Analyzing the GitHub Repositories of Research Papers

Michael Färber
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
michael.faerber@kit.edu

ABSTRACT

Linking to code repositories, such as on GitHub, in scientific papers becomes increasingly common in the field of computer science. The actual quality and usage of these repositories are, however, to a large degree unknown so far. In this paper, we present for the first time a thorough analysis of all GitHub code repositories linked in scientific papers using the Microsoft Academic Graph as a data source. We analyze the repositories and their associated papers with respect to various dimensions. We observe that the number of stars and forks, respectively, over all repositories follows a power-law distribution. In the majority of cases, only one person from the authors is contributing to the repository. The GitHub manuals are mostly kept rather short with few sentences. The source code is mostly provided in Python. The papers containing the repository URLs as well as the papers' authors are typically from the AI field.

KEYWORDS

GitHub, code repositories, papers, scholarly data, bibliometrics

1 MOTIVATION

The number of research papers has increased steadily in the past. This is particularly true for papers in AI-related fields, such as machine learning and computer vision. For instance, more than 60,000 papers have been published in the area of machine learning in each of the last years [2]. Furthermore, it has become increasingly common to provide links to source code repositories in the computer science research papers. In this way, research (i.e., approaches and evaluations) became more replicable and reproducible. So far, little effort has been performed on analyzing the status and characteristics of code repositories belonging to research papers. Existing works focused on measuring the importance and influence of GitHub repositories in general (e.g, [1, 3]) without considering research papers. In this paper, we analyze the GitHub repositories of research papers and outline characteristics of these repositories and the corresponding papers in which they are mentioned.

2 ANALYSIS

2.1 Data Set and Methodology

The Microsoft Academic Graph (MAG) [4] (as of Dec 13, 2019) already contains links to code repositories mentioned in research papers. Our preliminary analysis revealed that this set of URLs to code repositories is more complete and more precise than using an own approach of extracting source code repository URLs with an own implemented information extraction method from papers' full text. Specifically, the links to code repositories have already been tagged as being primary (linking to the actual repository) or being secondary (linking to additional repositories). From the set of primary URLs, 4,876 of them link to GitHub and are mentioned in papers belonging to computer science according to the MAG field of study assignment. Since GitHub is currently the most popular source code platform, we use this URL set as data basis. We were able to download 2,955 out of the 4,876 repositories. The remaining repositories were either unavailable or duplicates to already downloaded repositories.[1]

2.2 Analysis Results

Figure 1, 2, 3 and 4 illustrate the number of stars, forks, contributors and manual lengths respectively of the GitHub repositories. We can observe that the number of stars and forks of all GitHub repository in our collection is – with only a few exceptions – in the range of zero to ten. On average, there are four contributors per repository. In many cases, the repository is provided by a single account. For many repositories, the manual is kept very short leading to difficulties in terms of replicability and reproducibility.

Figure 5 and 6 illustrate the used programming languages and machine learning frameworks. We can observe that Python has emerged to be the most popular programming language in the repositories. Tensorflow has become very successful as framework.

Figure 7 shows the most frequently occurring fields of study (of all field of study hierarchies). Unsurprisingly, papers containing a link to a GitHub repository deal mostly with AI-related topics (e.g., machine learning, pattern recognition). Also, many repositories are linked in papers assigned to mathematics.

Table 1 and Table 2 show the journals and conference series in which GitHub repositories are mentioned most frequently. We can see that AI-related fields, such as computer vision, machine learning, and natural language processing, are well represented.

Considering the papers' authors being listed as corresponding repository contributors, we can observe that the papers' first authors own the most repositories (63%), followed by second authors (20%) and third authors (10%).

[1] The source code and data is available online at https://github.com/michaelfaerber/paper-github-analysis.

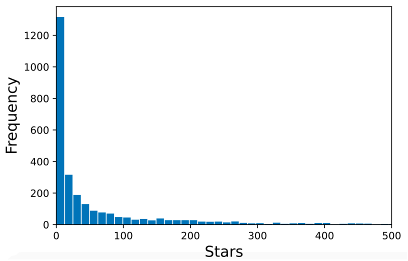

Figure 1: Distribution of repositories by the number of stars on GitHub.

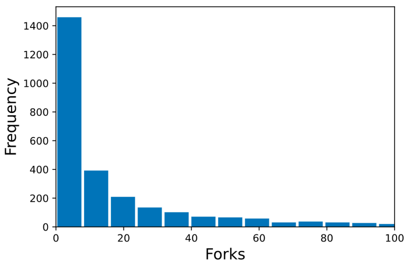

Figure 2: Distribution of repositories by the number of forks on GitHub.

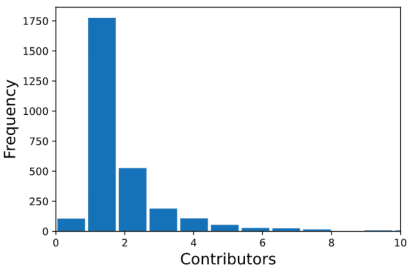

Figure 3: Distribution of repositories by the number of contributors.

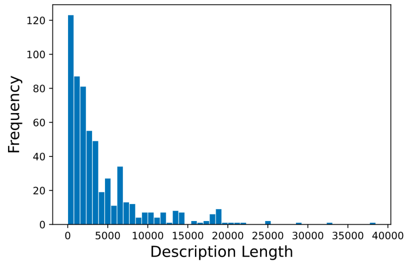

Figure 4: Lengths of the repository manuals.

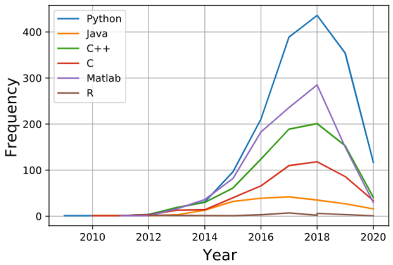

Figure 5: Programming languages used in the repositories by the year of the last repository commit.

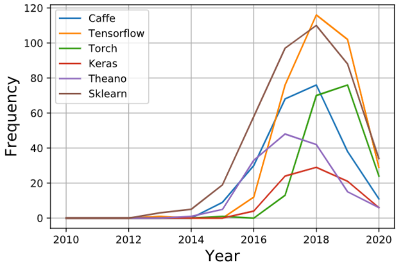

Figure 6: Machine learning frameworks used in the repositories by the year of the last repository commit.

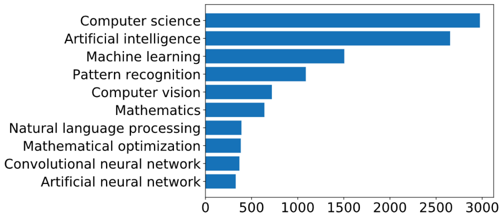

Figure 7: Top 10 fields of study.

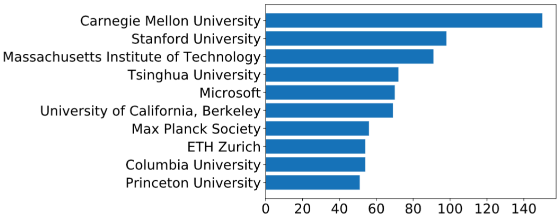

Figure 8: Top 10 affiliations.

Table 1: Top 5 journals (incl. pre-print).

Journal	#
arXiv	1,416
IEEE Transactions on Image Processing	42
IEEE Transactions on Pattern Analysis and Machine Intell.	27
bioRxiv	27
IACR Cryptology ePrint Archive	19

Table 2: Top 5 conference series.

Conference Series	#
Computer Vision and Pattern Recognition (CVPR)	190
Neural Information Processing Systems (NeurIPS)	84
European Conference on Computer Vision (ECCV)	79
International Conference on Computer Vision (ICCV)	76
International Conference on Machine Learning (ICML)	60

Figure 8 shows the affiliation from the papers' authors. We can observe that the CMU is particularly promoting open source publications on GitHub. Most universities are located in the US. Microsoft is the only company in the top 10 list.

3 CONCLUSION

In this paper, we presented a first analysis of GitHub repositories from research papers modeled in the Microsoft Academic Graph. Overall, we saw that providing links to GitHub repositories has become increasingly common. However, we observed a strong bias towards specific computer science areas (e.g., machine learning), papers' venues, as well as authors from specific institutions. In the future, we will investigate these biases in more detail.

REFERENCES

[1] Laura A. Dabbish, H. Colleen Stuart, Jason Tsay, and James D. Herbsleb. 2012. Social Coding in GitHub: Transparency and Collaboration in an Open Software Repository. In *Proceedings of the Computer Supported Cooperative Work (CSCW'12)*. 1277–1286.

[2] Michael Färber. 2019. The Microsoft Academic Knowledge Graph: A Linked Data Source with 8 Billion Triples of Scholarly Data. In *Proceedings of the 18th International Semantic Web Conference (ISWC'19)*. 113–129.

[3] Yan Hu, Jun Zhang, Xiaomei Bai, Shuo Yu, and Zhuo Yang. 2016. Influence analysis of Github repositories. *SpringerPlus* 5, 1 (2016), 1–19.

[4] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Paul Hsu, and Kuansan Wang. 2015. An Overview of Microsoft Academic Service (MAS) and Applications. In *Proceedings of the 24th International Conference on World Wide Web Companion (WWW'15)*. 243–246.