doi:10.3772/j.issn.2095-915x.2025.01.011

石油类高校高被引论文特征挖掘及影响因素研究
——以中国石油大学（北京）为例

开放科学（资源服务）标识码（OSID）

刘天琳　陆桃妹　张芹

中国石油大学（北京） 北京 102249

摘要：[目的 / 意义] 为提高石油类高校科研竞争力、国际影响力，梳理石油类高校高被引论文的整体情况，以供同行业高校在提升论文学术质量及实际应用价值方面参考。[方法 / 过程] 以中国石油大学（北京）2018—2023 年高被引论文为研究样本，总结了论文本身、期刊、作者及参考文献类型 4 类 21 个评价因素，综合运用随机森林方法，挖掘高被引论文的特征及影响因素。[结果 / 结论] 在论文本身方面，早期被引量、研究领域较重要；在期刊方面，期刊影响因子则是最为重要的因素；在作者方面，作者声誉反映了作者在学术界的地位和影响力，作者合作规模在一定程度上反映了研究团队的实力和合作能力；在参考文献方面，参考文献平均影响因子、参考文献的累计被引量两个指标需要关注。根据数据特征挖掘，提出了提升论文质量与创新性、优化论文发表策略等 4 方面的提升石油类高校高被引论文数量的策略。

关键词：石油类高校；高被引论文；特征挖掘；影响因素

中图分类号： G35; G255.2

Research on the Characteristics and Influencing Factors of Highly Cited Papers in Petroleum Universities: Taking China University of Petroleum (Beijing) as an Example

LIU Tianlin LU Taomei ZHANG Qin

China University of Petroleum (Beijing), Beijing 102249, China

Abstract: [Objective/Significance] In order to improve the scientific research competitiveness and international influence of petroleum universities, the overall situation of highly cited papers in petroleum universities is sorted out, so as to provide

基金项目　中国石油大学（北京）研究生教育质量与创新工程重点项目子课题“AIGC 时代学术诚信教育与 AI 学术伦理研究（yjs2024018）。

作者简介　刘天琳（1992-），硕士，馆员，主要研究方向为石油情报及知识产权信息分析；陆桃妹（1992-），硕士，馆员，主要研究方向为图书情报、知识产权分析，Email：1016195164@qq.com；张芹（1981-），博士，副研究馆员，主要研究方向为石油情报分析。

引用格式　刘天琳，陆桃妹，张芹 . 石油类高校高被引论文特征挖掘及影响因素研究——以中国石油大学（北京）为例 [J]. 情报工程，2025, 11(1): 119-127.

doi:10.3772/j.issn.2095-915x.2025.01.011

reference for the same industry universities in improving the content, academic quality and practical application value of papers. [Methods/Processes] The author took the highly cited papers of China University of Petroleum (Beijing) in the past six years from 2018 to 2023 as research samples, and summarized twenty-one evaluation factors in four categories: the papers themselves, journals, authors and reference types. The random forest method was comprehensively applied to explore the characteristics and influencing factors of highly cited papers. [Results/Conclusions] The results show that: In terms of the papers themselves, the early citation amount and research field are more important; In terms of journals, journal impact factor is the most important factor. As for the author, the author reputation reflects the status and influence of the author in the academic circle, and the scale of the author cooperation reflects the strength and cooperation ability of the research team to some extent. In terms of references, the average impact factor of references and the cumulative number of references need to be paid attention to. Based on data feature mining, the paper puts forward four strategies to improve the number of highly cited papers in petroleum universities, including improving the quality and innovation of papers and optimizing the publishing strategy.
Keywords: Petroleum Universities; Highly Cited Papers; Feature Mining; Influencing Factors

引言

在全球科研竞争日益激烈的背景下，高被引论文作为衡量科研影响力和学术质量的重要指标，受到学术界广泛关注。根据基本科学指标（ESI）数据库的界定，高被引论文指近十年间累计被引用次数进入各学科世界前 1% 的论文。高被引论文评价受多种因素共同影响，根据现有研究，论文本身、被引用论文的作者、摘要、期刊、领域和参考文献等是影响论文被引频次的主要因素 [1-2]；部分探究影响因素与被引频次的相互关系，如呈线性、非线性、U 型关系等 [3-4]。另外，研究课题新颖程度、文章类型、不同学科、国际合作关系也有影响 [5-6]。

现有研究成果是从全局角度开展高被引论文影响因素评价及被引量预测，研究样本并没有针对具体科研机构、具体专业领域，更没有涉及石油领域高校的深入分析，特别是在能源战略性领域的研究尤为不足。作为全球最大的能源消费国和生产国，中国的石油能源安全对国家发展具有重大战略意义。石油领域高校在保障国家能源安全、推动能源科技创新方面肩负着重要使命。在此背景下，深入研究石油领域高校的高被引论文特征及其影响因素，对于提升科研核心竞争力、增强国际学术影响力具有重要的现实意义。

基于此，本文聚焦石油行业重点高校——中国石油大学（北京），通过构建科学评价指标体系，深入分析高被引论文的特征及其影响因素。研究采用机器学习方法，系统挖掘影响高被引论文的关键因素，并在此基础上提出培育高被引论文的有效路径。本文不仅填补了石油领域高被引论文研究的空白，也为提升石油领域高校的科研水平和国际影响力提供了理论依据和实践指导。

1 中国石油大学（北京）ESI 高被引论文概况

1.1 数据来源

本文以中国石油大学（北京）2018—2023

年去除重复论文后的 223 条高被引论文作为分析样本，在 SCI、INCITES 网站中进行相关指标数据下载。

1.2 高被引论文特征分析

中国石油大学（北京）作为石油领域的重点高校，正在全面构建能源特色鲜明学科体系，争创世界一流学科，高被引论文呈现 5 个方面的特征：（1）高被引论文数量规模增长，2023 年该校产出 216 篇高影响力论文，总体数量比 2018 年增长 18%；（2）论文影响广泛，被中国、美国、印度、澳大利亚等 23 个国家或地区引用，国内、国外引用比为 7:3，自引、他引比约为 1:9；（3）优势学科明显，高影响力论文产出于本校优势学科，尤其以工程学、地球科学、化学三大学科产出最多；另外，在物理学、经济与商业、临床医学、生物与生化等学科中也有少量高影响力论文产出；（4）论文整体质量较高，高达 84% 的高影响力论文属于中科院分区 1 区；（5）国际合作广泛，在高影响力论文中，国际合作论文占比 52%，与 210 个国际机构有论文合作，且大多数是与美国等发达国家的国际合作。

通过对发文关键词进行聚类分析，可以发现高被引论文主要围绕多相热流体与热传递特性、非冷凝性气体与性能分析、分离与吸附、优化与模型恢复以及其他相关主题进行聚类，如图 1 所示。这些主题共同构成了石油领域的重要研究内容和技术挑战，主要研究方向包括：pore structure（孔隙结构）、photocatalysis（光催化）、shale gas（页岩气）、pyrolysis（热解）、natural gas hydrate（天然气水合物）、CO_2 emissions（二氧化碳排放）、heat-transfer characteristics（传热特性）等。例如，图中红色圆圈展示的与页岩气（shale gas）研究相关的关键词网络，中心是 pore structure（孔隙结构），这表明它是主要的研究领域。与“pore structure”直接相连的关键词包括“permeability”（渗透率）、“porosity”（孔隙度）、“fractal dimension”（分形维度）和“sorption”（吸附作用）。这些都与岩石物理性质和气体在多孔介质中的行为有关。其他重要术语如“mississippian barnett shale”（密西西比巴尼特页岩）和“ordos basin”（鄂尔多斯盆地）则指出了研究区域。关键词以不同深浅的红色表示，代表这些研究方向的重要性及研究频率，线条表示关键词之间的关联性，线条越粗意味着两个关键词之间的关系越密切。总体来看，反映了关于页岩气研究中各个关键领域的相互联系，强调了理解孔隙结构和其相关属性对于页岩气开采和应用的重要性。

为了从时间维度理清研究重点领域的演化过程，分析了关键词时区分布图，展示学校研究关注点的时间序列，如图 2 所示，该图能够从整体上反映研究路径的变化。图中圆圈大小代表该研究主题在时间段内的重要性和流行度。每个时间段展示的均为该时段内首次涌现的关键词，若这些新关键词与先前时间段中的关键词在同一篇文章中共现，则会通过线条相连，且前期关键词的计数增加 1，其对应的圆圈随之扩大。图 2 展示了 2010—2023 年关键主题之间的关系网络。

doi:10.3772/j.issn.2095-915x.2025.01.011

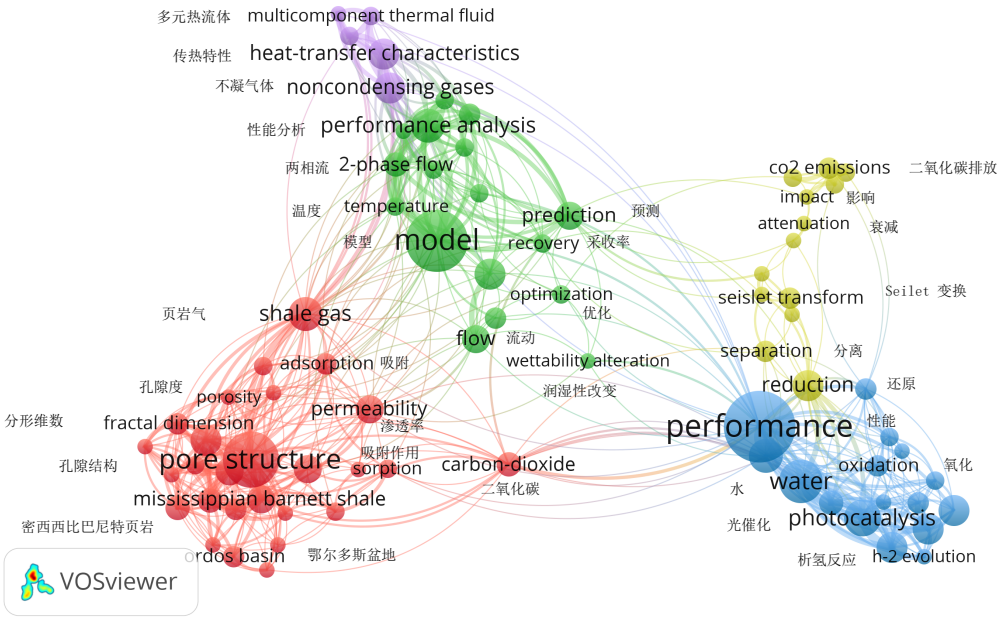

图 1　关键词聚类

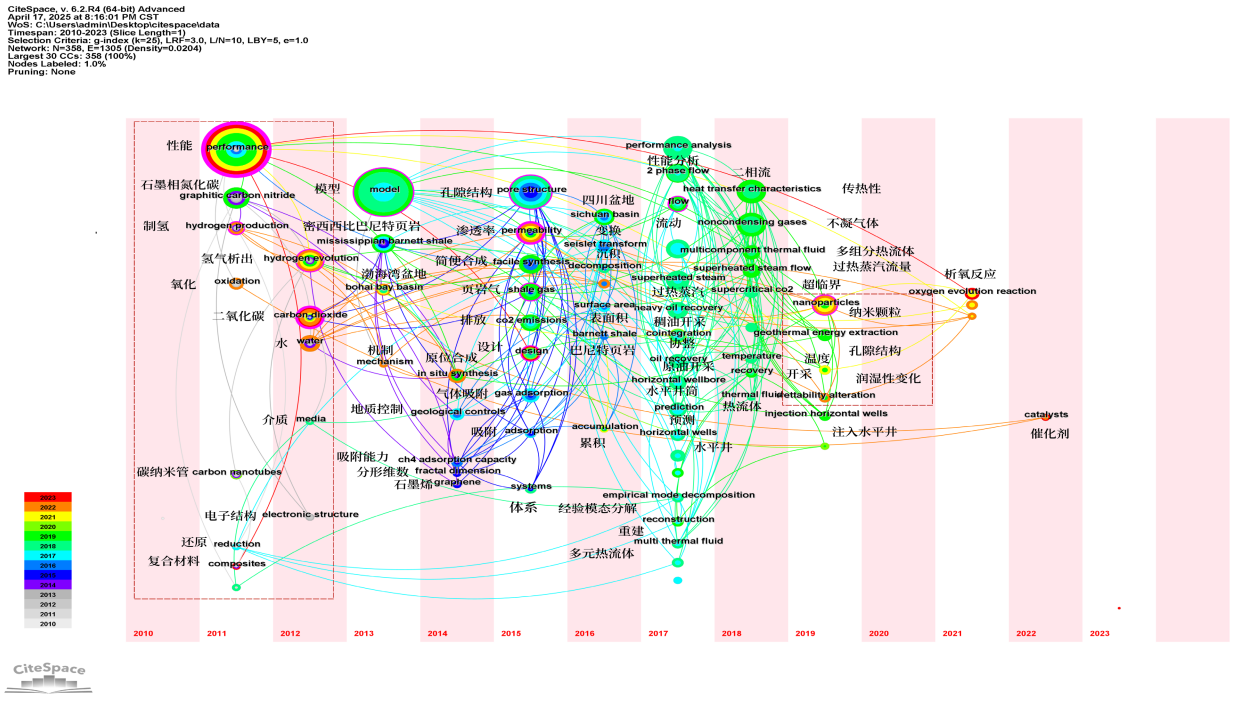

图 2　关键词时区图

进一步通过聚类分析，评价主题之间的内在联系和相互作用，这些主题涵盖了从研究理论、实验技术到实际应用等多个方面，为石油能源领域的研究和发展提供了重要的参考。例如图 2 中一个 2011 年出现的突出关键词“performance”（性能），该关键词对应的圆圈，由圆心向外逐步扩大，且由不同颜色组成。对照年份色标可以看出，这个关键词

在 2011 年至 2023 年间受关注程度的变化。可以看到，“performance”在 2019 年（绿色色标）和 2020 年（浅绿色标）期间尤为突出，表明这段时间对其研究的兴趣达到高峰。该词与同年出现的“composites”（复合材料）连线为红色，表明了研究的重点应用和扩展方向。“performance”在后续年份的圆形面积虽有所波动，但仍然保持在一定的重要位置；另外一个关键词“nanoparticles”（纳米颗粒）在 2019 年（绿色色标）开始出现，并在 2021 年（黄色色标）变得更加突出。这表明纳米颗粒作为一个新兴的研究主题正在逐步获得更多的关注。与“nanoparticles”相关的关键词“wettability alteration”（润湿性变化），揭示了纳米颗粒改变润湿性的可能性。

2 数据来源与指标选择

2.1 指标选择与数据预处理

石油领域论文具有“侧重解决实际生产问题、跨学科融合并涉及新技术新方法新材料、关注国际合作和交流”等特点，针对可以定量化的指标对 223 条高被引论文进行数据预处理，最终从论文本身、参考文献、作者、期刊 4 个方面，共选取 21 个定量指标。

2.2 基于随机森林的预测模型构建

传统单一模型往往难以应对复杂数据集中的噪声和多样性，而机器学习方法能够有效地解决这一问题，因此本文采用机器学习方法预测高被引论文的年均被引量。在众多的机器学习方法中，集成学习是一种有效提高模型性能的方法，其中，随机森林算法（Random Forest）因其简单易用、效果显著而成为常用的方法之一。该方法通过构建多个决策树并将其结果进行综合，从而完成高效的分类或回归任务。由于其在处理高维数据、缺失值和非线性关系方面的优越性能，随机森林已经广泛应用于生物信息学、金融预测、图像识别等领域。

表 1 指标选择

指标	描述
论文	研究领域：wos 学科分类个数
	论文总页数
	发表论文时间
	摘要字数
	关键词数量
	论文早期被引量：论文发表后两年内被引总次数
	是否有基金资助
	类型：综述性论文、研究型论文
参考文献	论文所引用的参考文献数量
	参考文献质量：参考文献期刊平均影响因子
	参考文献的年龄：参考文献的出版年份均值
	参考文献的累积被引：截至计算日期参考文献的累积被引
作者	作者声誉：第一作者 h 指数
	作者合作规模：参与作者数量
	国家合作规模：国家数量
	合作方式：国际合作、国内合作、组织内合作、组织外合作
	自引率：论文作者对论文的被引量占比
期刊	期刊质量：期刊影响因子
	是否开放获取
	发文量：发文当年刊载的论文数
	期刊的语言类型

随机森林算法由 Breiman[10] 于 2001 年提出，其核心思想是通过构建多个决策树并结合其投票或平均结果来降低模型的方差。基于算法的

doi:10.3772/j.issn.2095-915x.2025.01.011

基本设计原则，通过 Python 语言编写预测模型，流程图如图 3 所示。选择 80%（178 篇）的数据建立训练集，利用剩余 20%（45 篇）数据进行验证，评价模型准确性。

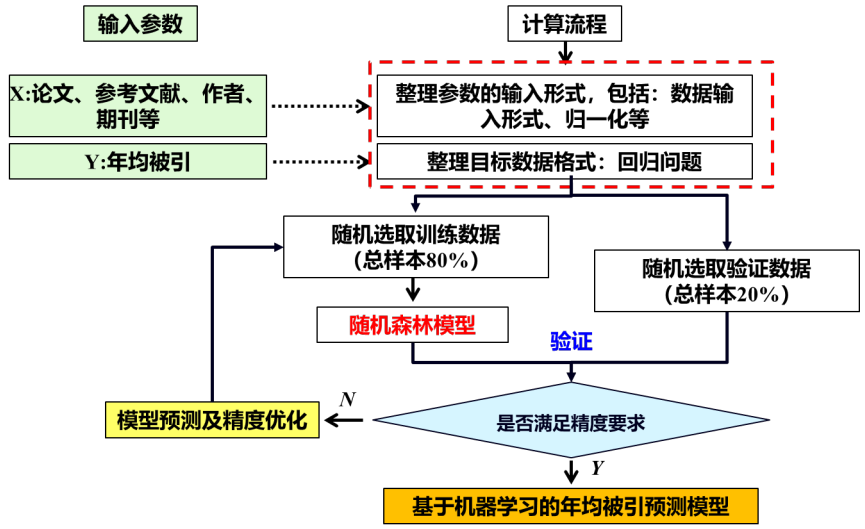

图 3 机器学习预测流程

预测值与实际值的比较如图 4 所示，图中虚线为相对误差 ±15% 的范围，实线为误差 0 的标准线，验证数据处于误差范围内，证明该模型能够准确预测论文的年均被引量。

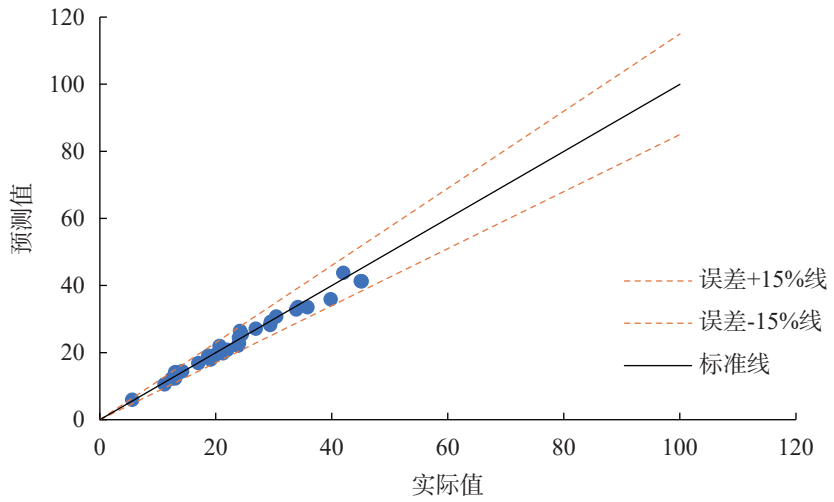

图 4 论文的年均被引量预测结果

3 影响因素研究

3.1 影响因素权重分析

基于随机森林预测结果，对 21 个指标进行重要性排序，如图 5 所示。在论文方面，论文早期被引量、论文的研究领域比较重要。论文早期被引量反映了论文发表后的影响力和被认可程度。在石油领域，早期被引量反映了论文

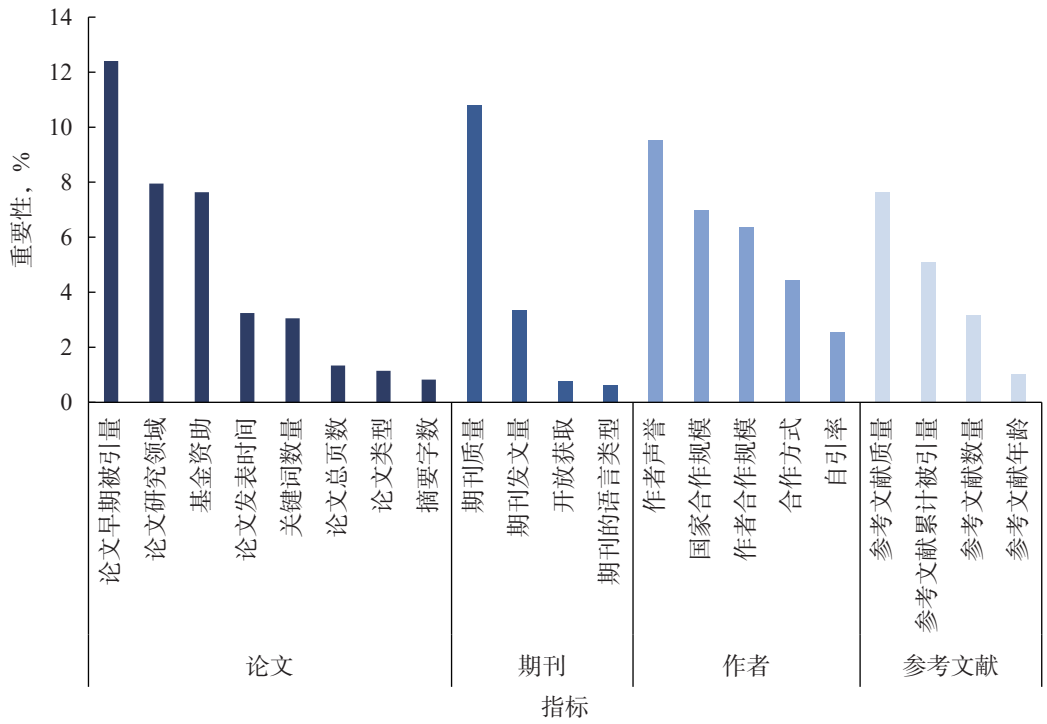

图 5　指标重要性

在发表后迅速被学术界和工业界关注的程度。石油领域的研究往往具有较强的应用导向，因此早期被引量高的论文通常涉及前沿技术或解决行业关键问题；石油领域的研究方向，如油气勘探、油田开发、石油化工等，对论文的被引量有显著影响。具有创新性和突破性的研究，如页岩气开发、深海钻井技术更容易受到广泛关注和引用。

在期刊方面，期刊影响因子是最为重要的因素。石油领域的高质量期刊（如《Journal of Petroleum Science and Engineering》）往往能够吸引高水平的论文投稿，并提升论文的学术价值和实践价值；石油领域的开放获取期刊能够加速研究成果的传播，特别是在国际合作中，开放获取有助于提升论文的可见性和引用率。

在作者方面，作者声誉反映了作者在学术界的地位和影响力，在石油领域，知名学者或研究团队的研究成果更容易被引用；作者合作规模在一定程度上反映了研究团队的实力和合作能力，石油领域的研究往往需要跨学科、跨机构的合作。大规模的合作团队通常能够整合更多资源，提升研究的深度和广度。

在参考文献方面，参考文献平均影响因子、参考文献的累计被引量两个指标需要关注。高质量的参考文献是论文研究深度和广度的体现，也是评估论文学术价值的重要依据。在石油领域，引用高被引文献能够增强论文的学术基础；累计被引量反映了参考文献的影响力和被认可程度，是评估论文研究基础扎实程度的重要指标。

与材料科学、计算机科学等领域相比，石油领域的研究更注重实际应用，因此论文早期被引量和研究领域的重要性更为突出。全球范

doi:10.3772/j.issn.2095-915x.2025.01.011

围内，高被引论文的影响因素可能更侧重于国际合作和跨学科研究，而石油领域的高被引论文则更依赖于行业内的创新性和实用性。

3.2 影响因素的作用机制

石油类高校高被引论文的影响因素及其作用机制是一个复杂而多维的问题，本文将对随机森林方法挖掘的影响高被引论文的因素进行分析。

（1）论文的学术水平是被引次数最重要的影响因素，这与内容质量密切相关。一篇高质量的论文通常具有创新性、前沿性和实用性，能够解决该领域内的关键问题或提出新理论和方法。这样的论文更容易被同行认可和引用，从而成为高被引论文。石油领域的高被引论文往往与石油行业的实际需求密切相关。例如，针对特定地质条件下的油气勘探开发技术、提高采收率的方法等，这些研究能够直接解决石油行业面临的实际问题，因此更容易受到关注和引用。

（2）期刊的学术影响力与论文的被引频次密切相关，具体而言，那些拥有较高影响因子的期刊，通常享有更高的学术地位，并能吸引更庞大的读者基础。在这样的期刊上发表论文，有助于论文迅速进入主流学术视野，提高被引率。例如，《Petroleum Exploration and Development》作为石油工程类期刊中科院分区 1 区、影响因子为 7 的高水平期刊，发表在该刊上的论文往往能够获得更高的被引次数。

（3）作者的学术声誉、研究经验和合作网络等是影响论文质量和被引频次的重要因素。知名学者或研究团队发表的论文往往更容易受到关注和引用，因为他们的研究成果属于石油领域的技术突破，推动石油行业的发展和进步。

（4）高被引论文往往引用了大量权威且相关的文献。石油领域的研究往往涉及多个学科的交叉和融合，如地质学、地球物理学、化学、工程学等。跨学科的合作与融合有助于形成综合性的研究成果，提高研究的深度和广度，从而增加论文的被引频次。这些文献不仅为论文提供了坚实的理论基础，还增加了论文的可信度和说服力。并且引用的范围广泛，涵盖了多个研究方向和领域，体现了论文的跨学科影响力，促进了不同领域之间的知识交流和融合。同时，这些论文在引用时往往深入挖掘了相关文献的深层含义和价值，进一步提升了论文的学术价值。

（5）其他因素。一些其他因素也会影响石油类高校高被引论文的数量，例如，研究主题的前沿性对论文被引次数具有重要影响；选择一个既符合学科发展趋势又填补现有研究空白的课题，能够显著提高论文的吸引力和被关注的可能性；论文的标题、摘要、关键词等元素也可以提高论文的可见度和被检索到的概率，从而增加被引次数。

4 提升石油类高校高被引论文数量的策略

为了增加高被引论文的数量，石油类高校需要注重提升论文内容质量、选择高声誉期刊发表、培养知名学者和研究团队、加强交流与

合作等方面的工作。

（1）提升论文质量与创新性

聚焦当前石油领域的热点和前沿问题进行研究，这些领域往往更容易获得关注和引用。强化开创性研究，提出新的理论、方法或技术，这样的研究成果更容易被同行认可和引用。注重论文的实用性和应用价值，具有实际应用价值的研究成果更容易被工业界和学术界所关注，从而增加被引用的机会。

（2）优化论文发表策略

选择高质量期刊，投稿时应选择在本领域内具有影响力的高质量期刊，这些期刊的读者群更广，引用机会更多。优化论文标题和摘要，一个简洁明了、引人入胜的标题以及清晰、准确的摘要能够吸引读者的注意力，提高论文的可见度和被引率。

（3）注重人才培养及热点跟踪

对研究团队提供经费及政策支持，鼓励学者参与科研项目。并及时跟踪研究领域的最新进展，根据最新进展调整研究方向和策略，保持研究成果的时效性和前沿性。对已发表的研究成果进行持续的关注和更新，可以进一步提高其影响力和被引率。

（4）加强学术交流与合作

积极参加学术会议，通过参加学术会议，可以展示研究成果，与同行进行交流，增加论文的曝光度和引用机会。与其他学者合作撰写论文，合作研究可以汇聚多方智慧和资源，提高研究质量，同时增加论文的引用机会。

石油资源的稀缺性和在全球能源结构中的重要地位，使得石油领域的研究具有更高的关注度和重要性。结合行业的实际需求，提升石油类高校高被引论文数量，需要从论文质量、发表策略、学术交流与合作、学术声誉等多个方面入手。这些策略的实施将有助于石油类高校在学术界中提高声誉和影响力。为石油领域科研论文的发展提供指导。

参 考 文 献

[1] 方红玲 , 张亚杰 , 徐自超 . 第一作者和合著者的生产力、影响力与论文被引频次的相关性对比研究 [J]. 中国科技期刊研究 , 2024, 35(2): 273-279.

[2] 常宗强 , 叶喜艳 , 张静辉 , 等 . 基于被引频次的期刊论文被引质量评估指标构建 [J]. 编辑学报 , 2023, 35(S1): 238-240.

[3] 许林玉 . 高被引论文核心影响因素判别研究 [J]. 信息资源管理学报 , 2023, 13(5): 137-148.

[4] 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, 19(9): 30-37.

[5] 毛璐 , 许鑫 , 邓璐彡 . 基于研究数据评价的引证优化：高被引数据集特征视角 [J]. 情报科学 , 2023, 41(2): 126-134, 142.

[6] 刘红煦 , 唐莉 . 获评高被引学者会提升学术产出与影响力吗 ?——来自整体与个体层面的双重验证 [J]. 科学学研究 , 2021, 39(2): 212-221.

[7] 马荣康 , 李真真 . 高被引还是零被引：基于论文被引的最佳科研合作规模研究——来自 Financial Times TOP 45 商学院期刊的证据 [J]. 情报学报 , 2020, 39(11): 1182-1190.

[8] 赵婉忻 . 引文视角下国内高被引论文的 Altmetrics 指标相关性研究 [J]. 情报理论与实践 , 2020, 43(11): 47-53.

[9] 涂静 , 李永周 , 张文萍 . 国际合作网络结构与高被引论文产出的关系研究 [J]. 图书馆杂志 , 2019, 38(7): 69-75.

[10] BREIMAN L. Random forests [J]. Machine Learning, 2001, 45(1): 5-32.