Reproducibility Report for ACM SIGMOD 2020 Paper: "Locality-Sensitive Hashing Scheme based on Longest Circular Co-Substring"

JOHN PAPARRIZOS, The University of Chicago, USA
CHUNWEI LIU, The University of Chicago, USA

The work on the reproducibility of this project is praiseworthy. All required dependencies and build steps are carefully noted in the provided git repository. A series of scripts allow to automatically rerun the experiments, reproduce the results, and recreate some of the plots in the paper. The reproduced results are similar to the values reported in the paper and, importantly, all relationships between the compared methods are maintained.

1 INTRODUCTION

This is a reproducibility report for the paper [1]. To summarize, the central results and claims of the paper are supported by the submitted experiments. The key figures have been reproduced accurately enough. The reproducibility scripts is easy to use and well-documented.

2 SUBMISSION

The reproducibility submission consists of the detailed instructions on project dependencies and how to rerun the experiments with a Makefile acting as a command-line entry point for the reviewer. Several Python scripts are provided for running experiments and recreating results. Paper figures can be generated automatically by figure scripts with detailed log files.

The submission contains:

- Github repository with code and scripts at: https://github.com/1flei/lccs-lsh
- Data sources at: https://1drv.ms/u/s!Ascf3jEjrVdxg6c6w7CutkF0TpXgpA?e=fjnR80

3 HARDWARE AND SOFTWARE ENVIRONMENT

Table 1 describes the resources used in the original paper used and our reproducibility effort.

Table 1. Hardware & Software environment

	Paper	Repro Review
CPU	Intel i7-3820	Intel(R) Xeon(R) Gold 6126
Cores	4	12
GHz	3.60	2.60
RAM	64GB	192GB
Platform	Ubuntu 16.04	Ubuntu 16.04
g++	8.3	8.4

4 REPRODUCIBILITY EVALUATION

4.1 Process

The experiments are reproduced on the four datasets (Msong, Sift, Gist, Deep) attached in this submission. The scripts run the proposed approaches and other baselines sequentially on those datasets. The plot scripts parse the produced files and generate figures shown in the paper. It was possible to follow the reproducibility instructions without the authors' help.

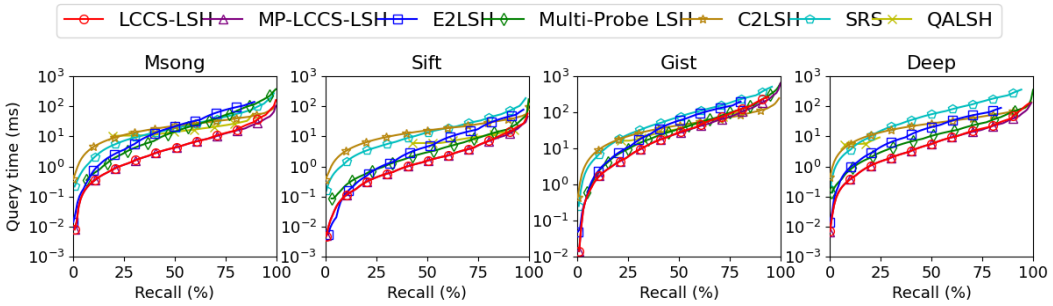

Fig. 1. Query time & recall under Euclidean distance (corresponding to original paper Fig 4)

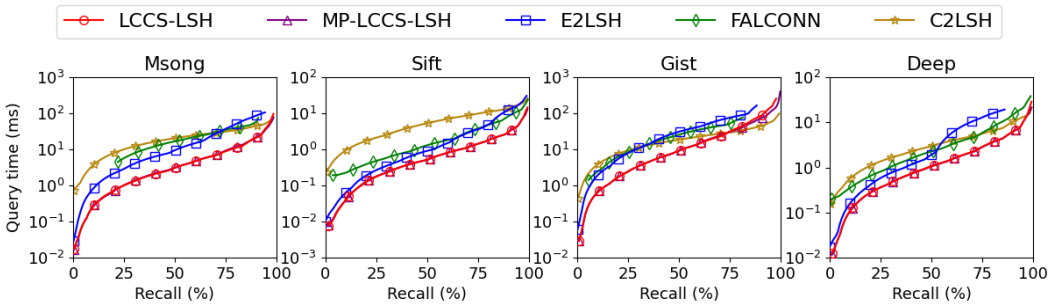

Fig. 2. Query time & recall under Angular distance (corresponding to original paper Fig 5)

4.2 Results

The following figures have been reproduced: Figure 1, Figure 2, Figure 3 and Figure 4. The obtained numbers and the visual plots appear to be close enough to the paper's reported values [1]. The deviation is attributed to the differences in hardware. Most importantly, the relationships between different baselines' performances match the ones reported and discussed in the paper.

5 SUMMARY

The major figures have been reproduced on the reproducibility platform. The ideas, claims, and findings supported by these figures are therefore reproduced as well.

REFERENCES

[1] Yifan Lei, Qiang Huang, Mohan Kankanhalli, and Anthony KH Tung. 2020. Locality-Sensitive Hashing Scheme based on Longest Circular Co-Substring. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 2589–2599.

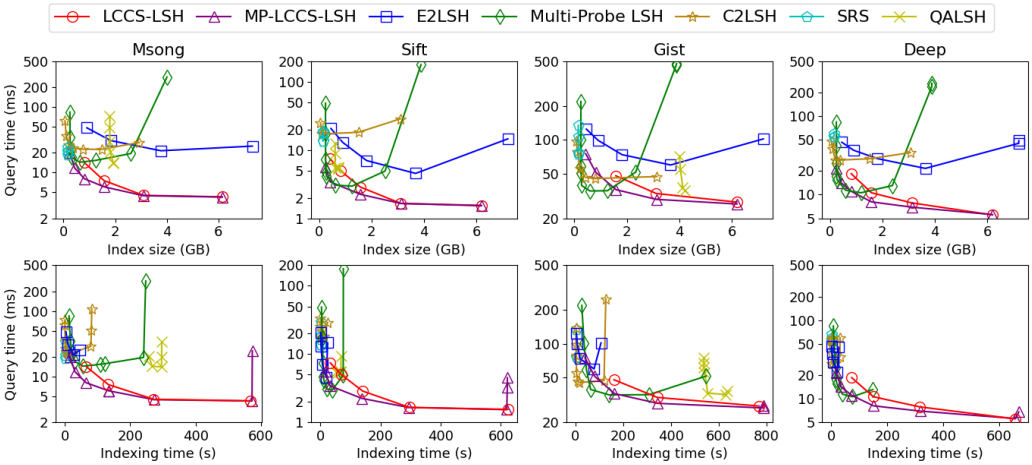

Fig. 3. Indexing time & recall under Euclidean distance (corresponding to original paper Fig 6)

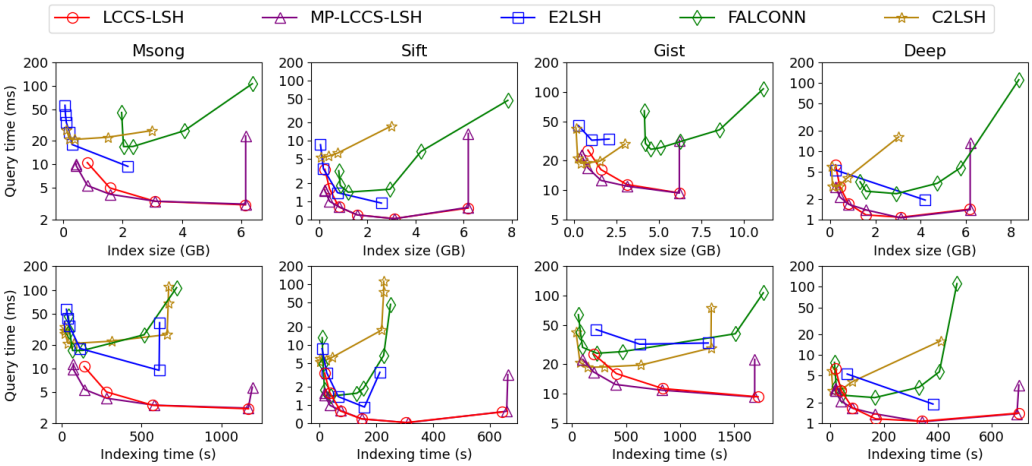

Fig. 4. Indexing time & recall under Angular distance (corresponding to original paper Fig 7)