Generative AI for Graph Data Repair

Advisors

:

- Angela Bonifati (Lyon 1 Univ., angela.bonifati@univ-lyon1.fr)
- Andrea Mauri (Lyon 1 Univ., andrea.mauri@univ-lyon1.fr)

Context

Graph data is ubiquitous in several application domains, such as life sciences, finance, security, logistics, and planning. To ensure the quality of the data, several approaches were developed to express graph constraints of different expressiveness to detect potential glitches and adopt automatic repairing strategies.

However, these methods require both to know the shape of the data (i.e., the schema) in advance, which may be not possible (or is totally absent) and to know the constraints. Defining the constraints can be difficult, time-consuming, and prone to error since it requires knowledge from domain experts.

Objectives

The objective of this project is to study how generative AI, especially Generative Adversarial Neural Networks and Large Language Models, can be used to data repair rules (e.g., consistency constraints) in graphs.

While other machine learning methods need a large quantity of annotated data to be trained (that is expensive or even impossible to obtain), self-supervised methods as the one mentioned above can learn patterns from the data and infer consistency rules that can be used to clean the data [1].

However, the challenge consists in the fact that the data itself may be dirty and, as a consequence, the model may learn wrong or inaccurate rules. For this reason, it is crucial to design and implement methods to detect those false rules and remove or correct them.

These methods can be either based on the discriminator component of a GAN [1] or human judgments (in a human-in-the-loop fashion)[3,4].

The project can be either a comparison between GAN and LLM or focus on a single technology (e.g., only LLM [8]). In this case, multiple models should be considered (e.g., GPT-4 vs. FLAN-T5).

Task

- Familiarize yourself with the concept of data cleaning [7] and graphs [2]. Possibly, select a subset of consistency rules to address [5,6].
- Investigate how a GAN and a LLM can be used to infer the consistency rules. What are the inputs? How does the graph need to be encoded?
- Design the experiment(s) to evaluate the effectiveness of the models you chose. What metrics should be considered? Which data? How difficult is it to detect and correct the wrong rules inferred by the models?

Requirements

- Good programming skills (Python, C++)
- Familiarity with machine learning technologies
- Familiarity with graph databases

References

[1] Peng, J., Shen, D., Tang, N., Liu, T., Kou, Y., Nie, T., ... & Yu, G. (2022). Self-supervised and Interpretable Data Cleaning with Sequence Generative Adversarial Networks. Proceedings of the VLDB Endowment, 16(3), 433-446.
[2] Angela Bonifati, George H. L. Fletcher, Hannes Voigt, Nikolay Yakovets: Querying Graphs. Synthesis Lectures on Data Management, Morgan & Claypool Publishers 2018
[3] Mauri, A., & Bozzon, A. (2021). Towards a Human in the Loop Approach to Preserve Privacy in Images. In IIR.
[4] Bozzon, A., Brambilla, M., Ceri, S., Mauri, A., & Volonterio, R. (2014, July). Pattern-based specification of crowdsourcing applications. In the International Conference on Web Engineering (pp. 218-235). Springer, Cham.
[5] Shimomura, L. C., Fletcher, G., & Yakovets, N. (2020, October). Ggds: Graph generating dependencies. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management (pp. 2217-2220).
[6] Parisi, F., & Grant, J. (2021). On Database Inconsistency Measures. In 29th Italian Symposium on Advanced Database Systems, SEBD 2021 (Vol. 2994, pp. 200-208). CEUR-WS.
[7] Wenfei Fan, Ping Lu: Dependencies for Graphs. ACM Trans. Database Syst. 44(2): 5:1-5:40 (2019)
[8] Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, Damien Graux:
Large Language Models and Knowledge Graphs: Opportunities and Challenges. CoRR abs/2308.06374 (2023)