Reproducibility Report for ACM SIGMOD 2021 Paper: "Putting Things into Context: Rich Explanations for Query Answers using Join Graphs"

MARC SPECKMANN, Leibniz Universität Hannover, Germany
ZIAWASCH ABEDJAN, Leibniz Universität Hannover, Germany

In this reproducibility report on the paper "Putting Things into Context: Rich Explanations for Query Answers using Join Graphs" we present our experience on the reproduction process. In particular, we evaluate the quality of the reproducibility submission and the artifacts and show that we could fully reproduce the results.

1 INTRODUCTION

The following is a reproduction of the paper "Putting Things into Context: Rich Explanations for Query Answers using Join Graphs" by Chenjie Li, Zhengjie Miao, Qitian Zeng, Boris Glavic and Sudeepa Roy [1]. The paper presents a new approach to explaining query results by augmenting provenance with information from other related tables. In the reproducibility study, we reproduced 14 experiments on 2 datasets measuring the runtime and comparing the query explanations. We fully reproduced the results of these experiments. The experiments were carried out partially on our machines with similar specs as discussed below as well as on their own machines as described in their paper.

2 SUBMISSION

The submission for the reproducibility of the paper contains everything needed for the reproduction. All the code has been made available on GitHub. The repository also contains a detailed README file describing all the steps to reproduce the paper. Furthermore, the repository contains three scripts that can be used to reproduce all figures and tables from the paper on the NBA dataset. The scripts for the experiments on the protected MIMIC become available upon request and the condition that one has undertaken the corresponding ethical training guide of PhysioNET.

In addition to the scripts, there are corresponding Docker containers that can be used to run the experiments with a few commands in an isolated environment.

All in all, the necessary information to reproduce the paper was available. List of submitted content:

- GitHub repository with code and scripts at:
  https://github.com/IITDBGroup/CaJaDe/tree/sigmod_reproducibility
- Detailed README file on the same GitHub page
- Prepared Docker containers
- Description and citation of the datasets, as well as instructions on how to access them.
- Details of the hardware used in the original experiments

3 HARDWARE AND SOFTWARE ENVIRONMENT

Information about the hardware environment for the experiments is given in Table 1. The left column of Table 1 shows the hardware used for the original experiments, and the right column shows the hardware used for reproduction. Due to limited access to the MIMIC dataset, the hardware from the original experiments was used to reproduce this part of the experiments.

The tool (CaJaDE) implemented for the experiments runs with Python 3.6 and works with PostgreSQL 10.14. All dependencies used by CaJaDE can be found in the setup.py file in the repository. As suggested by the authors the ideal environment for a successful deployment is Linux.

Table 1. Hardware & Software environment

	Paper	Repro Review
CPU	2 x AMD Opteron 4238	AMD EPYC 7543
Cores	2 x 6	32
Caches (per CPU)	L1 (288KiB), L2 (6 MiB), L3 (6MiB)	L1 (3MB), L2 (16 MB), L3 (256MB)
GHz	3.3 GHz	2,8 GHz
RAM	128GB DDR3	512GB DDR4
Storage	HDD	SSD & HDD

4 REPRODUCIBILITY EVALUATION

4.1 Process

The detailed description of the steps for reproduction made it easy to carry out the experiments. To run the experiments, only four steps were needed.

(1) Install docker + docker-compose
(2) Clone the GitHub repository
(3) Starting the docker containers via docker-compose
(4) Starting the experiment scripts.

The corresponding batch commands are given in the README file.

The results of the experiments were saved with the same names as the figures and tables in the paper so that the results could be easily compared.

Accessing the MIMIC dataset was a more complicated process as it requires special certification.

However, the research team offered an alternative approach. Under the promise not to query the MIMIC dataset, the corresponding experiments could be carried out on their server.

4.2 Results

All in all, we could fully reproduce all results presented in the paper. There are slight deviations in the query explanations, but these do not change the statement of the findings. The slight deviations because of the randomized initialization process can cause the differences.

Due to the different hardware, the measured times for the calculations on the NBA dataset differ. As our hardware was more powerful, the runtime results we obtained were generally by a factor $1.5 - 2.5$ faster. Yet the relative performance differences could be reproduced.

For the MIMIC experiments, we obtained nearly identical runtime results as those were carried out on the original hardware.

5 SUMMARY

In summary, the paper "Putting Things into Context: Rich Explanations for Query Answers using Join Graphs" can be fully reproduced. The available GitHub repository and the proper documentation by the research team contributed significantly to this. It is also worth mentioning that questions were answered quickly and comprehensively.

REFERENCES

[1] Chenjie Li, Zhengjie Miao, Qitian Zeng, Boris Glavic, and Sudeepa Roy. 2021. Putting Things into Context: Rich Explanations for Query Answers using Join Graphs. In *SIGMOD '21: International Conference on Management of Data, Virtual Event, China, June 20-25, 2021*, Guoliang Li, Zhanhuai Li, Stratos Idreos, and Divesh Srivastava (Eds.). ACM, 1051–1063. https://doi.org/10.1145/3448016.3459246