arXiv:2301.04369v1 [cs.DL] 11 Jan 2023

科学中的可复现性信号：初步分析

Akhil Pandey Akella
计算机科学系 北伊利诺伊大学 aakella@niu.edu

Hamed Alhoori
计算机科学系 北伊利诺伊大学 alhoori@niu.edu

David Koop
计算机科学系 北伊利诺伊大学 dakoop@niu.edu

摘要

可复现性是科学研究的一项重要特征；实验会被重新测试，分析也会被重复进行。当获得一致的结果时，人们对研究发现的信任会增加。尽管可复现性十分重要，但这些工作往往需要投入大量精力，而且一些已发表的研究发现可能由于疏忽或错误而无法复现。本文考察了发表于计算机科学会议和期刊的学术文章中的大量特征，并检验这些特征与可复现性之间的相关性。我们从三个不同来源收集数据，这些来源将出版物标记为可复现或不可复现，并采用统计显著性检验来识别其中能够揭示可复现性的特征。我们发现，学术文章的可读性以及通过超链接获取软件制品的便利性，是可复现学术文章中尤为明显的强信号。

1 引言

科学过程的透明度能够加速科学发现，并增强公众对科学驱动事务的看法。可复现性在促进这种透明度方面发挥着至关重要的作用；令人鼓舞的是，科学界已达成共识，致力于解决科学中的可复现性问题。政策制定者、政府机构、开源社区、同行评审期刊、会议以及整个学术界，都有共同责任推动可复现研究。若科学过程中缺乏信任与诚信，科学便无法得到有效传播。实际上，可复现科学在研究实验室、课堂、行业和学术界等重要场所都有直接影响。缺乏可复现研究，可能会妨碍人们深入理解原始研究者的思维过程，从而严重影响此前提及的各个群体中的相关人员。

可复现性的概念复杂且具有层次性，其中包含彼此不同但相互补充的问题。在我们尝试理解如何应对可复现性问题之前，必须首先对这一语境下该术语的含义作出某种定义。诸如（Gundersen 和 Kjensmo、2018年；Cohen 等人、2018年；Barba，2018年）等研究强调了可复现性的定义如何因不同研究和学科而异，以及不同定义如何导致混淆。因此，可复现性在 Gundersen 和 Kjensmo （2018年）中提出的灵活定义颇具吸引力：“独立研究团队依据原始研究团队所作的文档，使用相同的方法得出相同结果的能力。”研究界的各种参与者——例如出版商、会议组织者和期刊——共同努力，在实验过程中推广确保可复现性的良好实践，这令人振奋；但对于究竟什么构成“良好实践”，目前仍缺乏共识，这一点令人担忧。

本研究尝试通过统计显著性检验，理解科学结构（Thelwall，2019年）与可复现性概念之间的关系。为此，我们着重考察与可复现性相关的语言特征和结构特征的认识论不透明性（Newman，2015）。我们通过运行大量假设检验，识别影响学术文章可复现性的显著因素，以此实现这一目标。我们的目标是利用统计检验筛选出相关信号，从而帮助识别哪些文章需要投入更多（或更少）精力来复现。

2 相关工作

可复现性是一个重要概念，通常会影响广泛的群体（Mede等人，2020年；Hutson，2018年）。关于可复现性的文献涵盖范围广泛，涉及

不同学科（开放科学合作组织，2012年；Prinz等人，2011年；Be-gley and Ellis，2012年；Peers等人，2012年）大体上聚焦于两种做法：开展大型荟萃分析，复现大量学术文章；或开展定性研究，鼓励研究人员采用某种方法论。

我们的研究与那些试图量化影响可复现性的重要因素的研究相一致，例如（Raff，2019年）。识别出这些重要因素也有助于构建机器学习模型，以估计学术文章的可复现程度（Yang等人，2020年）。

3 数据

尽管科学出版物通常遵循相似的结构，但在思想的传达和表达方式上仍有很大的自由度。这种不受拘束的特性使作者能够围绕基本思想编织故事，而读者对某些特定思想的吸收，有时也与这些思想的呈现方式有关。我们感兴趣的是，出版物的结构是否揭示了其潜在的（不可）复现性。为探究这一问题，我们从三个不同来源汇集了一批学术文章，这些文章已被评估为可复现或不可复现。对于每篇文章，我们都收集了全面的元数据，并提取了结构特征和语言特征。这些文章集包括：

- **布朗大学**：Collberg等人（Collberg等人，2015）开展了一项元分析，涉及复现发表于 ACM 计算机科学会议和期刊上的学术文章所需的步骤。他们发现，受审查的学术文章中，近 50% 需要付出额外努力才能复现。布朗大学的计算机科学家发起了一项名为“考察计算机科学中的可复现性”的工作，以众包方式重新审视这项研究（Krishnamurthi，2015）。他们对原研究进行了元分析，并提出了新的见解。收集到的数据详细呈现了复现原始出版物中研究所需付出的努力。当前存储库提供了 207 篇论文的结果；其中 142 篇被归类为可复现，65 篇被归类为不可复现。

- **撤稿观察数据库（Retrac- tionDB）**：撤稿观察数据库存储了从会议和期刊撤回的学术文章相关信息（Oransky和Marcus，2010年）。它还记录了学术文章所属的学科或领域、文章发表的国家、出版商名称、期刊名称，以及最重要的——文章被撤稿的原因。我们利用该数据库查找了计算机科学领域中因研究结果不可复现等原因而被撤稿的所有学术文章，共有 34 篇论文符合这些条件。

- **带徽章的ACM论文**：计算机协会（ACM）推出了徽章，用于标示成功复现的出版物。我们从176篇被授予“结果已复现”徽章的文章入手。其中，90篇被授予“可复用工件”徽章，这些文章中有70篇还获得了“功能性工件”徽章。我们成功获取了64篇带有“结果已复现”徽章，且同时获得“可复用工件”和“功能性工件”徽章的论文。

我们利用三个来源各自提供的元数据来查找文章。在某些情况下，我们通过文章标题和作者姓名进行搜索，以获取文章的数字对象标识符（DOI），有时也会获取文章的统一资源定位符（URL）。如果无法明确确定这些信息，就会将该文章从数据集中剔除。利用DOI，我们得以获取更多元数据以及文章全文，全文通常为PDF格式。填写完元数据并获取全文后，我们共收集到305篇论文，其中206篇被归类为可复现，99篇被归类为不可复现。数据和代码将在发表时作为补充信息提供。

4 方法论

4.1 特征工程

考虑以下特征的动机源于（Gundersen 等人，2018年；Gundersen，2020年；Raff，2019年）中强调的共同直觉，以及 NeurIPS、ICML 等热门发表场所提供的清单。

表1：结构特征列表及其与目标变量的点二列相关

特征	p值
是否包含引言部分	0.0808
是否包含方法论章节	0.3112
是否包含结果部分	0.7006
页数	0.1630
图像数量	0.3571
表格数量	0.7187
算法数量	0.0654
超链接数量	0.0028
公式数量	0.4212

1. **结构特征：** 与学术文章结构相关的定量和定性信息。这包括特定章节是否存在的信息，以及给定学术文章中的表格、图或算法数量。我们开发了 Python 模块来解析学术文章的 PDF，以提取这些信息。相关特征及其对应的点二列相关见表1。

2.**语言特征：** 用于量化学术文章所用语言不同指标的语言指标，以区分不同作者的写作风格。这些指标包括词数、平均词长、平均句长、长度超过平均词长的词语出现频率、音节数，以及尤尔I词汇多样性度量（尤尔，2014）。这些特征属于计算语言学的通用范畴，且易于理解。此外，我们还考虑了复杂词，即给定文本中的多音节词数量。该特征通过 Python*TextBlob* 库提取。平均可读性通过计算多项可读性指标的平均值来衡量，这些指标包括弗莱什阅读易读度、SMOG 指数、Coleman-Liau 指数、自动可读性指数、Dale-Chall 可读性评分、Linsear Write 公式和 Gunning FOG 指数。我们从 $textstat$（一个 Python 软件包）获取可读性指标。此外，我们还收集了完整情感评分

给定学术文章的文本，并为相应文章附上情感标签（正面 = 1、负面 = 0）。我们采用类似的方法获取文章标题的情感标签。

表2：语言特征列表及其与目标变量的点二列相关

特征	p值
词数	0.5357
平均词长	0.2379
长于平均词长的词频	0.9804
复杂词	0.8394
音节数	0.7467
尤尔 I 度量的词汇多样性	0.1102
平均可读性	0.0000
文章情感	0.5659
标题情感	0.7335

我们通过编写 Python 程序收集了这些信息，这些程序使用了 *spaCy* 和 *NLTK* 等 Python 库来构建计算这些指标的方法。所有这些语言学指标均基于学术文章的全文。表格中列出了这些特征及其相应的点二列相关系数。 2。

4.2 点二列相关

可以使用相关性对因变量和自变量进行初步统计分析。由于目标是名义变量，因此不能使用 *Pearson*相关或*Spearman* 相关，因为这两种方法都假定目标变量是连续变量。点二列(Gupta，1960年)相关矩阵用于测量二分目标变量与连续变量之间的相关性。表1 和表2 中的结果是通过计算点二列相关系数及其对应的 p值获得的。

4.3 显著性检验

表格中提及的特征1 和2 结合了有序属性和名义属性。为了确定这些特征的显著性，我们必须采用不同的统计显著性检验，例如 *Mann-Whitney U* 检验（Mann 和 Whitney，1947年）和卡方检验（Yates，1934年）。

5 结果

我们对组合数据源进行了相关性计算和统计显著性检验，以确定在表明学术文章可复现性方面发挥重要作用的特征。如表格1 和2 所示，点二列相关**平均可读性**和**超链接数量**显示，只有这两项与可复现性显著相关。

Mann-Whitney U 和卡方 检验的结果表明，**平均可读性、超链接数量、算法数量、平均词长和尤尔词汇多样性度量**是具有统计显著性的特征，能够与可复现的学术研究相吻合，并以合理的确定性发出信号。更重要的是，学术文章的可读性以及软件制品的可访问性（无论其形式是代码仓库、伪代码还是算法）都可视为可复现性的有力指标。需要注意的是，这些信号并不量化或保证学术文章的可复现性，而是有助于识别哪些文章需要投入更多（或更少）精力才能复现。

表 3：数值特征的 Mann-Whitney U 显著性检验

特征	p值
尤尔 I 词汇多样性度量	0.0131
多样性	
词数	0.6547
平均词长	0.0003
长于	0.9171
平均词长的词频	
音节数	0.3910
复杂词	0.9596
平均可读性	0.0001
图像数量	0.2039
表格数量	0.9586
算法数量	0.0283
论文长度	0.5039
超链接数量	0.0011
公式数量	0.2148

我们的研究发现得到了统计实验结果的支持，例如点二列相关、卡方检验和曼-惠特尼U检验；p值（$p < 0.05$）则作为判断研究发现显著性的依据。您可以获取数据集、实验设置及其他内容的副本来自 GitHub代码仓库的软件制品。 [1]。

表 4：分类特征的卡方显著性检验

特征	p值
是否包含引言部分	0.1070
tion	
是否包含方法论部分	0.3728
tion	
是否包含结果部分	0.8617
文章情感	0.6646
标题情感	0.8495

6 讨论

科学结构涉及一个完善的过程：从准确有效的数据开始，接着详细描述实验流程，随后清晰地呈现结果。科学过程有许多原则，其中这些只是其中一部分。多年来，这些原则一直得到推广，使科学过程能够在同行评审这一制衡机制下蓬勃发展。从具体情境来看，学科、年份、科学研究类型等因素在确定复现文章所需的工作量方面发挥着重要作用。因此，在解读我们的研究发现时，我们构建的数据集是一个不可或缺的考量因素：学术文章的可读性，以及通过超链接获取软件制品的便利性，是可复现学术文章的重要特征。我们的动机是发现其他潜变量，在确定复现文章所需的工作量时将这些情境因素纳入考量。

7 结论与未来工作

在这项研究中，为了识别能够表明科学是否可复现的特征，我们进行了相关性分析和显著性检验。我们发现，在可复现的学术文章中，学术文章的可读性以及通过超链接访问软件制品的便利性是重要特征。我们的代码仓库将包含数据和实验，并在发表后开放。

未来，我们计划通过以下方式扩大研究范围：1）从 ACM 收集更多带徽章的数据；2）使用对抗样本检验我们的研究发现的有效性；3）观察引用可复现文章与不可复现文章所产生的影响

[1] https://github.com/reproducibilityproject/reproducibilitysignals

引用可复现文章与不可复现文章的情况

个。

8 致谢

本研究部分得到美国国家科学基金会（NSF）2022443号资助。

参考文献

Lorena A. Barba.2018. 可复现研究的术语.CG. Begley andLM. Ellis. 2012. 药物开发：提高临床前癌症研究的标准. 《自然》. K. Bretonnel Cohen, Jingbo Xia, Pierre Zweigen- baum, Tiffany Callahan, Orin Hargraves, Foster Goss, Nancy Ide, Aurélie Névéol, Cyril Grouin, and Lawrence E. Hunter. 2018. 自然语言处理中的可复现性的三个维度. In第十一届国际语言资源与评估会议论文集（*LREC 2018*），日本宫崎。欧洲语言资源协会（ELRA）。Christian Collberg, Todd Proebsting, and Alex M War- ren. 2015. 计算机系统研究中的可重复性与贡献。 亚利桑那大学技术报告，14:4。O. E. Gundersen. 2020. 可复现性的基本原则. arXiv 预印本 cs-LG。O. E. Gundersen, Y. Gil, and D. W. Aha. 2018. 关于可复现人工智能：迈向人工智能出版物中的可复现研究、开放科学与数字学术。*AIMag*，39(3):56–68。Odd Erik Gundersen and Sigbjørn Kjensmo. 2018. 现状：人工智能中的可复现性。载于 第三十二届人工智能促进协会人工智能会议。S.D. Gupta. 1960. 点二列相关系数及其推广。 《心理测量学》，25:393–408。M. Hutson. 2018. 人工智能面临可复现性危机。美国科学促进会。Shriram Krishnamurthi. 2015. 审视计算机科学中的可复现性。H. B. Mann and D. R. Whitney. 1947. 关于检验两个随机变量中一个在随机意义上是否大于另一个。 《数理统计年刊》，18(1):50–60。N. G. Mede, M. S. Sch\"afer, R. Ziegler, and M. Weißkopf. 2020. “重复性危机”进入公众视野：德国人对科学研究（不可）复现性的认知与看法。《公众对科学的理解》，9636。JulianNewman.2015. 认识论不透明性、确认整体论与技术债务：从实证软件工程视角看计算机模拟。载于*HaPoC*，第487卷，第256–272页。开放科学合作组织。2012. 一项旨在估计心理科学可复现性的开放式、大规模协作研究。 《心理科学展望》，7(6):657–660。I. Oransky and A. Marcus. 2010. 撤稿观察数据库。I. S. Peers, P. R. Ceuppens, and C. Harbron. 2012. 探寻临床前研究稳健性。 《自然综述：药物发现》，10:733。F. Prinz, T. Schlange, and K. Asadullah. 2011. 信不信由你：我们对潜在药物靶点的已发表数据能有多大程度的信心？ 《自然综述：药物发现》，10。E. Raff. 2019. 朝量化可独立复现的机器学习研究迈进一步。 《神经信息处理系统进展》第*32*卷，第5485–5495页。Mike Thelwall. 2019. 科学的修辞结构？对文章标题的多学科分析。 《信息计量学杂志》，13(2):555–563。Y. Yang, W. Youyou, and B. Uzzi. 2020. 利用人类和人工智能估算科学发现的深层可复现性。 《美国国家科学院院刊》，117:10762–10768。F. Yates. 1934年. 涉及小样本数的列联表与卡方检验。 《皇家统计学会杂志》，1(2):217–235。C Udny Yule. 2014. 文学词汇的统计研究。剑桥大学出版社。