Reproducibility Report for ACM SIGMOD 2022 Paper: "JEDI: These aren't the JSON documents you're looking for…"

TORSTEN GRUST, University of Tübingen, Germany

We have been able to faithfully reproduce the original paper's findings as well as the key performance results reported in its experimental section. The authors provided data files, a framework of scripts, and plotting routines that allowed the (near identical) reconstruction of two of the paper's main figures.

1 INTRODUCTION

The original paper studied efficient similarity search on collections of JSON (thus tree-shaped) documents, with a focus on JSON's array and object type constructors: the former imposes a strict ordering on its contained elements, while the latter establishes an unordered sibling relationship between fields. The authors introduce a notion of JSON tree edit distance (coined *JEDI*), its optimized implementation based on pruning, and a supporting index structure to facilitate tree similarity search over large JSON corpora.

An experimental study addresses the effectiveness of pruning and index support as well as the runtime of search operations with varied tree distance thresholds.

2 SUBMISSION

The paper is accompanied by a reproducibility setup that allows readers to replay its effectiveness and runtime experiments. A separate two-page README contains pointers to a source code repository and setup instructions:

- Code repository: `https://github.com/DatabaseGroup/jedi-experiments`
- Data repository: `https://github.com/DatabaseGroup/jedi-datasets`
- The README describes a Docker-based as well as a native setup to run the experiments. Readers that choose the latter route find a list of the required programming languages (C++11, Python 3, shell) and libraries in the README. In the present reproducibility report, we have chosen the Docker-based route.
- Shell scripts are provided to automatically download and pre-process the JSON corpora (hosted at the authors' site) required by the experiments.

3 HARDWARE AND SOFTWARE ENVIRONMENT

Find a comparison of the authors' original and our local machine setups in Table 1. The README estimated a running time of about 20 hours for the complete batch of experiments and this is exactly what we observed during our review.

4 REPRODUCIBILITY EVALUATION

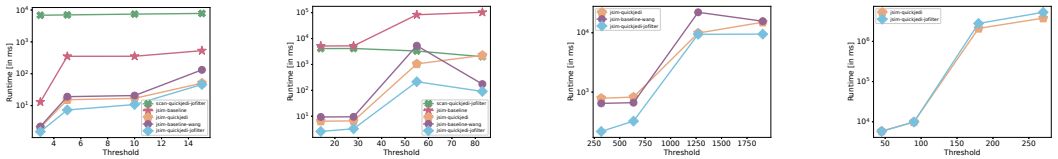
4.1 Process

We followed the Docker-based route described in Section D.1 of the README. Once the Docker image had been prepared (`docker build`), the actual experiments were performed in a single 20-hour batch. We used a multiplexing terminal emulator (`tmux`) to detach from and then re-attach to the host machine to check the experiment's progress messages—no attention or intervention is required during the entire run. If readers want to deviate from this "canned" form of prepared experiments, Python scripts are provided (see `scripts/execute-lookup.py`) to tweak configuration parameters.

Raw result data and formatted plots are found in the `results/data` and `results/plots` directories on the host (*i.e.*, outside the Docker image), once the batch run completes.

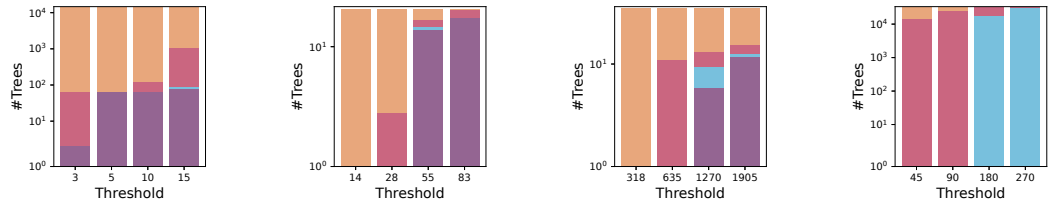
Table 1. Comparison of hardware and software setups.

	Paper [1]	**Repro Review**
CPU	Intel® Xeon E5-2630 @ 2.4 GHz	AMD® EPYC 7402 @ 2.8 GHz
CPU cores	2 × 8 (× 2 threads)	24 (× 2 threads)
primary memory	90 GiB	1.9 TiB
secondary storage	2 × 1.8 TiB HDDs	1.8 TiB Intel® SSD (6 GB/s)
operating system	Linux 4.19.0-14 amd64	Linux 5.15.0-53 x86_64



(a) FENF, $T_{50\%}$. **(b) Reddit,** $T_{75\%}$. **(c) StanDev,** $T_{75\%}$. **(d) NBA,** $T_{25\%}$.

Fig. 1. Reproduction (excerpt) of Figure 11 in [1]: Overall runtime of JSON similarity queries.



(a) FENF, $T_{50\%}$. **(b) Reddit,** $T_{75\%}$. **(c) StanDev,** $T_{75\%}$. **(d) NBA,** $T_{25\%}$.

Fig. 2. Reproduction (excerpt) of Figure 12 in [1]: Filter effectiveness pruned by JSIM index, label intersection, upper bound, and number of verifications.

4.2 Results

The reproducibility framework replays all experiments found in Section 6 of [1], contains additional JSON data sets, and adds experimental runs with further configurations parameters (*e.g.*, distance thresholds). Our local results faithfully match those reported in the key Figures 11 and 12 of the original paper. We have reprodced an excerpt of the plots generated on our setup in Figures 1 and 2.

5 SUMMARY

All experimental results turned out to be reproducible on our machine. The authors are to be commended for the care and effort that went into preparation, packaging, and documentation.

REFERENCES

[1] Thomas Hütter, Nikolaus Augsten, Christoph M. Kirsch, Michael J. Carey, and Chen Li. 2022. JEDI: These aren't the JSON documents you're looking for.... In *Proceedings of the ACM SIGMOD/PODS International Conference on Management of Data*. ACM, Philadelphia, PA, USA, 1584–1597.