Json Schema Extraction Report

Fabien Kavuganyi
University of Passau
Passau, Germany
kavuga01@ads.uni-passau.de

Artifact Availability:
find all related projects data on this repository.
https://github.com/Fabu1da/JSONSchemaDiscoveries

1 INTRODUCTION

In today's business landscape, data has emerged as the most pivotal asset. This was first articulated by British mathematician Clive Humby in 2006 when he famously stated "Data is the new oil" [4]. However, as the pace of the world accelerates, data's significance seems to eclipse even that of gold. The importance of having clean and reliable data cannot be overstated for companies, as it serves as a key driver for insights and future predictions[3].

This report details an experiment on JSON Schema Extraction[2], a format widely used by NoSQL databases to manage unstructured data. NoSQL stands out for its performance capabilities, especially in handling large volumes of data efficiently. Ensuring the integrity and verifying the accuracy of data stored in JSON format is crucial. This is achieved by establishing clear rules and constraints for the JSON data, thereby enhancing its reliability.

2 HYPOTHESIS

The core objective of this experiment is to reproduce the process of automated schema extraction, specifically in the context of a NoSQL database environment like MongoDB. This aims to determine if the processing times observed in the reproduced package match those reported in the original study, thereby affirming the method's claimed reliability and effectiveness. Additionally, this replication effort seeks to validate the accuracy and efficiency of the schema extraction process as applied to diverse datasets typical for a NoSQL context. It also intends to assess any scalability issues and the impact of different data volumes and structures on the performance of the schema extraction method. [1]

3 METHODOLOGY

Datasets	N JSON	RS	ROrd	TB/min	TT/min	TB/TT
Venues	2 mil	257	117	7,47	7,52	99,33%
Checkins	11 mil	2	2	35,27	35,52	99,29%
Tweets	17 mil	23	16	53,11	53,44	99,38%

Table 1: Original results from JSON Schema Extraction paper

N JSON - Number of JSON documents. RS - Raw schemas. ROrd - Raw schemas with ordered structure. TB - Time to obtain the raw schemas. TT - Total time.

We aim to refine the methodology of JSON Schema Extraction by leveraging the ease and flexibility of Docker containers on local machines. This approach enables us to replicate the original project environments meticulously, ensuring that our experimental setup mirrors the original conditions as closely as possible. By applying necessary patches and maintaining project functionality, we ensure a high degree of accuracy and reproducibility in our work.

Our experiment utilizes datasets that closely match the original study's in terms of check-ins, tweets, and venues, maintaining consistency and relevance in our data analysis. This careful selection underpins our comparative analysis, allowing for a direct assessment of the enhancements made to the schema extraction process.

4 ANALYSIS

In our reproducibility study, we diligently assembled a replication package, steering through a pathway marked by systematic steps and distinctive challenges. Our work encompassed acquiring datasets parallel to those in the original study, ensuring the NoSQL Database operated efficiently, attaining computational reproducibility, and deploying the original study's code within a custom-configured Docker environment tailored specifically to our investigative needs.

Additionally, we engaged in a detailed review of the original project documentation and artifacts, concentrating on assessing their accessibility and the quality provided. This critical appraisal was aimed not only at affirming the initial results but also at emphasizing the paramount importance of transparent and thorough documentation in aiding replication efforts, thus speaking to the wider implications for maintaining research transparency and upholding integrity within the scientific community.

In this context, I managed to carry out two of the three experiments suggested in the cited paper[1], with a particular focus on the Foursquare datasets. The conducted experiments included:

- An in-depth analysis of Firenze Venues
- A detailed examination of Firenze Checkins

Due to the limitations in data accessibility, the sample sizes for the Firenze Venues experiment were restricted to 332 and 900, respectively, and the tweets dataset was not publicly available. This adjustment not only highlights the flexibility required in conducting replication studies but also showcases our dedication to precision and the meticulous approach that defines our commitment to upholding the standards of scientific inquiry and replication fidelity.

5 SUMMARY

The objective of this experiment was to reproduce an existing project on JSON schema extraction, Developed by gbd-ufsc based on feeKosta's foundational project. During the replication process, I encountered a myriad of challenges, including errors and compatibility issues with versions, which necessitated a thorough review and deep dive into the source code. This critical examination led to the identification and implementation of necessary patches and modifications to rectify the issues and enhance the project's functionality.

To simplify and automate the process is also one of the task, a Docker container was employed, configured to execute a Python script specifically designed for this purpose. This approach not only facilitated a more efficient workflow but also ensured consistency and reproducibility in the results obtained from the schema extraction process.

Furthermore, the study's methodology was diligently crafted to mirror the original project's setup as closely as possible, while also incorporating improvements to address the identified shortcomings. This included a detailed analysis of the project's dependencies, software environment configurations, and the underlying algorithms driving the schema extraction. The aim was to not only replicate the original findings but also to contribute to the ongoing development and understanding of JSON schema extraction methodologies within NoSQL database environments.

The integration of Docker and automation scripts represents a significant advancement in the replication process, offering a scalable and replicable model for future research in this area. By documenting the challenges encountered and the solutions implemented, this study contributes valuable insights and enhancements to the field, paving the way for further exploration and refinement of JSON schema extraction techniques.

To evaluate the time for processing 332 records:

$$Time for 332 records = \frac{7.47 minutes}{2,000,000 records} 332 records \quad (1)$$

Upon computing, the processing time for 332 records was found to be approximately 0.00124 minutes, aligning closely with the experimental results obtained in this replication study which was 0.00109 minutes.

6 LIMITATIONS

Benchmarking on AWS EC2 instances serves as a critical evaluation, enabling us to measure the improvements in our methodology against a recognized standard. This comparison highlights the advancements in processing efficiency and data integrity achieved through our refined approach.

The original experiments were conducted on an Amazon EC2 t2.micro instance (Intel R Xeon R E5-2676 v3 @ 2.40GHz and 1GB of RAM). Meanwhile our experiments were conducted on a MacBook Air with an Apple M2 chip @ 3.49 GHz and 8GB of RAM.

The absence of the original dataset compromised the accuracy of the experiment, leading to a significant reduction in the number of records processed but yielding results that were roughly comparable to the original study. The estimated processing time for 2 million records in the original study was 7.42 minutes.

7 CONCLUSION

In conclusion, the replication study, despite facing several inconsistencies and challenges, has proven to be successful. This is evidenced by the improved processing efficiency observed in my reproduction package compared to the original study. The benchmark results, though not quantified in exact minutes, still highlight a noticeable enhancement, underscoring the successful mirroring of the original experiment's findings. This slight improvement in efficiency, while slight, confirms the effectiveness of the replication effort and underscores the value of the exact research in validating and enhancing previous studies. The replication not only reinforces the credibility of the original findings but also contributes to the body of knowledge with optimized processes.

REFERENCES

[1] Angelo Augusto Frozza, Ronaldo dos Santos Mello, and Felipe de Souza da Costa. 2018. An Approach for Schema Extraction of JSON and Extended JSON Document Collections. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*. 356–363. https://doi.org/10.1109/IRI.2018.00060

[2] John Keiser and Daniel Lemire. 2023. On-Demand JSON: A Better Way to Parse Documents? arXiv:2312.17149 [cs.DB]

[3] Ga Young Lee, Lubna Alzamil, Bakhtiyar Doskenov, and Arash Termehchy. 2021. A Survey on Data Cleaning Methods for Improved Machine Learning Model Performance. arXiv:2109.07127 [cs.DB]

[4] Christoph Stach. 2023. Data Is the New Oil - Sort of: A View on Why This Comparison Is Misleading and Its Implications for Modern Data Administration. *Future Internet* 15, 2 (2023). https://doi.org/10.3390/fi15020071