Causal Dataset Discovery with Large Language Models

Junfei Liu

B.S Computer Science, B.A. Economics, Class of 2024.

01 Abstract

This poster shows an investigation on application of newly emerging large language models on a novel causal task: discovering inter-table causal relations, which we name Causal Dataset Discovery. It is a challenging task especially within large-scale repositories, which requires identifying causal relationships in batches by analyzing columns across multiple tables within diverse datasets.

In this paper, we make the following contributions:

- We propose and formulate the data lake causal discovery problem.
- We propose a novel join-based causal discovery method for extracting potential causal link candidates across datasets between numerical variables and ultimately provide causal dataset search among large-scale datasets through large language models.
- We create a benchmark designed to evaluate causal dataset discovery solutions and provide empirical evaluations over benchmarks on our methodologies.
- We discuss the limitations of this study and suggest potential improvements and extensions of the causal discovery problem.

02 Problem Definition

Definition 1 (Correlation Link over Join)

In this work, we define that there is a correlation link between a numerical column X from a table T_1 and a numerical column Y from a table T_2 if X and Y have a correlation coefficient higher than a threshold C in table $T_1 \bowtie T_2$ where T_1 and T_2 are joined on categorical columns K_X from a table T_1 and K_Y from a table T_2, expressed as $(T_1\langle K_X, X\rangle, T_2\langle K_Y, Y\rangle)$.[1]

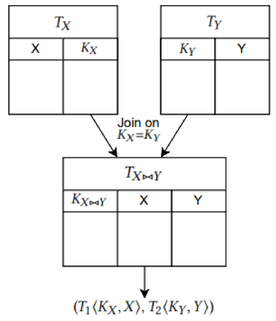
Definition 2 (Causal Link over Join)

A causal link over join exists between a numerical column X in T_Q and a numerical column Y in T_C if:

(1) There is a correlation link between X and Y in the joined table $T_Q \bowtie_{K_X=K_Y} T_C$ and this correlation exceeds a predefined threshold C.

(2) Post the application of causal discovery algorithms, the link between X and Y is confirmed as causal, with its direction being either clearly established or not determined.

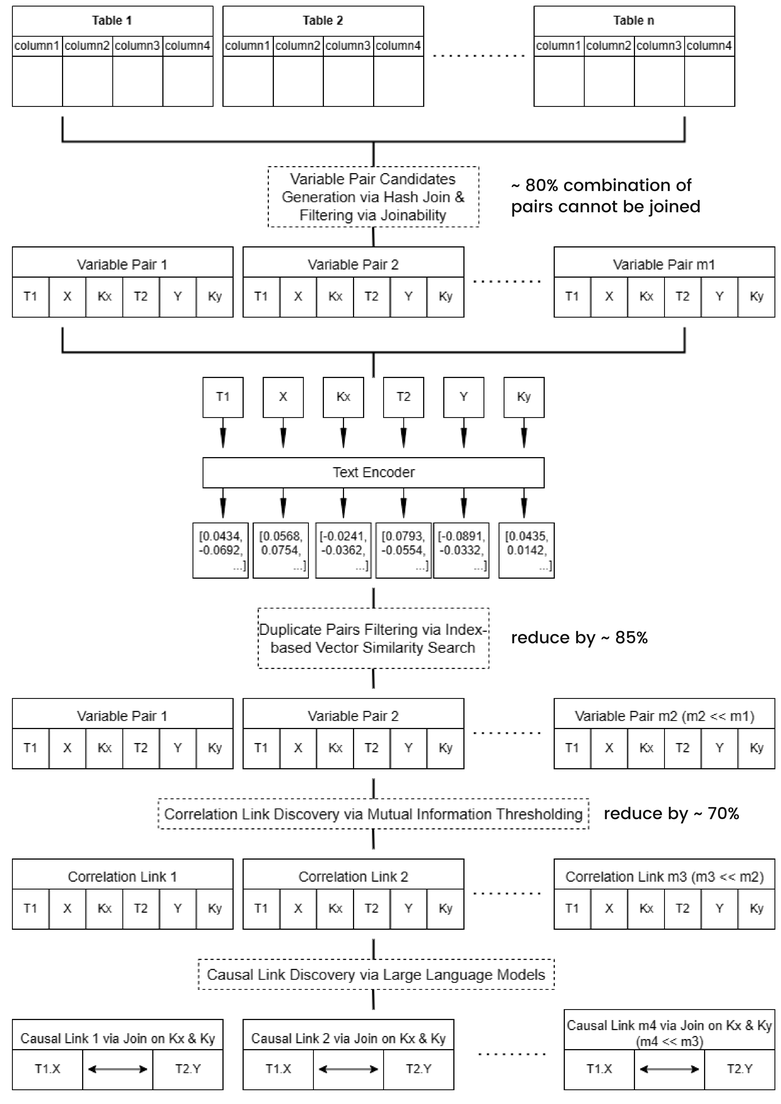
And column-level causal links over join can be expressed as $(T_1\langle K_X, X\rangle, \leftrightarrow, T_2\langle K_Y, Y\rangle)$, where the arrow represents the direction of causation. Additionally, we define a directed causal link exists between a table T_1 and a table T_2 if there exists a column-level causal link between an arbitrary numerical column X in table T_1 and Y in table T_2.



Definition 3 (Causal Dataset Discovery)

We introduce the causal dataset discovery problem as the search for the top-k datasets within the repository that have a directed causal link to the query dataset. Given a query dataset T_Q and a repository of datasets, the objective is to identify and rank the top-k datasets T_C within the repository by the number of directed causal links to T_Q. The desired output is a ranked list of datasets T_C, each contains a list of causal links $(T_Q\langle K_X, X\rangle, \leftrightarrow, T_C\langle K_Y, Y\rangle)$.

03 Methodology



Causal relations is naturally scarce. Although LLMs are powerful, they are meanwhile computation heavy and we can't just throw every possible combination of columns to it! To reduce the number of candidates, we did 1) filter out pairs that can't be joined; 2) reduce # of similar pairs; 3) identify correlation links that serve as potential causal links over join.

04 Experiments

Benchmarks are constructed on the remaining pairs after the joinability and similarity filtering. Ground truth is obtained through majority voting of 2 human labelers and GPT-4. We access the performance of different LLM strategies on Causal Dataset Discovery problem.

Table 1: Benchmark Statistics

	Benchmark#1	Benchmark#2	Benchmark#3
# of source tables	20	40	64
total # of pairs	1095	1265	1838
# of pairs pass MI threshold	312	405	668
# of Positive causal relations	61	12	78

Table 2: Causal Links Discovery Results

	Benchmark#1				Benchmark#2				Benchmark#3			
Measurement	**precision**	**recall**	**Accuracy**	**F-1 Score**	**precision**	**recall**	**Accuracy**	**F-1 Score**	**precision**	**recall**	**Accuracy**	**F-1 Score**
GPT-3.5 + default prompts	73.97	75.27	74.26	74.26	41.64	52.50	66.67	66.67	59.37	70.16	68.00	68.00
GPT-3.5 + CoT prompts	54.18	54.46	41.58	41.58	36.41	29.79	23.53	23.53	40.67	45.45	27.00	27.00
GPT-3.5 + 3-stage prompts	85.54	75.14	80.20	81.00	77.54	49.52	80.39	82.00	84.69	56.11	78.00	80.41
GPT-3.5 finetuned	81.43	78.68	80.20	80.20	75.16	66.76	86.27	86.27	92.69	87.09	94.00	94.00
GPT-4 + default prompts	**97.08**	**93.07**	**95.05**	**95.05**	**91.71**	**83.81**	**94.12**	**94.12**	**97.75**	**88.94**	**95.00**	**95.00**

Table 3: Causal Links Identification Results

	Benchmark#1				Benchmark#2				Benchmark#3			
Measurement	**precision**	**recall**	**Accuracy**	**F-1 Score**	**precision**	**recall**	**Accuracy**	**F-1 Score**	**precision**	**recall**	**Accuracy**	**F-1 Score**
GPT-3.5 + default prompts	80.0	89.80	84.16	84.62	47.62	83.33	74.51	60.61	58.00	**93.55**	77.00	71.60
GPT-3.5 + CoT prompts	52.75	**97.96**	56.44	68.57	25.58	**91.67**	35.29	40.00	32.95	**93.55**	39.00	48.74
GPT-3.5 + 3-stage prompts	**100**	66.67	84.00	80.00	**100**	33.33	84.00	50.00	**100**	37.93	81.44	55.00
GPT-3.5 finetuned	90.70	79.59	86.14	84.78	88.89	66.67	90.20	76.19	96.43	87.10	**95.00**	**91.53**
GPT-4 + default prompts	**100**	89.80	**95.05**	**94.62**	**100**	83.33	**96.08**	**90.91**	**100**	83.87	**95.00**	**91.23**

05 Discussion & Future Study

This study has several limitations and can be improved in future research:

- The benchmark size is limited due to the intensive manual labeling process it requires, which could undermine evaluation results' confidence for application in large data repositories like data lakes.
- There's a lack of data-driven methods to supplement LLMs for causal relation identification. It is suggested to explore causal discovery methods under the causal dataset discovery setting with potential missing variables and confounders while ensuring data correspondence.
- Algorithms that discover joinable tables could potentially be used to reduce the complexity of filtering tables via joinability of current pipeline.
- The effectiveness of LLMs in understanding and applying causality is debated, indicating a need for further research on their capabilities and limitations in causal dataset discovery tasks.
- ...

We also propose an extension of Causal Dataset Discovery problem called Causal Dataset Navigation problem, defined as constructing a navigation graph within a given repository, where the graph materializes all datasets as nodes and pairwise causal links as edges, which is equivalent to solving Causal Dataset Discovery Problem for each table in the repository.

Acknowledgement

I would like to extend my gratitude to Prof. Fatemeh Nargasian for her invaluable guidance and supervision, Mr. Shaotong Sun for his consistent contribution to this project, and Prof. Hangfeng for providing suggestions with his expertise in LLMs. This work would not have been possible without their dedication and commitment to excellence.