集成传统学术评价和Altmetrics指标的论文高被引预测研究

吴冰　齐思贤
（同济大学经济与管理学院，上海　200092）

摘要：随着Web 2.0和社交网络的发展，补充学术成果评价的Altmetrics指标应运而生，已有研究表明Altmetrics指标与被引频次之间存在相关性，但集成Altmetrics指标的论文高被引预测研究较少。因此，基于引用理论，将Altemetrics指标与学术层面指标相结合，构建论文高被引预测的指标体系；选取ESI高被引论文榜单，获取2022年4月经济与商业学科高被引论文合集，由此从Web of Science数据库获取论文集相关的学术层面数据，并从Altmetric LLP平台获取论文集相关的Altmetrics指标数据；经过数据清洗和预处理，共得到27 953篇论文数据，对比3种常用机器学习算法的论文高被引预测结果，得到最优的预测模型。研究结果表明：相较于仅使用学术层面指标，引入Altmetrics指标的论文高被引预测效果更优；Altmetrics指标中的在线阅读平台读者数对论文被引频次的影响最大，随后是学术层面指标中的期刊被引半衰期、论文首次被引两年内被引频次、一作总被引频次。研究可以为探究论文高被引的影响因素及其影响程度，完善学术成果的评价体系提供理论依据。

关键词：论文引用；高被引预测；替代计量学；引用理论；机器学习
中图分类号：G353.1；G203　　DOI：10.3772/j.issn.1673-2286.2023.09.004
引文格式：吴冰，齐思贤. 集成传统学术评价和Altmetrics指标的论文高被引预测研究[J]. 数字图书馆论坛，2023（9）：30-37.

学术论文作为科学研究成果的主要形式之一，是公认的知识传播的重要载体。随着科学技术的不断发展，全球范围内发表的学术论文数量逐年指数级增长。对学术论文影响力的评价关系着学术论文的影响力以及学术研究者自身价值，由此成为研究机构或团体研究能力的重要评判标准[1]。

高被引论文和高被引学者近年来引起广泛关注。不少学术服务权威机构依托自身学术数据库推出学术论文高被引影响力相关榜单，其中科睿唯安（Clarivate Analytics）每年发布的ESI（Essential Science Indicator）高被引论文和高被引学者榜单受到全球范围的广泛认可。基于旗下Web of Science数据库中的学术论文和引文数据，Clarivate Analytics构建了学术论文的科学绩效指标ESI，评选出不同学术领域以及学科中被引频次排名靠前的学术论文和学者。由此，国内外很多研究机构和组织都依托ESI，将高被引论文的数量和高被引学者的数量视为科研水平和科研实力的象征[2-3]。与此同时，伴随着Web 2.0的兴起和社交媒体平台的流行，社交媒体和网络平台及时传播与交流科研成果，从而对更广泛的社会公众产生影响，由此Altmetrics应运而生。以在线环境和网络平台的公开数据源为基础的Altmetrics指标可以为度量学术成果的社会影响力提供参考[4-5]，但目前融合Altmetrics指标的高被引影响因素研究较少。

本研究基于引用理论，首先将学术层面指标与Altmetrics指标结合，构建论文高被引预测指标；其次，选取Clarivate Analytics旗下知名的Web of Science数据库，以经济学与商业学科的论文集合为研究对象；最

收稿日期：2023-06-28

后，采用机器学习模型预测论文高被引，旨在充分挖掘论文高被引的影响因素及其影响程度。在理论上，一方面拓展了预测论文高被引的研究视角，另一方面有助于构建和完善学术成果的多维评价体系。在实践上，研究结果可以分别从学术层面和网络传播层面，为提升学术成果综合影响力和完善学术成果综合评价方法提供指导方向。

1 文献综述

1.1 论文高被引相关研究

从被引频次出发，当前对高被引论文的定义和划分采用绝对阈值和相对阈值两种方式[6]。以绝对阈值划分时，固定数值是评价高被引论文的基准，将一时段内被引频次在固定数值以上的论文认定为高被引论文。在绝对阈值划分方式下，判定变得直接高效，但有可能出现学科领域之间高被引论文的分布差异性，因为学术论文引用率高的学科领域会产生大量的高被引论文，而学术论文引用率低的领域内高被引论文会很稀缺。与绝对阈值划分方式不同，相对阈值的判别逻辑是以学术论文所在学科领域为比较范围，将该学科领域内被引频次相对较高的论文视为高被引论文。

随着论文高被引判定标准逐渐明晰，研究者从不同的学科领域出发，在统计分析的基础上，从作者维度、期刊维度和论文维度分析高被引论文的特征[7-8]，以深入探讨论文高被引的影响因素及其原因[9-11]。

1.2 Altmetrics

伴随社交媒体的发展，越来越多的研究人员通过使用社交媒体开展学术活动，Altmetrics应运而生，可以用于衡量学术研究成果的社会影响力[12]。Altmetrics为测度论文影响力提供补充评价指标，其数据来源广泛，有着高社会公众参与度，涵盖博客、新闻网站、政府平台、社交媒体和在线文献管理软件等不同平台，因此Altmetrics指标不仅可以从社交媒体角度解读学术成果的社会影响力，还具有数据源开放、数据获取免费、数据反馈及时和更新速度快的优势[13]。2010年以来，Altmetrics的数据集成服务商以及指标工具逐渐发展，由此研究者开始关注Altmetrics指标，研究和评估Altmetrics指标在衡量学术成果影响力方面的价值。

论文被引领域的Altmetrics相关研究主要有两大方向。①从Altmetrics指标数据出发，将其视为学术成果社会评价的数据来源，以构建评价学术成果影响力的Altmetrics综合指标体系，由此将来自Altmetrics平台的综合评分及其指标数据作为研究对象，验证了Altmetrics指标的合理性[14]。②将Altmetrics指标引入论文被引研究，探究Altmetrics指标与论文被引之间的关系以及影响机制，证实在不同学科领域中Altmetrics指标和论文被引频次呈现出一定的相关性，由此说明Altmetrics指标可以作为传统文献计量指标的重要补充。在此基础上，选取特定期刊的文献，自定义高被引阈值，融合Altmetrics指标和传统文献计量指标预测论文高被引。然而，研究对象的选取范围不广，研究数据的代表性有待提升[15]。

1.3 综述评述

被引频次是学术成果的重要评价指标，高被引论文引起广泛关注，由此围绕论文高被引问题，目前研究主要关注论文、期刊和作者这3个维度涵盖的学术层面影响因素，不断丰富和完善学术层面的指标及其内涵。随着Web 2.0和社交媒体的不断发展，来自网络和社交媒体平台的数据为学术影响力评价提供了补充，由此以Altmetrics指标为代表的社会影响力评价指标为学术成果评价提供了新的视角。虽然已有研究证实Altmetrics指标与论文被引频次之间存在一定的相关性，并进一步应用Altmetrics指标预测论文被引频次，但目前将学术层面因素与社会层面因素结合的论文高被引影响因素研究较少。

因此，本研究从ESI高被引论文出发，基于当前丰富的Altmetrics应用服务带来的开放、丰富和可获取的数据，将Altmetrics指标与论文学术层面的指标相结合，借助机器学习算法[16]，分别从学术层面和社会网络传播层面探究论文高被引的影响因素及其重要性，为完善学术成果的评价体系提供理论依据。

2 基于引用过程概念模型的论文高被引影响因素

2.1 引用过程概念模型

引用理论主要包括规范引用理论和社会建构主义

引用理论[17]。规范引用理论认为引用行为表示对同行的认可，更多的引用意味着更大的认可，由此引用主要取决于引用者对被引文献的感知价值。规范引用理论假设引用出于对同行的认可，但社会建构主义引用理论质疑这一假设的有效性，认为引用是复杂的过程，人们更倾向于引用由学科领域内被认为更权威或更有声望的作者发表的文章，作为研究结果和知识主张的论据支持。

综合规范引用理论和社会建构主义引用理论的实证研究，研究者提出了由三大核心要素组成的引用过程概念模型，包括被引文献、引用过程、施引文献[17]。作为核心要素之一的被引文献包含内容特征、作者特征、期刊特征以及感知价值4个部分，其中感知价值可以分为5类：认知价值、功能价值、条件价值、社会价值和情感价值。认知价值定义为作者对被引文献满足知识需求或信息需求的感知效用；功能价值定义为被引文献对施引文献做出贡献的感知效用；条件价值是指感知效用与社会群体或个人的特征有关；社会价值定义为特定社会群体对被引文献的感知效用；情感价值定义为被引文献引起的积极或消极情感的感知效用。

由此基于引用过程概念模型，针对被引文献这一要素，从内容特征、作者特征、期刊特征以及感知价值4个方面对论文高被引的影响因素展开讨论，其中：内容特征、作者特征、期刊特征为学术层面的影响因素；由于感知价值体现了被引文献的社会影响，可以用表征社会影响的Altmetrics指标来衡量。

2.2 基于引用过程概念模型的学术层面影响因素

在内容维度，论文内容质量是论文被认可的最重要因素，论文的外部特征从形式和内容方面概括和展示了论文的特点，对论文被引情况有一定程度的影响[11]，与此同时，论文的早期被引特征也对论文被引预测有重要作用[18-20]。由此，从特征完备性和代表性出发，选取具有代表性的论文层面的特征[21-22]，包括论文页数、作者数量、参考文献数量、首次被引的时间间隔、首次被引当年被引频次和首次被引两年内被引频次。

在作者维度[23]，作者的学术声誉和影响力对论文早期被关注有重要的影响，尤其第一作者的学术声望和产出[24-25]经常被认为是论文影响力的关键影响因素。随着对作者维度因素的挖掘，实证研究表明论文合作者的学术产出或学术声望对论文的被引和传播也有显著的影响[19]。由此，选取具有代表性的作者指标，统计在当前高被引论文发表之前作者的各维度指标值，包括一作论文数、一作总被引频次、一作H-index、合作者最大论文数、合作者最大总被引频次、合作者最大H-index。

在期刊维度，从双向选择的角度出发，在学术成果的传播过程中学术影响力高的期刊更容易吸引高质量的论文，同样高质量的论文也更倾向于在高影响力期刊上发表。研究发现，期刊的声誉和学术影响力对论文高被引起到决定性作用[10]。由此，选取具有代表性的期刊特征，包括期刊总被引频次、期刊影响因子、期刊五年影响因子和期刊发文数。其中，与影响因子相比，五年影响因子更能反映期刊的长期影响力，因为它考虑了引用时滞。此外，根据Web of Science数据库提供的评价指标，期刊维度的指标还包括期刊即时指数、被引半衰期、特征因子得分和影响力得分。其中，即时指数是指期刊当年发表论文的平均被引水平，衡量了期刊短期内的热度和受关注程度。

2.3 基于引用过程概念模型的Altmetrics指标

以在线环境和网络平台的公开数据源为基础，Altmetrics数据应用服务提供商提供多平台多渠道的数据收集服务，使得Altmetrics指标不断丰富和完善，为度量学术成果的影响力提供了补充性指标。

由于社交网络中的信息传播具有及时性和迅速性，学术成果在社交媒体平台中的被提及量、被收藏量等具有一定相关性[26]，为了避免同类型指标高度相关对论文高被引预测结果的影响，对各类指标进行保留或合并处理，选取具有代表性的Altmetrics指标，具体包括社交平台提及量、百科提及量、在线阅读平台读者数、搜索引擎检索量、开放新闻站点提及量、同行评议平台提及量。

3 数据获取

3.1 学术平台选取

选取Web of Science数据库中的ESI高被引论文为研究对象。首先，Web of Science是全球具有权威性的大型在线文献检索平台，数据库收录了万余种期刊中的

超千万篇论文。ESI一般以10年为计算周期，每两个月更新一次，从各个角度对国家/地区科研水平、机构学术声誉、科学家学术影响力以及期刊学术水平进行全面衡量，由此ESI高被引论文和高被引学者榜单为学界广泛接受和认可，具有权威性和代表性。其次，ESI划分了22个专业领域，根据每个领域的学术论文的被引用情况进行科学排名，提供筛选高被引和高热度论文的各种层次，有助于快速查找特定领域的高被引论文集合。最后，所有ESI高被引论文及其相关信息都可以在Web of Science中快速检索，因此高被引论文具有可得性。

3.2 Altmetrics指标平台选取

选取Altmetric LLP平台提供数据作为论文Altmetrics指标数据的来源。首先，Altmetric LLP平台是目前市场上最大的Altmetrics服务提供商，其旗下产品集成了从众多渠道收集到的数据，可全面衡量学术成果。在此基础上，通过加权将不同数据源集成到一起，得出论文的综合性指标，由此数据涵盖面广且具有代表性。其次，作为最早的Altmetrics服务提供商，Altmetric LLP平台的Altmetric Explorer和Altmetric API分别提供了DOI和PubMed ID等标识符来追踪学术成果，研究者可根据自身的需求申请相应的权限，进而获取所需的Altmetrics指标数据集合，因此数据可得性高。

3.3 论文集合的获取

选取来自Web of Science数据库中经济与商业学科领域的论文集合作为研究对象：一方面是由于这个领域的论文发表数量可观，另一方面是由于这个学科领域与Altmetrics指标表征的网络传播和社会影响紧密相关。

基于2022年4月的ESI榜单，筛选出经济与商业学科领域的高被引论文集合，以Web of Science数据库的主标识符WOS号为检索标识，识别出共3 340篇高被引论文，发表时间范围为2012—2022年，平均每篇被引178次。在此基础上，以DOI为标识，根据高被引论文及其作者信息，通过Web of Science数据库检索获取作者之前发表的所有学术论文；将DOI作为关键字关联Altmetrics指标数据，得到30 916条论文数据记录；进行数据清洗，排查异常值和重复值，最终得到27 953篇论文的数据，其中高被引论文共有4 403篇，非高被引论文共有23 550篇。

4 论文高被引预测

4.1 描述性统计分析

对论文维度的指标进行描述性统计，如表1所示。在论文早期被引特征中，论文首次被引的时间间隔最小值为0，即发表当年即被引用，论文整体首次被引的时间间隔平均为1.57年。论文首次被引当年被引频次平均为3.36次，2020年发表于*The New England Journal of Medicine*的文章“Use of CAR-Transduced Natural Killer Cells in CD19-Positive Lymphoid Tumors”首次被引当年被引频次最高，为164次，因其作者涉及社科领域而被收录至本数据集。论文首次被引两年内被引频次最大为601次，是2021年发表于*Asian Economic Papers*的“The Global Macroeconomic Impacts of COVID-19：Seven Scenarios”。由此可见，与疫情相关的研究引起了社会的广泛关注。

对作者维度的指标进行描述性统计，如表2所示。合作者最大论文数均值是一作论文数均值的近3倍，说明高被引论文的合作者处于持续发表论文的状态。在论文影响力方面，第一作者之间总被引频次差距很大，而H-index基于被引论文数计算，因而标准差相对较小。

对期刊维度的指标进行描述性统计，如表3所示，不同期刊的影响力水平相距甚远。*Nature*期刊总被引频次最大，总被引频次为915 939次。*Scientific Reports*期刊发文数最大，总发文数为21 179篇。

表1 论文维度指标的描述性统计数据

类 别	最小值	最大值	平均值	标准差
论文页数/页	1	198	18.29	12.39
作者数量/个	1	1 313	3.72	8.39
参考文献数量/篇	1	781	59.68	39.18
首次被引的时间间隔/年	0	30	1.57	0.90
首次被引当年被引频次/次	1	164	3.36	4.47
首次被引两年内被引频次/次	1	601	10.87	15.69

表2 作者维度指标的描述性统计数据

类 别	最小值	最大值	平均值	标准差
一作论文数/篇	1	2 508	29.83	56.07
一作总被引频次/次	0	282 904	2 519.64	5 913.71
一作H-index	0	216	13.39	13.42
合作者最大论文数/篇	1	3 090	74.48	119.73
合作者最大总被引频次/次	0	324 883	6 282.41	12 635.28
合作者最大H-index	0	216	25.72	19.88

表3 期刊维度指标的描述性统计数据

类 别	最小值	最大值	平均值	标准差
期刊总被引频次	6	915 939	51 806.32	138 381.45
期刊影响因子	0	91	6.02	6.15
期刊五年影响因子	0	89	7.45	6.79
期刊即时指数	0	259	1.76	9.31
期刊发文数	62	21 179	620.94	2 208.06
期刊被引半衰期	0	37	4.11	5.49
期刊特征因子得分	0	12	0.65	2.27
期刊影响力得分	0	42	10.86	5.64

对Altmetrics层面的6个指标进行描述性统计，如表4所示。Altmetrics指标数据来自多个开放社交媒体和网络平台，不同平台数据的覆盖度各不相同。在社交平台被提及、在开放新闻站点被提及、在搜索引擎被检索、在百科被提及、在在线阅读平台被阅读、在同行评议平台被提及的论文分别有3 806（13.62%）、679（2.43%）、747（2.67%）、27 953（100.00%）、27 953（100.00%）、8 722（31.20%）篇。百科和在线阅读平台数据覆盖度最高，但标准差较大，说明论文在这两个平台上的影响力有较大差异。

表4 Altmetrics指标的描述性统计数据

类 别	最小值	最大值	平均值	标准差
社交平台提及量/次	1	511	2.47	14.75
开放新闻站点提及量/次	1	54	2.32	4.22
搜索引擎检索量/次	1	224	3.42	13.09
百科提及量/次	3	45 804	65.05	582.85
在线阅读平台读者数/个	0	74 303	247.07	699.51
同行评议平台提及量/次	1	656	7.31	21.96

4.2 预测与分析

4.2.1 预测模型应用

以随机的方式将完整的27 953篇论文的数据按7∶3的比例划分成训练集和测试集，模型在训练集基础上学习后，对测试集数据进行预测，并通过k折交叉验证的方式获得稳定的模型效果，以进行模型的评估。

集成模型中，随机森林（Random Forest，RF）、Adaboost和LGBM（Light Gradient Boosting Machine）目前被广泛用于预测论文被引[11]，并且预测效果良好，因此应用这3个模型预测论文高被引。首先，设计两个预测论文高被引的方案：方案一仅使用学术层面指标，方案二组合学术层面指标和Altmetrics指标。接着，比较两个方案预测结果的均方根误差（Root Mean Square Error，RMSE）[27]，发现增加Altmetrics指标后，RF、Adaboost和LGBM的RMSE分别降低了16.8%、12.5%、20.3%。因此，相较于仅使用学术层面指标，结合Altmetrics指标的预测效果更优。最后，根据方案二组合学术层面指标和Altmetrics指标，分别采用RF、Adaboost和LGBM预测论文是否高被引，3个模型的评估指标[17]如表5所示。LGBM对论文高被引的预测效果优于其他两个模型，这是由于LGBM能较好控制模型复杂度，并能同时处理数据稀疏和数据集样本不均衡问题。因此，LGBM将用于Shap值分析，进一步探究各指标对论文高被引的影响程度。

表5 模型的评估比较结果

模 型	数据集	Accuracy	F1 Score	AUC
RF	训练集	0.930 4	0.717 4	0.966 1
	测试集	0.914 9	0.659 2	0.943 0
Adaboost	训练集	0.930 6	0.734 9	0.961 8
	测试集	0.918 7	0.692 0	0.947 8
LGBM	训练集	0.962 5	0.857 7	0.989 7
	测试集	0.925 1	0.712 5	0.954 1

注：AUC，Area Under the Curve。

4.2.2 Shap值分析

Shap值通过衡量特征的边际贡献度对模型进行解释，同时又能以可视化的形式对模型进行全局和局部分析，因此可用于解释模型中各个特征的贡献[28]。

应用LGBM预测论文高被引，得到的Shap值如表6所示。在线阅读平台读者数、期刊被引半衰期、首次被引两年内被引频次对论文高被引的贡献排名前3，Shap值分别为1.41、0.27、0.20。在线阅读平台读者数对预测论文是否高被引的正向影响最大，说明读者数越多，论文越有可能成为高被引论文。Altmetircs指标中，开放新闻站点提及量和搜索引擎检索量对预测结果影响

甚微，Shap值为0.01。百科提及量和参考文献数量的Shap值分别为−0.05和−0.25，这说明百科提及量和参考文献数量都会对论文被引产生负面影响[29-30]。

表6 各指标的Shap值

维 度	指 标	Shap值
论文	论文页数	+0.11
	作者数量	+0.09
	参考文献数量	−0.25
	首次被引的时间间隔	+0.01
	首次被引当年被引频次	+0.01
	首次被引两年内被引频次	+0.20
作者	一作论文数	+0.02
	一作总被引频次	+0.14
	一作H-index	+0.05
	合作者最大论文数	+0.03
	合作者最大总被引频次	+0.11
	合作者最大H-index	+0.06
期刊	期刊总被引频次	+0.08
	期刊影响因子	+0.02
	期刊五年影响因子	+0.04
	期刊即时指数	+0.02
	期刊发文数	+0.03
	期刊被引半衰期	+0.27
	期刊特征因子得分	+0.01
	期刊影响力得分	+0.09
Altmetrics	社交平台提及量	+0.07
	开放新闻站点提及量	+0.01
	搜索引擎检索量	+0.01
	百科提及量	−0.05
	在线阅读平台读者数	+1.41
	同行评议平台提及量	+0.04

4.3 研究结论与实践建议

4.3.1 研究结论

首先，研究组合学术层面指标和Altmetrics指标，提升了对论文高被引的预测效果，这是由于Altmetrics指标代表着论文在社会网络中的传播影响力。尤其对于经济与商业学科领域的研究论文，通过利用社交媒体促进社会层面的传播与交流，可提高被引用的可能性。

其次，研究发现Altmetrics指标中的在线阅读平台读者数的Shap值最大，随后是期刊被引半衰期、论文首次被引两年内被引频次、一作总被引频次。在线阅读平台读者数是论文高被引的最重要影响因素，这是因为在线阅读平台的研究者数量众多，阅读数越多意味着社会影响越大，论文也就更容易被研究者关注并引用。

再次，研究数据集合中的论文在社交媒体、开放新闻站点、同行评议平台、搜索引擎中的数据覆盖度虽然不足100%，但是相关指标对论文被引仍产生一定的影响，并且其影响程度与数据覆盖度为100%的论文首次被引的时间间隔、论文首次被引当年被引频次和期刊特征因子得分相同。由此说明，在开放新闻站点或搜索引擎中被提及或检索对论文而言十分重要。

最后，Altmetrics指标中，百科提及量对论文高被引有负面影响，这可能是由于百科提及有一定的滞后性，纳入的论文未及时受到关注而未能被进一步引用。与此同时，参考文献数量对论文高被引也有负面影响，这说明高被引论文通常引用适当数量的参考文献。

4.3.2 实践建议

在提升学术成果综合影响力方面，无论是学者个人还是研究团体、研究机构，都应重视在社交媒体和在线阅读平台上的交流，打通学术平台与社交媒体平台，由此扩大在社会网络传播方面的影响力，提高综合的知名度，以进一步提升在学术界的影响力。

在完善学术成果综合评价方法方面，除了需要关注学术层面的评价指标，还应关注学术成果在社交媒体平台中的传播力和影响力。学术成果在社会群体中产生的影响有着累积效应，由此应将社会大众对学术成果的关注度纳入评价范围，进一步完善学术成果评价体系，识别并传播更具价值和影响力的学术内容。

5 结语

在研究视角方面，以往研究集中关注学术层面特征。随着Altmetrics的发展，虽然已有研究在构建Altmetrics指标体系的基础上研究网络传播对论文被引的影响，但是将学术层面指标与Altmetrics指标相结合，研究论文高被引影响因素的研究较少。因此，本研究基于引用理论，整合学术层面指标与Altmetrics指标，构建预测论文高被引的综合指标体系，并通过实证研究说明了集成Altmetrics指标的必要性和可行性。

在研究数据方面，以往研究通常单独使用Web of Science数据库或Altmetrics平台数据，本研究选取Web of Science数据库作为学术层面指标的数据来源，

选取Altmetric LLP平台作为Altmetrics指标的数据来源，由此整合学术数据源与网络传播数据源，促进了领域的融合。

在研究数据集方面，以往研究通常针对某一特定期刊或者特定年份的论文集，因而数据集的规模和代表性都存在一定的局限性。本研究选取经济与商业学科ESI高被引论文集，反向获取高被引论文第一作者的所有学术论文，将两个集合集成在一起作为研究对象，使得数据集更具多样性和代表性，由此更好地发挥机器学习算法的优势，有效探究论文高被引的影响因素及其影响程度。

本研究的不足之处包括以下两个方面。首先，主要选取了Web of Science数据库中的经济与商业学科论文，未来研究可以选取来自不同领域、由不同学术数据库收录的论文，以获得普适性的结论。其次，Altmetrics指标数据主要来自Altmetric LLP平台，虽然这个平台提供了较为全面的网络传播开源数据，但是平台仅提供2010年后的公开数据，未来研究可以考虑根据论文发表时间段动态建模。

参考文献

[1] AMJAD T, REHMAT Y, DAUD A, et al. Scientific impact of an author and role of self-citations[J]. Scientometrics, 2019, 122: 915-932.

[2] 段丹，梁柏静，于文文，等. 基于Altmetrics视角的学术论文被引频次影响因素分析和预测[J]. 图书馆杂志，2020，39（4）：102-112.

[3] FISCHER A. Research evaluation and scientific publications: quantity or quality?[J]. Bulletin De L Academie Nationale De Medecine, 2022, 206（7）: 898-901.

[4] 宋丽萍，王建芳，付婕，等. Altmetrics对研究的社会影响力评价效果研究[J]. 信息资源管理学报，2022，12（5）：123-129.

[5] PERES M F P, BRASCHINSKY M, MAY A. Effect of Altmetric score on manuscript citations: a randomized-controlled trial[J]. Cephalalgia, 2022, 42: 1317-1322.

[6] 石丽，秦萍，李小涛. 高被引论文的跨学科性与Altmetrics指标相关性分析[J]. 情报理论与实践，2021，44（5）：60-65，91.

[7] DORTA-GONZÁLEZ P, SANTANA-JIMÉNEZ Y. Characterizing the highly cited articles: a large-scale bibliometric analysis of the top 1% most cited research[J]. Malaysian Journal of Library & Information Science, 2019, 24（2）: 23-39.

[8] HO Y S, SHEKOFTEH M. Performance of highly cited multiple sclerosis publications in the Science Citation Index expanded: a scientometric analysis[J]. Multiple Sclerosis and Related Disorders, 2021, 54: 103112.

[9] LEE D H. Author-related factors predicting citation counts of conference papers: focusing on computer and information science[J]. Electronic Library, 2020, 38: 463-476.

[10] JONES S, ALAM N. A machine learning analysis of citation impact among selected Pacific Basin journals[J]. Accounting & Finance, 2019, 59（4）: 2509-2552.

[11] ABRISHAMI A, ALIAKBARY S. Predicting citation counts based on deep neural network learning techniques[J]. Journal of Informetrics, 2019, 13（2）: 485-499.

[12] WOOLDRIDGE J, KING M B. Altmetric scores: an early indicator of research impact[J]. Journal of the Association for Information Science and Technology, 2019, 70（3）: 271-282.

[13] KARIMIPOUR N, SARKISYAN A, SMITH K E, et al. The Altmetric era in eating disorder research: assessing the association between Altmetric scores and citation scores for articles in the International Journal of Eating Disorders[J]. The International Journal of Eating Disorders, 2020, 53（12）: 2073-2078.

[14] AKELLA A P, ALHOORI H, KONDAMUDI P R, et al. Early indicators of scientific impact: predicting citations with altmetrics[J]. Journal of Informetrics, 2021, 15（2）: 101128.

[15] DEMACHKI É, DE MELO MARICATO J. Coverage of data sources and correlations between altmetrics and citation indicators: the case of a Brazilian portal of open access journals[J]. Serials Review, 2022, 48（1/2）: 151-166.

[16] ALOHALI Y A, FAYED M S, MESALLAM T, et al. A machine learning model to predict citation counts of scientific papers in otology field[J]. BioMed Research International, 2022, 2022: 1-12.

[17] TAHAMTAN I, BORNMANN L. Core elements in the process of citing publications: conceptual overview of the literature[J]. Journal of Informetrics, 2018, 12（1）: 203-216.

[18] 王超，马铭，李思思，等. Altmetrics视角下颠覆性技术的社会影响力探测研究[J]. 情报理论与实践，2022，45（1）：93-104.

[19] VAGHJIANI NILAN G, VATSAL L, NIMA V, et al. Social media and academic impact: do early tweets correlate with future citations?[J]. Ear, Nose, & Throat Journal, 2021: 1455613211042113.

[20] GELZER EMILY R，LAFORGE MICHEL P，BECKER JUSTINE A，et al. Getting cited early：influence of visibility strategies，structure，and focal system on early citation rates[J]. The Journal of Wildlife Management，2022，86（4）：e22214.

[21] SEDAGHAT A R. Distribution of article citation frequency，citation skew，and impact factor in otolaryngology journals[J]. Otolaryngology-Head and Neck Surgery，2023，168（1）：101-104.

[22] ABRAMO G，D'ANGELO C A，FELICI G. Predicting publication long-term impact through a combination of early citations and journal impact factor[J]. Journal of Informetrics，2019，13（1）：32-49.

[23] HOOD A S C，SUTHERLAND W J. The data-index：an author-level metric that values impactful data and incentivizes data sharing[J]. Ecology and Evolution，2021，11（21）：14344-14350.

[24] THELWALL M. Are successful co-authors more important than first authors for publishing academic journal articles?[J]. Scientometrics，2023，128（4）：2211-2232.

[25] HWANG A，ARBAUGH J B，BENTO R F，et al. What causes a business and management education article to be cited：article，author，or journal?[J]. The International Journal of Management Education，2019，17（1）：139-150.

[26] 高楠，宋官钰，徐少明. Altmetrics指标与引文量相关性分析：以LIS领域为例[J]. 情报理论与实践，2022，45（5）：53-60.

[27] HODSON T O. Root-mean-square error （RMSE） or mean absolute error （MAE）：when to use them or not[J]. Geoscientific Model Development，2022，15（14）：5481-5487.

[28] GHOSHROY D，ALVI P A，SANTOSH K C. Unboxing industry-standard AI models for male fertility prediction with SHAP[J]. Healthcare，2023，11（7）：929.

[29] BINMAKHASHEN G M，AL-JAMIMI H A. Evaluation of machine learning to early detection of highly cited papers[C]//2022 7th International Conference on Data Science and Machine Learning Applications（CDMA）. 2022：1-6.

[30] MENG Y，YANG N H，QIAN Z L，et al. What makes an online review more helpful：an interpretation framework using XGBoost and SHAP values[J]. Journal of Theoretical and Applied Electronic Commerce Research，2021，16（3）：466-490.

作者简介

吴冰，女，博士，副教授，研究方向：信息管理与信息系统，E-mail：ww_bing@163.com。
齐思贤，女，硕士，研究方向：信息管理与信息系统。

Research on High Citation Prediction of Papers by Integrating Traditional Academic Evaluation and Altmetrics Indicators

WU Bing　QI SiXian
(School of Economics and Management, Tongji University, Shanghai 200092, P. R. China)

Abstract: With the development of Web 2.0 and social networks, Altmetrics indicators have emerged as supplementary evaluation of academic achievements. Previous studies have shown that there is a certain correlation between Altmetrics indicators and citation frequency, but there is limited research on high citation prediction of papers that integrates Altmetrics indicators. Therefore, based on the citation theory, this study combines Altmetrics indicators with academic indicators to construct an indicator system for predicting high citation in papers. Then, the ESI highly cited paper list is selected to obtain highly cited paper collection of April 2022 in economic and business discipline. Thereby, academic data relating with the paper collection are obtained from the Web of Science database, and Altmetrics indicator data relating with the paper collection is obtained from the Altmetric LLP platform. After data cleaning and preprocessing, data of 27 953 papers are obtained, and the evaluation results of three common machine learning algorithms are compared to get the optimal model. The research results indicate that compared with using academic indicators alone, the integration of Altmetrics indicators yields better prediction of highly cited papers. The number of readers on online reading platforms has the greatest impact on the citation frequency of papers among Altmetrics indicators, followed by the journal's citation half-life, the number of citations within the first two years after a paper's initial citation, and the total number of citations as the first author among academic indicators. This study can contribute to exploring the factors influencing high citation and their respective impact levels, providing a theoretical basis for improving the evaluation system of academic achievements.

Keywords: Paper Citation; High Citation Prediction; Altmetrics; Citation Theory; Machine Learning

（责任编辑：王玮）