RepEng: Reproducibility of the Project JSON Schema Discovery

Konrad Drees
University of Passau
Passau, Germany
drees03@ads.uni-passau.de

Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/kon-drees/RepEng-JSONSchemaDiscovery.

1 INTRODUCTION

This reproduction study looks at the research and experiments detailed in the paper "An Approach for Schema Extraction of JSON and Extended JSON Document Collections"[1]. The original study presented a methodology for schema discovery in NoSQL databases, with a focus on JSON and Extended JSON documents. NoSQL databases operate without a fixed schema due to their flexibility in data format handling [2]. Therefore NoSQL databases have high availability and elasticity capabilities but also introduces challenges in data organisation, data analysis and retrieval [3]. One of the NoSQL Databases are Document-oriented Databases, which store and retrieve documents with simple and complex attributes mainly in JSON (JavaScript Object Notation) or Extended JSON formats[1]. This Project aims to methodically validate the methodology and evaluates the reproducibility of the original paper by replicating the experiments.

2 REPLICATION OF PROCESSING TIME EVALUATION IN SCHEMA EXTRACTION

This reproduction study aims to replicate and confirm the findings of the original paper regarding the Processing Time Evaluation of the JSON Schema Discovery approach. Specifically, it validates the reported efficiency of this approach in accurately extracting schemas from JSON documents within NoSQL databases. The focus will be on assessing whether the replication of the original experiments shows similar processing times, thereby confirming the method's reliability and efficiency as reported in the initial study.

2.1 Methodology

Table 1: Results for Foursquare Datasets of the original Paper [1].

Collection	N_JSON	RS	ROrd	TB	TT	TB/TT
venues	2 mil	257	117	7,47 min	7,52 min	99,33%
checkins	11 mil	2	2	35,27 min	35,52 min	99,29%
tweets	17 mil	23	16	53,11 min	53,44 min	99,38%

For evaluating the processing time of JSON Schema Discovery, this reproduction study will be conducted using a Docker container on a local machine. The experimental setup will replicate the original environment to the extent possible, given the hardware and software differences. The datasets used will be similar to those in the original study, including tracked data from tweets, check-ins, and venues from Foursquare. The Docker environment will ensure a controlled, consistent, and replicable setting for comparing processing times against those reported in the original study on an Amazon EC2 instance

Table 1 shows the results of the experiments of the original study [1]. The criteria for confirming the experiment involve matching the original study's reported ratio **TB/TT** between Time to Obtain the raw schemas (TB) and Total Time (TT). A similar TB/TT ratio would indicate successful replication of processing efficiency.

2.2 Analysis of the Reproducibility Process

In order to evaluate reproducibility of the original study, a replication package will be created. The analysis contains the methodological steps taken and the challenges encountered, such as obtaining equivalent datasets, running the NoSQL Database, ensuring computational reproducibility, and using the original codebase within a local Docker environment.

The analysis will also evaluate the documentation and artefact availability from the original study.

REFERENCES

[1] Angelo Augusto Frozza, Ronaldo dos Santos Mello, and Felipe de Souza da Costa. [n.d.]. An Approach for Schema Extraction of JSON and Extended JSON Document Collections. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)* (2018-07). IEEE. https://doi.org/10.1109/iri.2018.00060

[2] Pramod J. Sadalage and Martin Fowler. [n.d.]. *NoSQL distilled.* Addison-Wesley. Hier auch später erschienene, unveränderte Nachdrucke.

[3] Martin Wischenbart, Stefan Mitsch, Elisabeth Kapsammer, Angelika Kusel, Birgit Pröll, Werner Retschitzegger, Wieland Schwinger, Johannes Schönböck, Manuel Wimmer, and Stephan Lechner. [n.d.]. User Profile Integration Made Easy: Model-Driven Extraction and Transformation of Social Network Schemas. In *Proceedings of the 21st International Conference on World Wide Web* (Lyon, France, 2012) *(WWW '12 Companion)*. Association for Computing Machinery, New York, NY, USA, 939–948. https://doi.org/10.1145/2187980.2188227