词很快消失而没有成长期,经历引入期和成长期后的新关键词才进入发展期,而发展期可能面临增长、平稳、衰减三类走势。因此,结合不同时期配置不同权重可作为以后改进的一种策略。然而,仅仅依靠关键词设计学术测评新方法仍然可能具有缺陷,今后可望结合智能信息分析继续深入研究。

4 小结与建议

关键词是学术概念的体现,关键词映射概念演化,学术主体提出的概念以关键词形式记

录在学术客体中,学界若认同则会产生较大关键词流量和通量,从而形成流量和通量越大越接近主流的测评指向。以一个学科内的关键词向量和关键词通量为基准,其向量通量谱和累积通量谱分别经h截断和g截断后可分别定义基于h核和g核主流关键词的主流率和主流指数,用作测量对主流学术贡献的参考。合并CNKI、万方、维普三大中文数据库元数据和WoS、Ei里中文元数据作为实证数据平台提供的中观层面实例支持这些论点。

建议在实践中逐步改良完善关键词分析法,使之与引文分析法互为补充,以促进学术发展进步。

致谢:感谢中国工程院项目基金对数据平台建设的支持、华东师范大学“中国哲学社会科学国际化研究前沿”战略项目和中央高校基本科研业务费专项资金资助(编号14370301),并感谢上海交通大学图书馆谭旻博士、浙江大学图书馆薛霏馆员等多次参加合作研讨。

参考文献

- [1] Garfield E. Citation indexing: its theory and application in science, technology and humanities [M]. New York: Wiley, 1979
- [2] 叶鹰. 国际学术评价指标研究现状及发展综述[J]. 情报学报, 2014, 33(2): 215-224. (Ye F Y. The research progress and developing perspective of assessment indicators [J]. Journal of the China Society for Scientific and Technical Information, 2014, 33(2): 215-224.)
- [3] 叶鹰. 图书情报学的学术思想与技术方法及其开新[J]. 中国图书馆学报, 2019, 45(2): 15-25. (Ye F Y. An essay on the academic thoughts and technical methods with their renewing in library and information science [J]. Journal of Library Science in China, 2019, 45(2): 15-25.)
- [4] 冯璐, 冷伏海. 共词分析方法理论进展[J]. 中国图书馆学报, 2006, 32(2): 88-92. (Feng Lu, Leng Fuhai. Development of theoretical studies of co-word analysis [J]. Journal of Library Science in China, 2006, 32(2): 88-92.)
- [5] 刘启元, 叶鹰. 文献题录信息挖掘技术方法及其软件 SATI 的实现——以中外图书情报学为例 [J]. 信息资源管理学报, 2012, 1(1): 50-58. (Liu Qiyuan, Ye F Y. A study on mining bibliographic records by designed software SATI: case study on library and information science [J]. Journal of Information Resources Management. 2012, 1(1): 50-58.)
- [6] Chen C. Mapping scientific frontiers [M]. London: Springer-Verlag, 2013 (2nd ed.).
- [7] Chen G, Xiao L, Hu C P, et al. Identifying the research focus of library and information science institutions in China with institution-specific keywords [J]. Scientometrics, 2015, 103(2): 707-724.
- [8] Chen G, Xiao L. Selecting publication keywords for domain analysis in bibliometrics: a comparison of three methods [J]. Journal of Informetrics, 2016, 10(1): 212-223.
- [9] Didegah F, Thelwall M. Which factors help authors produce the highest impact research? Collaboration, journal and document properties [J]. Journal of Informetrics, 2013, 7(4): 861-873.
- [10] Uddin S, Khan A. The impact of author-selected keywords on citation counts [J]. Journal of Informetrics, 2016, 10(4): 1166-1177.
- [11] 叶鹰, 张力, 赵星, 等. 用共关键词网络揭示领域知识结构的实验研究 [J]. 情报学报, 2012, 31(12): 1245-

1251. (Ye F Y, Zhang P L, Zhao S X, et al. An experimental study on revealing domain knowledge structure by co-keyword networks [J]. Journal of the China Society for Scientific and Technical Information, 2012, 31(12): 1245-1251.)
- [12] Li H, An H, Wang Y, et al. Evolutionary features of academic articles co-keyword network and keywords co-occurrence network: based on two-mode affiliation network [J]. Physica A: Statistical Mechanics and its Applications, 2016, 450: 657-669.
- [13] Uzzi B, Mukherjee S, Stringer M, et al. Atypical combinations and scientific impact [J]. Science, 2013, 342(6157): 468-472.
- [14] Lee Y N, Walsh J P, Wang J. Creativity in scientific teams: unpacking novelty and impact [J]. Research policy, 2015, 44(3): 684-697.
- [15] Wang J, Veugelers R, Stephan P. Bias against novelty in science: a cautionary tale for users of bibliometric indicators [J]. Research Policy, 2017, 46(8): 1416-1436.
- [16] Tahamtan I, Bornmann L. Creativity in science and the link to cited references: is the creative potential of papers reflected in their cited references? [J]. Journal of Informetrics, 2018, 12(3): 906-930.
- [17] Carayol N, Llopis O, Lahatte A. Capturing scientific novelty through paper keyword combinations [C]//Ràfols I, Molas-Gallart J, Castro-Martínez E, et al. Proceedings of the 21st International Conference on Science and Technology Indicator. València, Spain: Universitat Politècnica de València, 2016.
- [18] Hirsch J E. An index to quantify an individual's scientific research output [J]. Proceedings of the National Academy of Sciences of the USA, 2005, 102(46): 16569-16572.
- [19] Egghe L. Power laws in the information production process: Lotkaian Informetrics [M]. Amsterdam: Elsevier, 2005.
- [20] Ye F Y, Rousseau R. Probing the h-core: an investigation of the tail-core ratio for rank distributions [J]. Scientometrics, 2010, 84(2): 431-439.
- [21] Egghe L. Theory and practice of the g-index [J]. Scientometrics, 2006, 69(1): 131-152.
- [22] 赵星, 高小强, 郭吉安, 等. 基于主题词频和 g 指数的研究热点分析方法 [J]. 图书情报工作, 2009, 53(2): 59-61, 7. (Zhao Xing, Gao Xiaoqiang, Guo Ji'an, et al. Research focus analysis based on the frequency of topic words and g-index [J]. Library and Information Services. 2009, 53(2): 59-61, 7.)
- [23] 赵星, 李盛庆, 叶鹰, 等. H 型指数和 H 型测度研究 [M]. 北京: 科学出版社, 2018: 177-182. (Zhao Xing, Li Shengqing, Ye F Y, et al. H-type indices and H-type measures [M]. Beijing: Science Press, 2018: 177-182.)

黄 晨 浙江大学图书馆研究馆员。浙江 杭州 310027。

赵 星 华东师范大学信息管理系教授, 博士生导师。上海 200241。

卞杨奕 浙江大学图书馆数据分析师。浙江 杭州 310027。

张家榕 南京大学信息管理学院博士研究生。江苏 南京 210023。

张 慧 南京大学信息管理学院博士研究生。江苏 南京 210023。

叶 鹰 南京大学信息管理学院教授, 博士生导师。江苏 南京 210023。

(收稿日期: 2019-08-06)