A Framework for Crowd-based Causal Analysis of Open Data

Introduction: Many organizations provide open data currently. Important insights can be got through analyzing open data. In the data analyzing tasks, causal relationship analysis is a complex one. We proposed a framework of crowd-based causal analysis which combined the intelligence of the crowd with the state-of-the-art machine learning methods. The proposed framework gives consideration to the effect of possible confounding in causal analysis by collecting explanations of correlation between variables. To verify the collected explanations, a causal discovery workflow was proposed in which conditional independence test and further causal discovery methods are used. We did experiments using data of world bank and open government data. Several interesting causal relationships have been got through analyzing the collected explanations and data using the proposed framework. The experimental results showed that the proposed framework was efficient when doing causal analysis of open data.

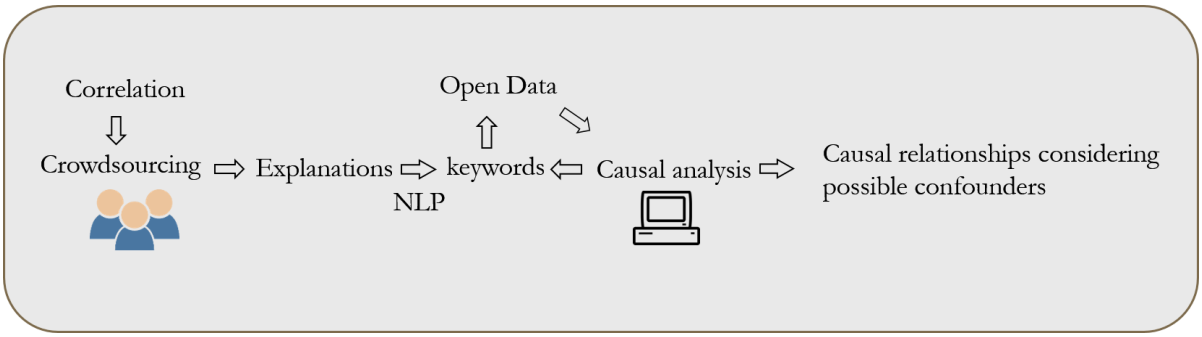

Fig. 1 Proposed Framework

Details of the proposed framework:

- Explanation generation
  - General crowdsouring platform
- Keywords extraction
  - Causal relationship extraction method[1]
  - Latent Dirichlet allocation (LDA) model[2]
- Possible confounder test
  - (Conditional) independence test[3]
  - Intrinsic dimension estimation[4]
- Causal direction learning
  - Information geometric causal analysis (IGCI)[5]

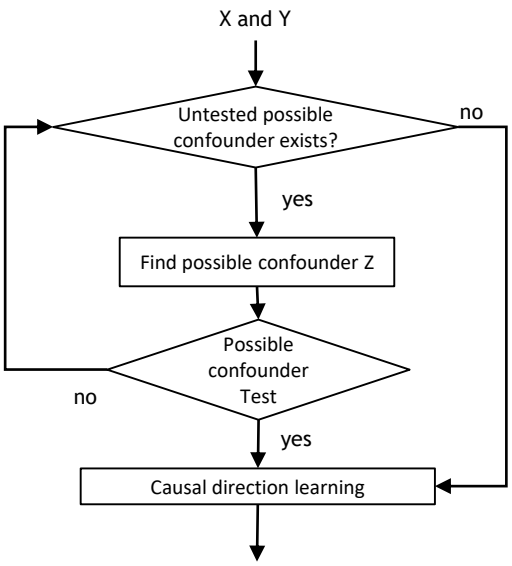

Fig. 2 Proposed causal analysis workflow

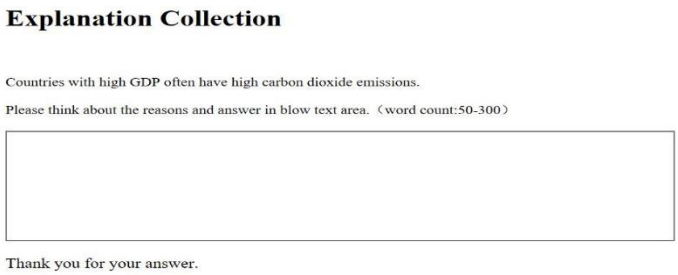

Fig. 3 Crowdsourcing task example

Experimental results:

World bank dataset:

GDP ← Access to electricity → Carbon dioxide emissions

GDP ← Exports of goods and services → Carbon dioxide emissions

GDP ← Imports of goods and services → Carbon dioxide emissions

GDP ← price level ratio of PPP conversion factor (GDP) to market exchange rate → Carbon dioxide emissions

References

1. T. Kudo and Y. Matsumoto, “Japanese dependency analysis using cascaded chunking,” proceedings of the 6th conference on Natural language learning-Volume 20, pp.1-7, Association for ComputationalLinguistics,2002

2.M. Hoffman, F.R. Bach, and D.M. Blei, “Online learning for latent Dirichlet allocation,” Advances in Neural Information Processing Systems,pp.856-864,2010.

3. Zhang, K., Peters, J., Janzing, D. and Schölkopf, B., “Kernel-based conditional independence test and application in causal discovery,” arXiv preprint arXiv:1202.3775, 2012.

4. Lee, J.A. and Verleysen, M., Nonlinear dimensionality reduction. Springer Science & Business Media, 2007

5. Janzing, D., Mooij, J., Zhang, K., Lemeire, J., Zscheischler, J., Daniušis, P., Steudel, B. and Schölkopf, B., “Information-geometric approach to inferring causal directions,” Artificial Intelligence, 182, pp.1-31, 2012.