

DOI:10.12154/j.qbzlgz.2021.06.005

高关注论文视角下研究前沿识别多方法融合研究

高楠^{1,2} 周庆山¹ 赵蕴华²(¹北京大学信息管理系 北京 100871; ²中国科学技术信息研究所 北京 100038)

摘要: [目的/意义] 引文分析法与 Altmetrics 分别代表了不同侧重点下的学术评价方法,为聚焦研究前沿来源文献的遴选范围,将结合二者用于论文的综合影响力评价。不同的研究前沿识别方法优缺点不同,为更全面、客观地识别,将在4种不同的研究前沿识别方法得到的结果基础上,探索融合不同前沿识别结果的途径。[方法/过程] 选取领域的高关注量文献集进行共被引分析、耦合分析,分别筛选出高关注、高被引文献集和高关注、低被引文献集进行词频分析,并在混合这两种文献集的基础上进行共词分析,对比不同研究前沿识别方法所得结果的异同,之后采用改进的线性加权融合模型进行前沿结果融合。[结果/结论] 通过对人工智能领域进行实证分析,验证了本文方法的可行性。研究发现,耦合分析比共被引分析能识别到更多的研究前沿,二者识别到的前沿结果中有约40%的论文相同,与 Jarneving 的研究结果相近;词频分析法识别结果的特点是多样且碎片化,共词分析识别结果的特点是少但更明确。融合后得到的识别结果包括人工智能算法与应用两类,其中,虚拟现实、中风康复、人脸识别、智能安防、智能医疗是时下最具潜力的人工智能应用领域。

关键词: 研究前沿 Altmetrics 关注度 融合

Research on Multi-method Fusion of Research Front Identification from the Perspective of High-Profile Papers

Gao Nan^{1,2} Zhou Qingshan¹ Zhao Yunhua²(¹Department of Information Management, Peking University, Beijing, 100871;²Institute of Scientific & Technical Information of China, Beijing, 100038)

Abstract: [Purpose/significance] Citation analysis method and Altmetrics respectively represent academic evaluation methods in different ways. In order to select the source of frontier documents precisely, the two methods will be combined for the comprehensive impact evaluation of papers. Different research front identification methods have different advantages and disadvantages. In order to identify more comprehensively and objectively, first we use four different research front identification methods to get different identification results, then try to explore the way to integrate the identification results of different research front methods. [Method/process] Co-citation analysis and coupling analysis are carried out on the basis of high-attention documents, and word frequency analysis is carried out respectively on the basis of high-attention, high-cited documents and high-attention, low-cited documents. Besides, co-word analysis is carried out on the basis of the previous two document collections. After that, the similarities and differences of the results obtained by different research front recognition methods are compared. Then, the improved linear weighted fusion model is used to merge the frontier results. [Result/conclusion] Through empirical analysis in the field of artificial intelligence, the feasibility and effectiveness of this method are verified. The study found that coupling analysis could identify more research fronts than co-citation analysis. About 40% of the research fronts identified by the two methods are the same, which is similar to Jarneving's research results. The characteristics of the recognition results of word frequency analysis are diverse and fragmented, the characteristics of coword analysis are fewer but more clear. The research front recognition results obtained by the fusion analysis method include two aspects, which are the algorithms and applications of artificial intelligence. Among them, virtual reality, stroke rehabilitation, face recognition, intelligent security, and intelligent medical care are the most potential artificial intelligence application fields.

Keywords: research front Altmetrics paper attention fusion

1 引言

从德国工业4.0战略到欧盟“地平线2020”科研计划,再到中国“十三五”国家科技创新规划,各国都积极加强科技创新部署并制定了一系列科技发展战略。在此背景下,及时探测、识别并追踪领域的前沿热点,综合分析和把握重要领域科技创新发展的最新脉搏,成为情报研究的重要任务。而随着开放获取运动的蓬勃发展,以及以“海量、同步、共生、智慧”为代表的Web 4.0时代的到来^[1],信息传播渠道和获取方式越来越多元化,社交媒体等各类Web平台数据成为探测研究前沿不可或缺的信息源。

Altmetrics通过收割学术论文在社交媒体平台上的下载、评论、分享、点赞、收藏、转发等指标表现,利用互联网大数据来表征论文的知名度、可见度和影响力^[2],其具备数据范围广、指标全面丰富、受众群体多样化、时效性强、传播范围广、评价过程透明等特点^[3]。而作为表征论文学术影响力的传统经典方法——同行评议和引文分析法,因难以避免其引文统计滞后性、评价方式单一、学科偏倚性、马太效应等缺陷^[4],已难以满足学界日益增长的评价需求,因此需要引入新的计量方式,向多元化指标评价体系转变。

同行评议与引文分析法侧重于面向专业领域的学术影响力评价,其中,引用指标的产生主要涉及学术交流中发生引证行为的阶段,而Altmetrics则侧重于面向大众评议的社会影响力评价,Altmetrics指标主要产生于大众传播交流的过程,二者分别揭示了学术成果的交流深度和传播广度,共同构成了学术成果影响力评价的完备性体系。相关研究如Noorden等^[5]通过大规模数据分析,发现Twitter提及量与引文量没有关系,即Twitter提及量不能反映学术影响力,其研究结论被Nature网站认可并报道。此外,Kousha等^[6-9]分别从不同数据来源、不同数据类型、不同学科、不同评价指标等角度,结合相关性分析、因子分析、主成分分析等方法,对被引量与Altmetrics指标间的相关性进行判断,均得到了二者相关性较弱的结论,即论文的高被引量与高Altmetrics分值没有必然相关性。引文分析法与Altmetrics分别代表了不同侧重点下的学术评价的方法,二者相互独立、互为补充,共同反映出学术成果不同方面的影响力,故在评价时应结合起来应用。

研究前沿(Research Fronts)是文献计量学里的一个专有名词,最早由Price^[10]于1965年提出,他认为研究前沿由领域的30~50篇最新发表的高被引文献及其相关研究主题来反映。经过近55年发展,研究前沿已衍生

出多种定义和识别方法,但目前国内外学者尚未形成统一的研究前沿定义和判别标准,通常是依据其采用的数据源、分析方法和研究目的来定义。现汇集研究前沿的多种定义,如表1所示。

表1 研究前沿概念汇总

概念分类	作者	年份	研究前沿	识别方法	
将一组高被引文献或共被引聚类文献定义为研究前沿	Price ^[10]	1965	将一个领域内的30~50篇近期发表,且有较高被引次数的文献集称为研究前沿。	直接引用	基于引文的探测方法
	Small&Griffith ^[11]	1974	进行共被引聚类,将引文网络中聚合的高被引文献称为研究前沿。	共被引分析	
	Garfield ^[12]	1991	共被引聚类数和引文的总和。	共被引分析	
将耦合聚类文献定义为研究前沿	Persson ^[13]	1994	提出知识基础和研究前沿,研究前沿是知识基础的施引文献簇,即高被引文献簇的引文文献簇。缺点:研究前沿和知识基础界限模糊,使得部分文献在共被引频率阈值较高时属于研究前沿,阈值较低时属于知识基础。	耦合分析	耦合分析
	Morris等 ^[14]	2003	克服Persson理论缺点,首先确定一组时间和组成单元都固定基础文献簇,引用这个基础文献簇的另一个文献簇则构成研究前沿。	耦合分析	
将突发、热点词汇、共词聚类定义为研究前沿	Suji等 ^[15]	1998	统计在同一文献中两个词汇共同出现的次数,以此为基础对这些词进行聚类分析,从而反映出这些词之间的亲疏关系,进而分析这些词所代表的学科和主题的结构变化。	共词分析	基于内容主题的探测方法
	Robert等 ^[16]	1991	具有不同知识和社会背景的许多科学研究者共同关注的一组相关的研究问题和概念	词频分析	
	Kleinberg ^[17]	2003	通过词频密度变化的突破检测算法探测研究前沿。	词频分析	
	Blei等 ^[18-19]	2003	先后提出LDA、DLDA模型,通过分析文档中词语分布情况,运用概率统计的方法探测前沿主题。	概率主题模型	
面向国家重点领域或重大战略需求的具有前瞻性、先导性、理论性、探索性的研究内容	刘小平等 ^[20]	2012	未来的科技前沿是指政府的科技规划战略路线图;当前的科技前沿,是指世界科技强国的资助机构通过各类计划项目最新资助的战略投资重点领域。	定性分析法	

共被引分析、耦合分析、词频分析和共词分析法是探测研究前沿中普遍应用的方法,每种方法各有其优缺点,可根据不同的情况和应用环境选择不同的方法。为使研究前沿的识别结果更具可靠性和客观性,在采用多种研究前沿识别方法的基础上,探索融合不同方法所得前沿识别结果的路径。在筛选研究前沿的来源文献集时,考虑到Altmetrics指标及时性的特点,更有利于捕捉领域的前沿热点,因此,选取领域的高关注量文献集进行共被引分析、耦合分析;考虑到经典的引文量与新兴的Altmetrics指标在文献影响力评价上的相互独立性与互为补充性,分别筛选出高关注、高被引文献集和高关注、低被引文献集进行词频分析,并在混合这两种文献集的基础上进行共词分析。

2 关注度计量工具选择

综合使用范围与发展成熟度,目前主流的4款Altmetrics 计量工具分别是:Plum Analysis、Altmetrics.com、ImpactStory 和 PLoS ALM,4种Altmetrics 计量工具的优势与劣势如表2所示。Altmetrics.com收集了19种原始数据源平台上的替代计量指标值,截至2020年8月,其数据库覆盖了900万个研究产出(包括期刊论文、数据集、图像、白皮书、研究报告等)的超过6000万次提及^[21],此外,其最具颠覆性的地方是——不含任何引用数据,并对不同平台来源的数据与各类指标均赋予了一定权重,通过自动化加权算法给每篇论文一个总的关注量分值(Altmetric Score),这是其特色功能。Plum Analysis 数据收集能力最强,共收集了26种数据源平台上的替代计量指标数值,既包括新兴的替代计量指标,也包括传统的引文指标;涉及的计量对象种类最多,不仅包括期刊论文,还包括书籍、学位论文、博文、专利、数据库、源代码、软件、视频等共计35种,能够较好地覆盖学术成果的各类形式^[22]。ImpactStory 搜集的部分替代计量指标来自 Altmetrics.com 和 PLoS ALM,共涉及17种原始替代计量指标数据源平台,存在不支持模糊检索、运行速度较慢等问题^[23]。PLoS ALM 是 PLoS 出版商通过计量自家平台出版的研究论文的表现,来提供一组 Article-Level Metrics 的指标数据,共覆盖23种产生替代计量指标数值的数据源平台^[24]。

表2 Altmetrics 计量工具的优势与劣势

替代计量工具	优势	劣势
Plum Analysis	底层数据覆盖范围最广,数据收集能力最强,指标种类最多。	收费;信息检索功能欠缺。
Altmetrics.com	提供文献的综合影响力指标 Altmetric Score,数据加工最为严格,数据来源可靠性强;对个人用户免费。	评价对象单一,仅对文献的影响力进行测评,不涉及其他成果类型的(如软件)影响力数据;由于无引文数据,不适宜对研究人员的影响力进行测评。
PLoS ALM	免费;除数据下载和管理功能外,还提供数据清洗和标准化的功能。	仅提供在 PLoS 平台出版的文献的替代计量指标;无可视化。
ImpactStory	将替代计量指标数值进行百分比的转换,对指标进行了归一化处理,便于不同类型指标的比较;提供总影响力测度,适合对单个研究人员进行影响力评价。	收费;无可视化,技术支持文档少。

综合来讲,Altmetrics 计量工具收集多样化数据的能力越强,覆盖的底层数据源越广,得到的文献社会影响力评价结果越客观。因此,选择 Plum Analysis 作为本文测度论文社会影响力的平台。Plum Analysis 推出的 PlumX 指标大致分为5类:提及(Mention)、获取(Capture)、使用(Usage)、社交媒体(Social Media)、引用(Citation)。表3列出了 PlumX 的指标类型和来源。

表3 PlumX 指标类型、名称与数据来源

指标类型	指标名称	数据来源
Usage	Downloads, Views, Collaborators, Book holdings, Document delivery, Clicks, Abstract views, Link outs, Plays, Full text views	Airiti Library, Bepress, Dryad, DSpace, EBSCO, ePrints, Figshare, Github, Institutional Repositories, CABI, OJS Journals, PLOS, PubMedCentral, WorldCat Vimeo, YouTube, SoundCloud 等
Capture	Bookmarks, Followers, Forks, Readers, Favorites, Exports/ Saves, Watchers, Subscribes, Groups	Slideshare, Youtube, CiteULike, Delicious, GitHub, Mendeley, SoundCloud, Goodreads, EBSCO 等
Mention	Blog mentions, Comments, Forum topic count, News mentions, Q&A Site Mentions, Blog posts, Wikipedia articles, reviews	Facebook, Reddit, Slideshare, Vimeo, Youtube, GitHub, Wikipedia, Amazon, Goodreads, SourceForge 等
Social Media	Tweets, +1's, Likes, Shares, Recommendations, Ratings	Tweets, Facebook, SourceForge, Vimeo, YouTube, Amazon, Goodreads, Reddit, Figshare 等
Citation	Patent Citations, Citation Indexes, Clinical Citations	USPTO, PubMed, Scopus, CrossRef, RePEc, Scielo, SSRN 等

指标覆盖率,即指标值非空的文献数量占总文献量的比重,是衡量指标应用程度的重要标准之一^[25]。相较于被引量,由于 Altmetrics 指标的种类多样、来源广泛,各类原始替代计量数据平台在大众中的普及程度差距较大,因此,不同类型的 Altmetrics 指标在覆盖率上差异较大,且大部分指标的覆盖率均很低。Alperin^[26]分析拉丁美洲国家的论文发现,仅 Twitter、Facebook、Mendeley 的指标覆盖率就超过了2%。因此,本文筛选指标覆盖率超过2%的 PlumX 指标进行文献的社会影响力即关注量的测度。

3 研究前沿识别方法与结果融合

3.1 基于引用关系的研究前沿识别

基于引用关系研究前沿探测方法包括三类:直接引用、共被引分析和耦合分析法,由于直接引用的探测结果精确度最差,因此,学界更多采用共被引分析和耦合分析法来进行研究前沿识别。共被引分析和耦合分析的核心步骤均是以文献间的共被引及耦合关系为基础,构建引文相似度矩阵,然后引入一定的聚类算法得到引文聚类簇,通过对各簇中论文的标题、关键词进行归纳总结,来为各研究前沿簇命名。其中,论文关注量的计算以 PlumX 指标体系中覆盖率超过2%且非 Citation 类型的13种指标求和得到,高关注文献的遴选借鉴 ESI 高被引论文的选取标准,将关注量排名前1%的论文集定义为高关注论文集。此外,现有的具备共被引分析功能的软件多是从原始论文集的参考文献集中抽取论文间的共被引关系,参考文献集也同时作为施引文献集,这种分析结果明显存在缺陷,论文间的被引关系仅是被部分揭示,超出参考文献集的施引文献及

施引关系不能被展示到最终结果中。本文针对高关注文献集专门进行施引文献追踪下载,以使共被引分析结果更准确。

相比于共被引分析,耦合分析识别到的研究前沿结果准确度更高,且数量更多,在时效性上也更好。但文献间的耦合关系一旦形成不会随时间变化,而文献间的共被引关系随时间变化同样具备变化性,因此,就对引文网络结构的揭示力度而言,共被引分析要强于耦合分析^[27]。

3.2 基于主题词汇的研究前沿识别

基于主题词汇的研究前沿预测主要包括词频分析法、共词分析法和主题模型法。本文主要采用前两种方法进行研究前沿识别。词频分析法主要是通过词频筛选出领域热点词,从而反映出一段时间内受领域学者高度关注的研究主题。共词分析的核心步骤是以文献中的词共现为基础,构建关键词共现矩阵,然后引入一定的聚类算法得到关键词聚类簇。其中,高关注、高被引文献集即关注量排名前1%的论文集与被引量排名前1%的论文集的交集,体现了受学界认可和关注的热点前沿主题;按照帕累托法则^[28],将被引量排名在后80%的文献集作为低被引文献集,高关注量文献集与低被引文献集的交集即高关注、低被引文献集,反映了已经得到学界充分关注,但其学术影响力尚未得到扩散的前沿主题,这类主题具有“睡美人”的潜质。

词频分析法的优点在于直观、便捷,而共词分析法能将碎片化的主题词聚集为主题关联网络,有助于揭示研究领域的内容结构特征。共被引分析、耦合分析、词频分析和共词分析法,在进行研究前沿预测时各有利弊,为求更客观、全面地对实证领域研究前沿进行揭示,本文将在分别采用这4种方法进行研究前沿识别的基础上,采用改进的线性加权融合模型进行结果融合,并对得到的研究前沿按照一定的测算标准进行排序。

3.3 改进的线性加权融合模型

为对基于引用关系和主题词汇分析法得到的预测结果进行综合处理,本文采取的融合公式为: $F=af(x)+bf(y)+cf(z)+d(w)$,式中F为研究前沿预测结果; $f(x)$ 、 $f(y)$ 、 $f(z)$ 、 $f(w)$ 分别为基于共被引分析、耦合分析、词频分析和共词分析法得到的研究前沿预测结果; a 、 b 、 c 、 d 分别为各研究前沿预测方法对预测结果的影响权重值, a 、 b 、 c 、 d 的和为1。不同的分析方法所预测的研究前沿各有侧重,但也存在相同或相似的情况。因此,最终的研究前沿预测结果是由多个前沿分支及其准确程度组成的。F是融合后研究前沿方向及其前沿可能性分值,数值越高,说明该研究前沿方向越重要、发展潜力越大,未来

成为核心研究前沿的概率越大。

$$F = [abcd] \begin{bmatrix} f(x) \\ f(y) \\ f(z) \\ f(w) \end{bmatrix} = \begin{bmatrix} ax_1 + bx_1 + cx_1 + dx_1 \\ \vdots \\ ax_i + bx_i + cx_i + dx_i \\ by_1 + cy_1 + dy_1 \\ \vdots \\ by_j + cy_j + dy_j \\ cz_1 + dz_1 \\ \vdots \\ cz_k + dz_k \\ dw_1 \\ \vdots \\ dw_t \end{bmatrix}^T$$

为了更科学地设置线性融合模型中的权值,引入强制评分法(Forced Decision, FD)。作为确定评价对象权值系数的常用方法之一,强制评分法通过对评价对象间的两两强制对比,来对评价对象的重要性进行打分,包括“0-1”评分法和“0-4”评分法两种类型。“0-1”评分法,即对比的两个评价对象中,更为重要的打1分,另一个打0分;若评价对象与自己相比,则不得分,用“X”表示。“0-1”评分法虽然操作更为简便,但评价对象间的差距难以拉开层次,尤其是对于一些重要性相对较弱但有存在必要性的评价对象并不适宜。“0-4”评分法将参与对比的两两评价对象进行3种层级的划分,最重要的打4分,最不重要的打0分;相对重要的打3分,相对不重要的打1分;同等重要时,各自打2分。强制评分法比较适用于评价对象间差异性不大,且数量不多的情况。

4 人工智能领域实证分析

本研究的论文数据来源于Web of Science(WOS)核心合集的Science Citation Index Expanded (SCI-E)数据库,检索时间范围为2014—2019年,检索时间是2020年5月15日,经过筛选后,共得到58 659篇人工智能(Artificial Intelligence, AI)领域相关SCI论文。

4.1 Altmetrics 指标筛选

如表4所示,选取PlumX指标体系中覆盖率超过2%且非Citation类型的指标,共13种,以这些指标值的和作为文献的关注量。这13种所选指标的指标覆盖率最高的3种是Abstract Views、Readers和Link-outs,取值区间波动最大3种的是Full Text Views、Clicks和Abstract Views,不小于某项指标均值的文献量均在150篇以上。

对AI领域的所有SCI论文进行关注量计算,并按照关注量降序排列,选取排名前1%的文献构建高关注文集,共586篇;在WOS平台追踪AI领域论文的被引数量,选取被引量排名前1%的文献构建高被引文集,共586篇;对高关注文集和高被引文集作交集,得到高关注、高被引文集,共41篇;对高关注文集与被引量排

表4 PlumX 指标类型、名称与数据来源

指标类型	指标名称	指标来源	最小值	最大值	均值	指标覆盖率	不小于均值文献量
Usage	Abstract Views	EBSCO	1	31878	112.39	91.57%	4694
Captures	Readers	Mendeley	1	10018	43.30	79.49%	4841
Usage	Link-outs	EBSCO	1	6525	16.56	70.00%	3195
Captures	Exports-Saves	EBSCO	1	2677	14.54	52.32%	2376
Social Media	Tweets	Twitter	1	1572	15.26	27.23%	1133
Usage	Full Text Views	EBSCO	1	25303	112.97	23.14%	983
Social Media	Shares, Likes & Comments	Facebook	1	6252	34.16	9.46%	414
Mentions	Blog Mentions	Blog	1	149	1.43	6.38%	261
Usage	Clicks	Bitly	1	48393	74.63	5.02%	108
Mentions	News Mentions	News	1	108	4.64	3.44%	177
Usage	Full Text Views	PLoS	247	113604	4763.44	2.11%	136
Usage	Full Text Views	PubMed Central	14	14705	736.44	2.11%	166
Captures	Readers	CiteULike	1	20	1.70	2.07%	153

名后 80% 的文集作交集,得到高关注、低被引文集,共 62 篇。

4.2 共被引分析

根据 3.1 节与 4.1 节所述方法,基于筛选的 AI 领域的 586 篇高关注文献,以及文献间的共被引关系,构建高关注量文集的共被引相似度矩阵,矩阵的行列交叉处以文献间的共被引次数来表示,接着,采用社会网络分析法来揭示文献间的关联关系,通过 Ucinet 工具进行可视化,连续调节阈值直到能获得比较清晰的聚类结果,此时阈值为 10,得到图 1。图中每个点代表一篇论文,通过 WOS 对文献标注的唯一标识符 UT 号表示,连线的粗细表示两篇文献间共被引次数的高低。

从图 1 可见,共被引分析共得到了 AI 领域的 4 大类研究前沿聚类簇。深度学习与智能医疗是规模最大的前沿团簇,共涉及 17 个成员点,其中,关注量超过 1 万次的论文有 4 篇,这些论文涉及的主题包括:人工智能伦理学、深度学习与糖尿病视网膜病变分期研究、深度学习与转录因子、深度学习综述。此外,前沿团簇中深度学习技术在医疗领域的应用还包括:医学图像分析、病变检测、健康信息学、计算生物学、遗传学与基因组学。虚拟现实与中风康复前沿团簇共涉及 7 个成员节点,节点 WOS:000333678100041 的关注量最高,达 36392 次,研究主题是虚拟现实技术在成人脑卒中(又称脑中风)治疗中的应用。

4.3 耦合分析

同理,基于高关注文献间的耦合关系,构建高关注量文集的耦合相似度矩阵,矩阵的行列交叉处以文献间的耦合次数来表示,连续调节阈值到 10,得到比较清晰的聚类结果,如下页图 2 所示。

耦合分析共得到了 AI 领域的 6 大类研究前沿聚类簇。深度学习与智能医疗是规模最大的前沿团簇,共涉及 15 个成员点,其中,关注量超过 1 万次的论文有 3 篇,这些论文涉及的主题包括:深度学习与长短期记忆、机器学习与急性缺血性卒中的预后、深度学习综述。此外,前沿团簇中深度学习技术在医疗领域的应用还包括:脑震荡分类、急性手术疼痛管理、医学图像分析、健康信息学。与虚拟现实疗法相关的研究团簇有 2 个,共涉及 8 个成员节点,其中,关注量超过 1 万次的论文有 2 篇,研究主题分别是虚拟现实技术在成人脑

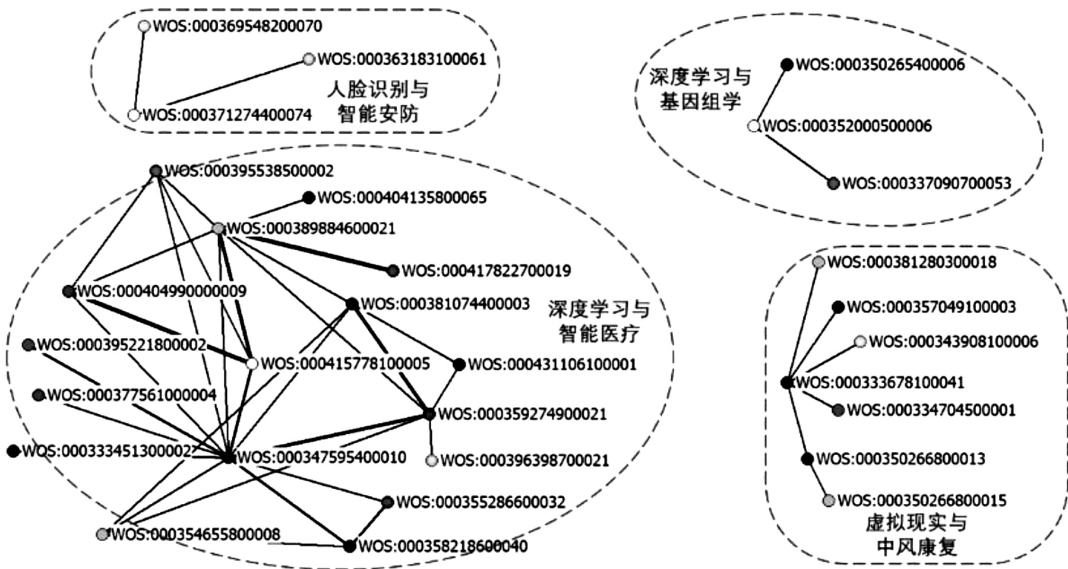
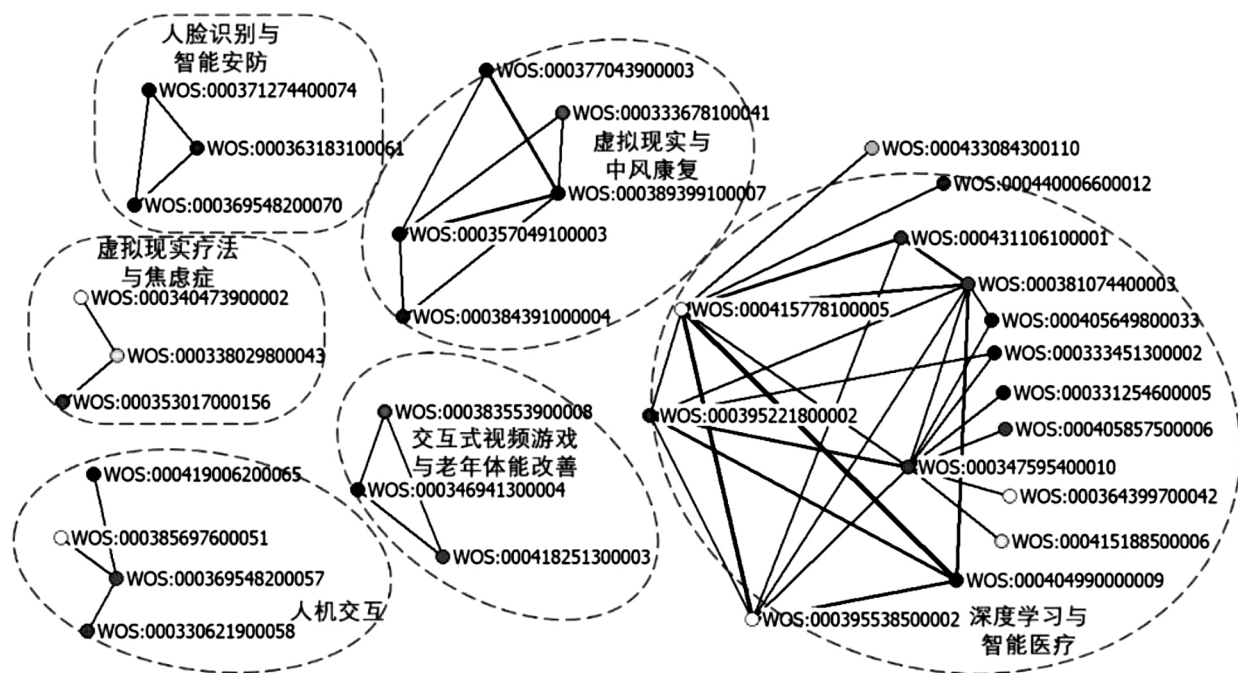


图1 基于共被引分析的2014—2019年人工智能领域研究前沿



卒治疗中的应用、虚拟现实暴露疗法与焦虑障碍分析。人脸识别与智能安防是两种引文分析法均识别到的前沿团簇,与人机交互相关的研究团簇有2个,共涉及7个成员节点。

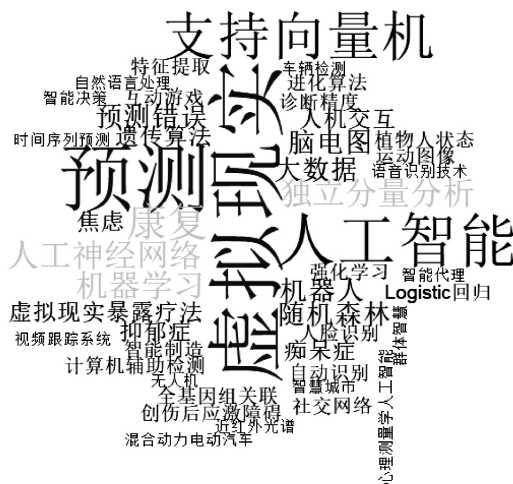
4.4 词频分析

对62篇高关注、低被引文集进行词频统计与热点词云展示。首先,对部分没有作者关键词的文献,使用WOS给定的机标关键词进行补充;然后,对关键词列表中的同义词进行归并,并给出规范化的统一表示,去除无意义的停用词;最后,为使结果易读性更高,对英文关键词进行翻译(以下同)。从图3中可见,高关注、低被引文集识别出的热点关键词包括虚拟现实、预测、人工智能、支持向量机、康复等。

对41篇高关注、高被引文集进行词频统计与热点词云展示,从图4中可见,高关注、高被引文集识别出的热点关键词包括深度学习、机器学习、虚拟现实、中风康复、无监督特征学习等。

4.5 共词分析

对高关注、高被引文集与高关注、低被引文集进行合并,基于合并后的文集进行共词分析。在进行共词分析前,同样需要进行关键词规范化处理、过滤停用词、英文关键词翻译。由图5可见,共词分析得到了6个前沿聚类簇,其中,点的大小表示词频高低,连线粗细表示共现强度前沿聚类簇1中,联系最紧密的词组是



机器学习、卷积神经网络和深度学习,该前沿簇主要是人工智能算法在医学成像、计算机辅助监测方面的应用;前沿聚类簇2中,人工智能是其中最大的点,该前沿簇主要是人工智能在机器人及预测方面的应用;前沿聚类簇3中,虚拟现实是其中最大的点,该前沿簇主要是虚拟现实技术辅助心理类疾病治疗,如焦虑症;前沿聚类簇4主要是通过强化学习等人工智能算法,辅助物理运动来进行康复治疗,如帕金森;前沿聚类5主要是虚拟现实技术,辅助运动障碍类疾病或脑损伤类疾病

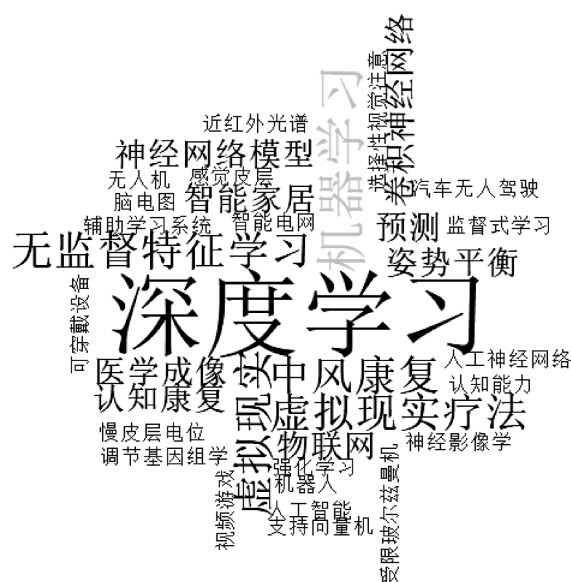


图4 基于词频分析的2014—2019年人工智能领域高关注、高被引主题词云

进行运动恢复;前沿聚类6主要是以智能游戏的形式,应用于缓解或治疗老年痴呆及中风康复等疾病。

4.6 结果融合与对比

通过强制评分法来确定不同的研究前沿识别方法的权重,以便于后续线性加权融合模型的应用。从表5

可见,共被引分析法(A)、耦合分析法(B)、共词分析法(C)、词频分析法(D)权重系数分别是 0.375、0.375、0.040、0.210。

表5 基于强制评分法的不同前沿识别方法权重确定

分析方法	A	B	C	D	合计	权重系数
A	X	2	4	3	9	0.375
B	2	X	4	3	9	0.375
C	0	0	X	3	1	0.040
D	0	1	3	X	5	0.210
合计					24	1.000

根据3.3节的融合模型,得到人工智能领域2014—2019年准确度排名前20位的研究前沿。从下页表6可见,识别到的准确度排名前20位的研究前沿大致分为两类:一类是各种人工智能算法,包括深度学习、卷积神经网络、机器学习、强化学习、支持向量机、无监督特征学习;另一类是人工智能技术在不同领域的应用,包括虚拟现实、中风康复、人脸识别、智能安防、焦虑症、人机交互、视频游戏、老人体能改善、基因组学、机器人、虚拟现实疗法、计算机辅助检测。

研究前沿准确度最高的虚拟现实技术(Virtual Reality, VR)是一种可以与人的视觉、听觉、触觉等感官建立联系的环境模拟系统,使用户可以与虚拟环境从社交、语言、行动等方面进行沉浸式交互^[29]。因其具有趣味性、有效性、安全性,能够调动患者的主观能动性、可

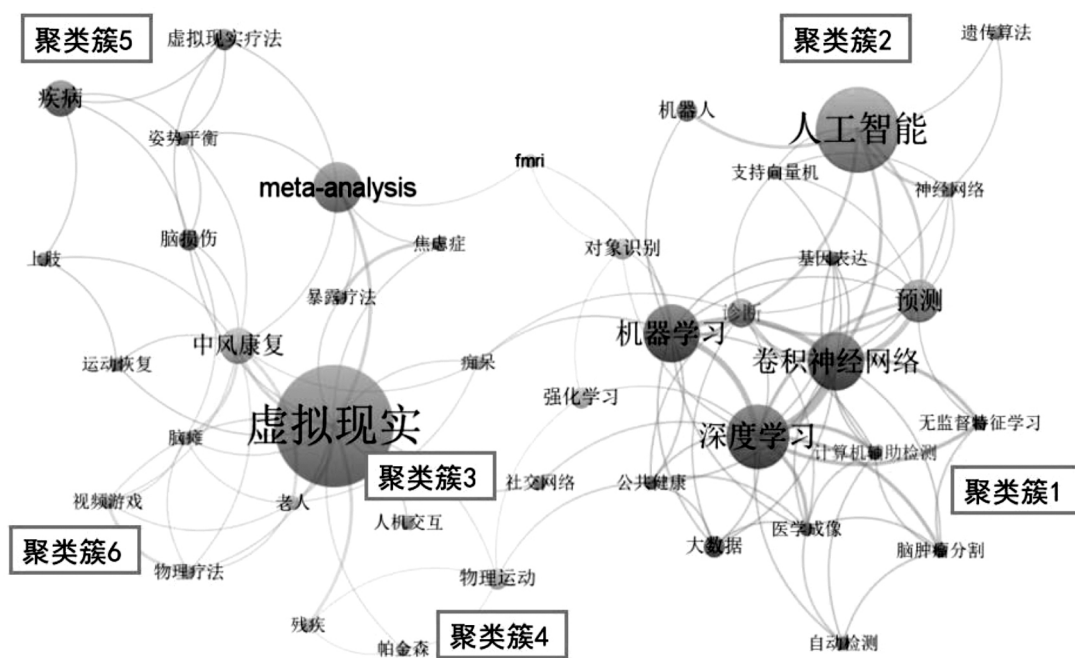


图5 基于共词分析的2014—2019年人工智能领域研究前沿

表6 2014—2019年人工智能领域研究前沿预测结果

序号	研究前沿	准确度	序号	研究前沿	准确度
1	虚拟现实	1	11	卷积神经网络	0.44
2	深度学习	0.98	12	基因组学	0.375
3	中风康复	0.98	13	机器人	0.25
4	人脸识别	0.77	14	机器学习	0.25
5	智能安防	0.75	15	强化学习	0.25
6	智能医疗	0.75	16	人工智能	0.25
7	焦虑症	0.605	17	支持向量机	0.25
8	人机交互	0.605	18	虚拟现实疗法	0.25
9	视频游戏	0.605	19	计算机辅助检测	0.23
10	老人体能改善	0.585	20	无监督特征学习	0.23

重复性、可个性化设置等特点,在医学领域有巨大的应用潜力^[30]。VR在医学领域的应用包括临床培训、康复训练、辅助诊断、心理治理、症状控制等方面,涉及的病症主要集中在神经系统类疾病,如脑卒中、帕金森、多发性硬化、脑瘫、神经病理性疼痛、认知功能障碍等,其他的病症还包括脑损伤、慢性阻塞性肺疾病、听力障碍、脊髓损伤、髋关节置换等^[31]。

通过表7的对比发现,耦合分析得到的前沿团簇比共被引分析要多,这也与张嘉彬^[32]的研究结果类似。此外,共被引分析和耦合分析识别到的前沿结果中有13篇论文相同,占各自识别结果论文总量的40%左右,Jarneving^[33]的研究显示耦合聚类的前沿簇有36%与共被引聚类的前沿簇重合,而共被引聚类的簇类有29%与耦合聚类簇重合,这都证明了操作类似的两种引文分析法,在揭示引文结构特征时存在一定的相似性。

表7 基于共被引分析与耦合分析法的研究前沿重合率

识别方法	阈值次	前沿团簇/个	文献数量/篇	相同文献量/篇	相同文献占比
共被引分析	10	4	30	13	43%
耦合分析	10	6	34	13	38%

为了便于结果对比,对高关注、高被引文集和高关注、低被引文集基于词频分析法识别到的结果均截取其前35个前沿热点主题词,对比两种文献集所得主题的异同,发现相同主题词仅有8个,包括:人工智能、预测、支持向量机、人工神经网络、机器学习、机器人、脑电图、强化学习(见表8)。可见,两种不同视角下研究人员关注的研究主题大部分是不相同的,反映了设置不同视角文献集的必要性。合并高关注、高被引文集和高关注、低被引文集所得70个主题词,删除重复项后共得到62个相异的主题词,与共词网络中涉及的38个主题词进行对比,发现相同的主题词比例大幅上升,达到45%。

表8 基于词频分析法与共词分析法的研究前沿识别结果对比

识别方法		涉及主题词数量/个	相同主题词		相异主题词	
			数量	占比	数量	占比
词频分析	高关注、低被引	35	8	23%	27	77%
	高关注、高被引	35	8	23%	27	77%
共词分析		38	17	45%	21	55%

5 结论

引文分析法与Altmetrics分别代表了不同侧重点下的学术评价的方法,前者侧重于面向专业领域的学术影响力评价,后者则侧重于面向大众评议的社会影响力评价,本文创新性地将二者结合起来,用于研究前沿来源文献集的遴选。此外,本文在各环节的指标选取与平台选择上,进行了充分地对比与严谨地筛选。在关注量与被引量的测算上,由于WOS平台对于期刊和论文引文的加工更为严格,故论文的被引量以WOS平台为准;通过比较Plum Analysis、Altmetrics.com、ImpactStory和PLoS ALM 4款主流的Altmetrics计量工具的优缺点,基于数据收集能力和数据覆盖范围考虑,最终选择了Plum Analysis作为测度论文社会影响力的平台。对于论文关注量的测算,则在选取了PlumX指标体系中覆盖率超过2%且非Citation类型的13种指标后,通过多项指标求和得到。

不同的前沿识别方法所得识别结果有些许差异,为使研究前沿的识别结果更具可靠性和客观性,本文在通过多种方法识别研究前沿的基础上,创新性地采用改进后的线性加权融合模型,来对不同方法所得前沿识别结果进行融合。这种融合算法优点在于不限于被融合方法的类型,能够将不同识别方法所得结果均进行集成,确保了识别结果的全面性。此外,通过给研究前沿赋予一定分值,来反映不同前沿主题的重要性和未来潜力不同,一定程度上缓解对专家智慧的过度依赖,但缺点在于不便于对融合结果进行验证。

本文还对相似的研究前沿识别方法所得结果进行了对比,以分析其异同点。共被引分析共得到了AI领域的4大类研究前沿聚类簇,耦合分析得到了6大类,耦合分析比共被引分析能识别到更多的研究前沿,此外,二者识别到的前沿结果中有40%左右的论文相同,与Jarneving的研究结果相近。高关注、高被引文献集和高关注、低被引文献集识别到的热点前沿主题少部分重合,大部分相异。词频分析法探测到的热点前沿主题虽多,但结果更加碎片化,主题词之间是孤立的,没有联系,而共词分析能够通过一组聚类的词,来对前沿主题进行更明确充分地揭示。

本文的局限性在于研究前沿的识别采用的是单一数据源,并未能结合其他类型数据源,如专利、基金项目等,使得前沿识别的结果不具有全面性,未来将进一步考虑融合多源数据进行研究前沿识别。

参考文献

- [1] 张庆普,陈 茫.Web4.0时代的情报学创新探究[J].情报学报,2016,35(10):1048-1061.
- [2] 邱均平,余厚强.论推动替代计量学发展的若干基本问题[J].中国图书馆学报,2015(1):4-15.
- [3] 刘丽敏,王 晴.我国图情领域 Altmetrics 研究评述及展望[J].情报杂志,2016,35(4):131-136,124.
- [4] 秦 奋,高 健.基于ESI高被引论文信息科学与图书馆领域 PlumX 与引文计量对比分析[J].现代情报,2019,39(3):157-165.
- [5] Noorden R V. Twitter Buzz about Papers does not Mean Citation Later[EB/OL]. [2014-08-19].<http://www.nature.com/news/twitter-buzz-about-papers-does-not-mean-citation-later.1.14354>.
- [6] Kousha K, Thelwall M. Can amazon.com reviews help to assess the wider impacts of books?[J]. Journal of the Association for Information Science and Technology, 2016, 67(3): 566-581.
- [7] Waltman L, Castas R. F1000 recommendations as a potential new data source for research evaluation: a comparison with citations[J]. Journal of the Association for Information Science and Technology, 2014, 65(3): 433-445.
- [8] 邱均平,陈丽婷.高社会影响力论文的特征分析——以2016—2018年 Altmetrics TOP100 论文为例[J].图书馆理论与实践,2019(8):20-25.
- [9] 秦 奋,高 健.基于Scopus数据库的 Altmetrics 指标与引文计量对比分析[J].情报学报,2019,38(4):377-383.
- [10] Price D. Networks of scientific papers[J]. Science, 1965, 149: 510-515.
- [11] Small H G, Griffith B C. The structure of scientific literatures? I: identifying and graphing specialties[J]. Science Studies, 1974, 4:17-40.
- [12] Garfield E. Research fronts[J]. Current Contents, 1994, 41(10): 3-7.
- [13] Persson O. The intellectual base and research front of JASIS 1986-1990[J]. American Society for Information Science, 1994, 45(1):31-38.
- [14] Morris S A, Yen G, Wu Z, et al. Time line visualization of research fronts[J]. American Society for Information Science and Technology, 2003, 54(5):413-422.
- [15] Sujit Bhattacharya, Prajit Basu. Mapping a research area at the micro level using co-word analysis[J]. Scientometrics, 1998, 43(3):359-372.
- [16] Robert Braam, Henk Wed, Van Raan. Mapping of science by combined co-citation and word analysis II: dynamical aspects [J]. Journal of the American Society for Information Science, 1991, 42(4):252, 264.
- [17] Kleinberg J. Bursty and hierarchical structure in streams [J]. Data Mining and Knowledge Discovery, 2003,7(4):373-397.
- [18] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003(3):993-1022.
- [19] Blei D M, Lafferty J. Dynamic topic models[C]. Proceedings of the 23rd International Conference on Machine Learning. New York: Acm, 2006:113-120.
- [20] 刘小平,冷伏海,李泽霞.国际科技前沿分析的方法和途径[J].图书情报工作,2012(12):60-65.
- [21] <https://www.altmetric.com/about-our-data/how-it-works/>.
- [22] 张 洋,叶 月,张宗翔,等.基于GBDT的学术会议替代计量学评价模型研究[J].情报学报,2019,38(11):1150-1159.
- [23] 吴胜男,赵蓉英. Altmetrics 应用工具的发展现状及趋势之分析[J].图书情报知识,2016(1):84-93.
- [24] 金贞燕,侯景丽,孙华丽. Altmetrics 数据整合分析工具的现状特点及相关问题研究[J].情报理论与实践,2019,42(4):89-95, 70.
- [25] 刘晓娟,赵卓娟,宋婉姿. Altmetrics 在学术评价中的可用性研究综述[J].数字图书馆论坛,2017(8):38-45.
- [26] Alperin J P. Geographic variation in social media metrics: an analysis of Latin American journal articles[J]. Aslib Journal of Information Management, 2015, 67(3):289-304.
- [27] 高 楠,赵蕴华,彭鼎原.基于引用关系与词汇分析法的研究前沿识别研究——以人工智能领域为例[J].情报杂志,2020, 39(4):44-50,13.
- [28] 王 玮.帕累托原则与阮冈纳赞图书馆学五定律[J].图书与情报,2002(1):7-9.
- [29] 张珂嘉,王志丹,魏予昕.虚拟现实技术在孤独症患儿中的应用进展[J].护理研究,2020,34(6):1026-1029.
- [30] 宋杨杨,陈校云,张曙欣,等.虚拟现实技术在我国康复医学领域中的研究进展[J].中国医疗设备,2019,34(1):128-132.
- [31] 吴阶平医学基金会,虚拟现实产业联盟,赛迪智库电子信息研究所.虚拟现实医疗应用白皮书(2019年)[N].中国计算机报,2019-12-09(008).
- [32] 张嘉彬.以书目耦合及共被引探讨不同引用区间之研究前沿:以OLED领域为例[D].台湾:台湾大学,2011:20-22.
- [33] Jarneving B. A comparison of two bibliometric methods for mapping of the research front[J]. Scientometrics, 2005, 65(2):245-263.

[作者简介]高 楠,女,1992年生,北京大学信息管理学博士研究生,中国科学技术信息研究所馆员。

周庆山,男,1965年生,北京大学信息管理学教授,博士生导师(通讯作者)。

赵蕴华,女,1967年生,中国科学技术信息研究所研究馆员。

收稿日期:2020-11-06