基于 OpenDigger 开源实践的学科交叉性洞察

——来自本科生角度的开源生态分析报告

柯宇 王可楠

华东师范大学 不知道叫什么才队

【摘要】当前互联网与信息技术领域，甚至信息通信领域来到了一个数据膨胀、资源溢出的状态。相关产业业者在这样的环境下，亟待一种可以交流资源、分享经验、展示项目的稳定路径。而开源社区作为一种现存的在线方式，深受广大相关领域工程师、科学家以及爱好者的认同。但作为互联网与信息技术事业的入门人，譬如广大爱好者和学生等在开源社区中摸爬滚打的人们，在开源社区中的发言权是较小的、声音也微乎其微；而官方开源社区的教程对于中国学生、中国入门者的建设性作用也有限。因此基于 OpenDigger 以及 OpenRank 所展示出的一些问题，结合开源社区入门者、特别是中国入门者在庞大日志数据中所反映出的在开源社区中遇到的问题，本文将从开源社区的指标估量与评价、对于入门者的友好程度以及中国目前开源事业遇到的瓶颈三个方面从学生的角度进行讨论。

【关键词】OpenSource；OpenDigger；Amateur；Student；Massive Data

【正文】

一．开源社区日志收集处理、展示、分析及其评价维度的优势和可改进点

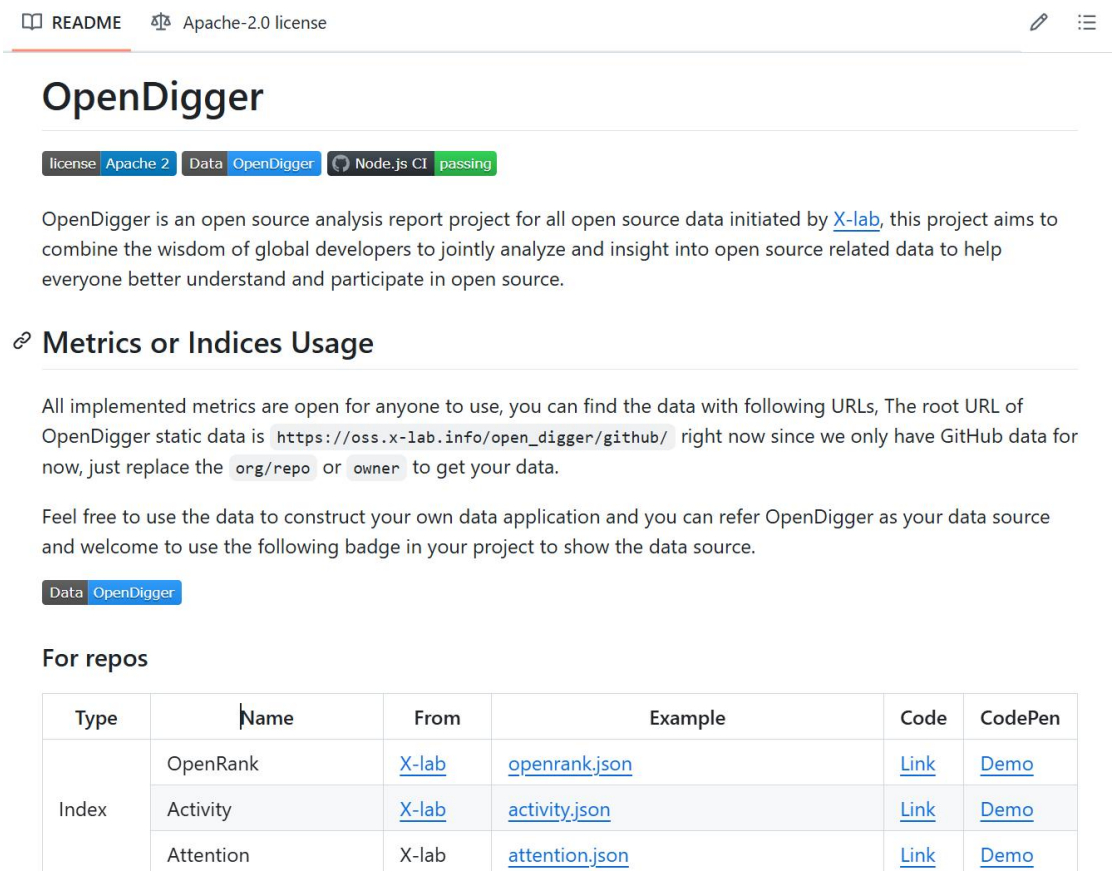
在本次使用 OpenDigger 分析 Github 上各个项目仓库与各个开源参与者的开源实践情况时，尽管这样的工具能够具有很强大的挖掘潜力、能够让广大开源参与者做出可视化大屏进行展示、也为初入开源社区参与开源事业的人们提供了最基础的贡献机会，但其中的部分细节引起了思考。

诚然，一个基于中国本土的开源挖掘工具无疑增加了中国开源社区在国际开源社区的话语权。在国家政策、国家机构的倾斜和帮助下，OpenDigger 搭建了 OpenRank 评价体系，以一个被开源平台认可的标准对所有的开源仓库、开源贡献者进行量化评估并且发布排名，展示了在开源社区中也可以使用科学量化的力量进行探究与实践；而开源 Top 榜单的发布也在相关产业中甚至是社会上引起反响，能够创造出良好的开源创造氛围，吸引更多的青年朋友和广大相关从业者参与到开源事业当中；从其直接效果上看，OpenDigger 挖掘工具为开源参与者或是社区爱好者的数据分析以及可视化展示提供了庞大的数据集，节省了这些参与者

的数据挖掘成本。

但不可否认的是，在笔者看来，OpenDigger 仍然存在了许多问题。

分析一个开源数据挖掘项目，不妨从数据的量级入手。因为 OpenDigger 尽管是开源工具，但庞大的数据量级仍然需要一个稳定的云数据库或是云服务器进行维护，而使用需要成本输出的云服务器就不可避免地会涉及到权限的可及性(accessibility)的问题。在其他开源参与者使用数据时，可能没有办法通过接入数据库进行操作，因而导致获取的数据的用途范围被大大减小。对于大部分如 Github、Gitee 等开源网站上的数字挖掘工具来说，其中并不会提供对于完整的数据集于调用者本地环境在 SQL 层面的处理路径，只会提供 API 调用路径。尚且不考虑配置环境中所需的成本（如本次实验中部分其他参与者调用 Click_House 数据库所配置环境付出的成本），其权限问题以及开放权限后增生维护成本就已然是一个让许多使用者头痛的难题。因此，尽管开源数据挖掘工具提供了一个路径，但其他开源参与者于数据处理仍存在使用困难情形也尚待解决。



README Apache-2.0 license

OpenDigger

license Apache 2 Data OpenDigger Node.js CI passing

OpenDigger is an open source analysis report project for all open source data initiated by X-lab, this project aims to combine the wisdom of global developers to jointly analyze and insight into open source related data to help everyone better understand and participate in open source.

Metrics or Indices Usage

All implemented metrics are open for anyone to use, you can find the data with following URLs, The root URL of OpenDigger static data is `https://oss.x-lab.info/open_digger/github/` right now since we only have GitHub data for now, just replace the `org/repo` or `owner` to get your data.

Feel free to use the data to construct your own data application and you can refer OpenDigger as your data source and welcome to use the following badge in your project to show the data source.

Data OpenDigger

For repos

Type	Name	From	Example	Code	CodePen
Index	OpenRank	X-lab	openrank.json	Link	Demo
	Activity	X-lab	activity.json	Link	Demo
	Attention	X-lab	attention.json	Link	Demo

图 1-1 Open-Digger 使用方法截图 直接达到数据层面具有复杂性

数据的庞大体量会导致的第二个问题是经济与资源成本的问题。尽管在 OpenDigger 管理者举办的活动中，开源贡献者可以获得部分样例数据，并且以局部概括整体的方式进行开

源统计学分析，但这样的样例数据对于缺乏资源的个人开源参与者仍然是一个不小的开支。举例来说，RTX4070 显卡加上 16GB 的 RAM 在开源活动中是较难去支持样例数据的 CSV 文件的打开与处理的，而部分 Python Module 对于调用时间的限制也使得数据的清晰与分析难以有效进行。从量化分析的角度，以本地对 Open-Digger 一个月的样例数据进行处理为例，对于 log_2020_01.csv 文件中所提取出的 1,000,000 行命名为 Washed4.csv 文件进行数据的读取、过滤以及划分操作操作，需要长达 10.2s。

```
# 读取CSV文件
df = pd.read_csv('Washed4.csv')

# 剔除create_description为空的行
df_cleaned = df.dropna(subset=['create_description'])

# 按repo_id分组，统计每个repo_id的行数
repo_count = df_cleaned.groupby('repo_id').size().reset_index(name='count')

# 定义count的区间
bins = [0, 1, 5, 10, 20, 50, 100, 1000]  # 自定义区间
labels = ['1 or less', '2-5', '6-10', '11-20', '21-50', '51-100', '>100']  # 给每个区间一个标签

# 将count分类
repo_count['count_range'] = pd.cut(repo_count['count'], bins=bins, labels=labels, right=False)

# 对count_range进行分组统计每个区间的repo_id数量
range_count = repo_count.groupby('count_range').size().reset_index(name='repo_count')

# 将结果保存为新的CSV文件
✓ 10.2s
```

图 1-2 实验过程中的操作截图

而 Washed4.csv 文件仅是 open-digger 一个月的操作数据中千分之一（1/1000）的一部分，尽管通过 csv 文件而不是数据库层面进行时间统计会有所偏差，但整体数据处理的复杂性和繁琐性可见一斑。那么进而，从后台管理成本的角度上看，因为 OpenDigger 是一个基于中国高校的开源工具，因此其维护成本应当是有限的，并且很大一部分可能依靠着校企合作机制进行。对于数据库、服务器的租用或是对于硬件的采购，对于一个实验室项目来说，一定会是一个不小的开支。而加上人员成本，开支难以估量。

而对于使用数据的开源实践者来说，想要可视化地展示数据，不可避免地也会需要用到第三方的资源展示平台，这仍然会是一份新的开支与资源利用。因此，笔者在开源实践中尝试只通过 HTML\CSS\JavaScript 三件套完成可视化展示模块。唯有做到可持续性、可再创造性，数据的使用才会是极致的。

此外，开源社区活动的多样性会使得开源数据挖掘存在一定的难度。以 OpenDigger 为例，依笔者愚见，OpenDigger 中挖掘的指标多样性会导致数据出现冗杂与部分无效性。譬如在样例数据中，我们可以清晰地发现有一部分的数据永远都是空列或者是空行。例如 Create_Event 的 Merge_at 指标。在逻辑上，Create_Event 数据出现这一指标也是不合适的。

这样问题因为在对日志数据进行统计的过程中没有预先对不同的 Event 类别进行处理导致的，也根源于对于一个完美数据挖掘工具的追求所需的庞大工作量和人员、资源的有限性之间的矛盾导致的。类似的问题出现在 Merge_Login_Type 这一指标上，所有的类别全部都是 Bot。但是通过观察，部分的提交信息数据中是有不同语言的提交信息以及错别字等少量人类思维痕迹在的，因此可以认为这可能是某个机器学习识别过程、或是程序上的笔误。

最后，对于 OpenRank 指标，笔者认为也可以加入更多的维度进行考量。在实验中可以发现，开源贡献者的分布与经济社会的发展状况也存在着很大的联系：在经济发达的地区，开源贡献者的密度是较高的，而且具有一定的经验和实践机会优势。因此可以在不侵犯隐私的前提下，以一种非歧视的视角，加入对于开源贡献者的地区进行评分赋值。例如来自南美或者非洲的开源贡献者，可以对其贡献值赋予更大的比重，考量该贡献对于贡献者所在国家的贡献幅度的比重。

二．对于 OpenDigger 开源工具以及 OpenRank 指标的维护建议

首先，应当吸引更多的开源爱好者、开源工作者参与到 OpenDigger 这一项目当中来，更大发挥好所收集到的数据的作用。很显然的是，在人员充足的情况下，对于开源社区日志进行挖掘时，可以直接将不同的 Event_Type 在初期就分为不同模块，并且在导出时可以通过不同的 CSV 文件进行导出，避免部分行数据的冗余。

其次从经济与资源成本的角度看来，基于高校背景的数据挖掘工具搭建是远远不够的、

图 2-1 华东师范大学数据科学与工程学院校企合作案例

可能也是难以维持的。在已有的平台和背景下，高校和 OpenDigger 项目应当抓住更多的校企合作机会、吸引到更多企业的参与，在资源与资金充足的条件下，才能更好地维持项目的可持续性以及保障其对开源社区的贡献。

图 2-2 OSGraph 的企业成功案例

以上两点都有一个先决的前提条件，就是要开始注重开源事业的社会性。开源和人工智能、算法等应用性强的领域有所区别，它是广大人民可以直接参与、门槛较低的互联网子学科。“开源不仅仅是一个技术层面的概念，更是一种文化和精神的象征，激励着全球开发者加入到软件开发与创新过程中。”正如福建省开源数字技术研究院秘书长叶伟华所说，开源创新正在深刻改变技术创新和产业发展的格局。因此，开源先锋者需要在社会上广泛且有效地宣传开源事业这一在当下互联网和数据热潮中仍然具有强大潜力的事业，并且将其从高校、科研院所以及从业者的专属变为广大人民都可以去了解、去参与的一个事业。这其中更离不开国家相关部门的指导、引领和宣传。

从社会性上看，OpenRank 作为属于中国的第一个开源评价指标，也应该具有强大的社会关照力和普世的包容性。如上文所说，Open-Digger 以及 OpenRank 可以考虑在分数中加入一个分数与地区发达程度成反比的加分项，能够对来自落后地区的技术人员起到更大的激励作用。作为中国的第一个开源标准，将人的因素添加到指标的考量里面去，将促进开源事业发展作为人的目的，会受到社会上更广大人民的欢迎。

三．开源社区对于入门者的友好程度分析

正如前文所提及，尽管开源社区是一个互联网、信息技术、数据科学的专业社群，但是其目的在于技术、也在于人。一个成功的社区从来不能只是一家之言，需要有对新加入进社区的参与者的一定的倾斜与关照。

从笔者在分析与可视化展示中所提取和预测的数据看来，开源社区中的贡献和活动呈现出了两个极端的态势。对于大部分参与者来说，难以在开源社区中有自己的贡献、发出自己的声音；而对于“技术大咖”们来说，发言是一件微乎其微的事情，目标是如何将自己的产品做出新意。而对于开源社区新人来说，想要经营好自己的项目更是难上加难。

首先，开源社区对于新人的邀请不应只是口口相传。对于大部分初学者，特别是中国初学者来说，在一开始可能连 Github 是什么都不了解，更何谈参与呢？因此中国学生使用 Github 或者其它的国内平替版开源社区的入门应当在广大高校、甚至中学有更加成体系、成系统、标准化的教程。大学乃达道之学也，开源是还未涉足真正事业的初学者和未来工程师们的第一条引道。

其次，开源社区应当在学生认证的基础上对学生用户有更多的倾斜。学生认证在 Github 这一开源代码托管平台上已然成型，但如 OpenDigger 这样的开源数据挖掘工具难以获得将学生群体从广大用户群体中分离出来的数据，因此许多的开源实践者和教育实践者无法对学生群体画像进行有针对性的分析或是形成可视化结果。学生是一个具有先进性和积极性的群体，通过对于学生用户群体提交日志以及对学生群体画像的仔细刻画和深入分析，更能对未来的开源事业有着更加准确、精准的预测作用；同时，开源教育仍然是一项新的潮流，并没有一个固定的范式或是体系的形成。而一份针对于学生群体的数据，有利于开源教育探索事半功倍。

而对于中国的学生群体和信息行业入门者来说，使用 git 和 Github 等面向国际的开源社区有着自己的特殊情况。首先，是网络的不稳定在一定程度上阻碍了参与；其次是语言上的分割与隔阂，以及中文开源教育资源的不足和不流通导致了大部分的中国开源参与者在入门时没有办法得到充分的学习，只能依靠网路上的资源进行自己的探索。以 Github 手册为例，其有中文版本，但许多细节不如英文版本说得明确与完整，或是本地的 git 程序或是 turtle 上传程式仅有英文版本，出现了语言上的脱节。

图 3-1 中文版 Github 手册中“洋腔洋调”的中文

这对初学者，特别是中国初学者，并非一个友善的举措。而如 Gitee、GitCode 等国内代码托管平台普遍面临着参与者较少、优秀品质的项目较少的问题，距 Github 这一平台仍然具有一定的进步空间。

因此，在开源领域中，贡献者们不仅可以在统计角度上做出贡献，也可以通过产出对 Github 手册、Git 流程等官方文件的完整、确实的翻译作品，或是设计并实践检验出一套符合中国教育现实、符合科学教育原则的新颖教育体系，将源源不断的潜在人才引入开源与互联网、信息技术、数据产业的大门，发出更大的光芒。

四．总结：中国开源事业仍处于上升期

笔者作为学生，也是开源事业的参与者。尽管在开源实践中发现了一些可以改进得更好、更加人性化的瑕疵点，但中国开源事业仍然让所有的开源参与者感受到一种向上的勃勃生机与活力。

尽管没有成型的教育体系和资源体系，但正是因为全球都缺少这样的体系，中国的开源贡献者，包括标准制定者、参与者、实践者才有更大的施展才华和创造力的空间。在可视化体系中，实践者们可以采用大屏、网站等方式进行展示；在理论体系板块，实践者可以从不同角度通过报告、论文阐述自己的观点，并制作幻灯片展示；在教育实践板块，可以用开源资源创造平台、用开源数据分析预测前景，促进下一代信息技术人才和数据人才的培养。

而属于中国自己的开源社区建设也在紧锣密鼓的进行中。尽管大部分的开源实践仍然在 Github 上进行，但 Gitee、GitCode 等中国平台的建设在中国大陆已经占有了话语权。接下来，在 OpenDigger、OpenRank 等开源数据工具的助推和协助下，中国标准体系有更大可能会带领中国开源社区走向世界。

【参考文献】

[1] 韩凡宇,毕枫林,张琰彬,等.OpenPerf：面向开源生态可持续发展的数据科学基准测试体系[J/OL].计算机学报,1-17[2024-12-18].http://kns.cnki.net/kcms/detail/11.1826.tp.20241101.1453.011.html.

[2] 蒋雅琛.开源之路在何方？[N].福州日报,2024-12-09(008).

[3] 齐佳音,张国锋,王伟.开源数字经济的创新逻辑:大数据合作资产视角[J].北京交通大学学报(社会科学版),2021,20(03):37-49.DOI:10.16797/j.cnki.11-5224/c.20210706.009.

[4] 王哲,薛澜,赵静.开源创新组织的创新成本演化与动态治理机制[J/OL].公共管理评论,1-31[2024-12-18].http://kns.cnki.net/kcms/detail/10.1653.d0.20241022.1057.002.html.

[5] 任正非：华为技术有限公司首席执行官开源社区致力于培育肥沃的“黑土地”[J].中国商人,2024,(08):216-217.

[6] 李锐,史依颖,冯冠霖.开源生态的四大挑战与繁荣之道[J].中国工业和信息化,2024,(07):20-24.DOI:10.19609/j.cnki.cn10-1299/f.2024.07.013.