Finding Root Causes for Outliers in Semantically Annotated Sensor Data

Tim Strobel[1,2][0009−0001−7876−9915], Tim Pychynski[1], and Andreas Harth[2][0000−0002−0702−510X]

[1] Bosch Center for Artificial Intelligence {Tim.Strobel,Tim.Pychynski}@bosch.com
[2] Friedrich-Alexander-Universität Erlangen-Nürnberg {Andreas.Harth}@fau.de

Abstract. Causal inference creates insights into observational data. Such insights could explain an outlying value to perform Root Cause Analysis. But how can causal inference be used with semantically annotated observations? The following demo showcases how to use semantically annotated sensor data for causal inference. The method's implementation uses an agent pattern interacting with a knowledge graph.

Keywords: Sensor Data · Root Cause Analysis · Causal Inference.

1 Introduction

Causal inference explains events in observations (e.g. an outlying value) in more detail than standard statistical analysis [4]. Therefore, it is a helpful tool for performing Root Cause Analysis (RCA) to understand the cause of an undesired event. Our demo[3] shows how to use semantically annotated data to create causal insights into observations. Our contribution includes annotating a causal model based on observations described with the Semantic Sensor Network Ontology (SOSA) [2]. In addition, we show a localized, recursive causal model evaluation using a knowledge graph in the Resource Description Framework (RDF).

2 Causality

According to Pearl [4], a causal graph encodes the dependencies of a system's variables and the direction of their influences on each other. A causal graph represents variables with nodes and dependencies with directed edges [4]. We can define causal mechanisms for each node based on the encoded structure. A causal mechanism describes the behavior of the associated variable with respect to its parent variables in the graph. Based on a causal model, causal inference infers explanations about observed data. A ground truth causal model encodes the physical mechanisms used to generate an observation, but nature hides this ground truth causal model from us [4]. To uncover parts of the ground truth causal model, we use the knowledge of a system's domain expert or algorithms from causal discovery.

[3] Link to the video showing the demo: https://www.tim-strobel.de/eswc24/causality-agents-demo-24.webm

3 Approach

Our approach demonstrates how to add causal dependencies and mechanisms to SOSA observations in an RDF graph. Based on the built RDF graph, we demonstrate causal inference in a recursive, localized manner to find root causes for outliers. The implementation follows a *Simple Reflex Agent* [5] pattern. The agents retrieve data from an RDF graph via SPARQL queries, process the data with pre-set rules, and save inferred conclusions into the graph.

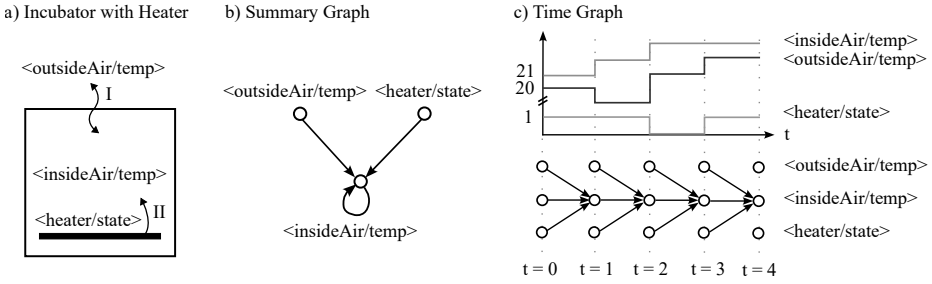

Fig. 1. Simplified incubator example of Feng et al. [1]

We use the running example *Incubator* described by Feng et al. [1] in a simplified version displayed in Figure 1a. The *Incubator* has three *SOSA:Observable Property*, which are <*insideAir/temperature*>, <*outsideAir/temperature*>, and <*heater/state*>. To add the causal knowledge to the RDF graph containing the SOSA annotated observations, instances of *SOSA:ObservableProperty* are also assigned to the class *CausalNode*. We can describe a causal dependency in the RDF graph by linking two causal nodes via a *CausalEdge*. A causal mechanism can be added using the class *CausalMechanism*. We can explain outliers through causal inference by adding causal nodes, edges, and mechanisms to the observations in the RDF graph.

RCA uses a set of observations for each time-step for every *SOSA:Observable Property* (as displayed in the time-dependent graph in Figure 1c). We sample observations using a simulated ground truth causal model for the *Incubator*. To build this ground truth, we use the assumptions by Feng et al. [1]. This model defines heat exchange between <*insideAir*> and <*outsideAir*> (see Figure 1a/I). In addition, heat is exchanged between <*heater*> and <*insideAir*> (see Figure 1a/II). To distinguish the direction of influences, we assume the following: Changing <*insideAir/temperature*> will not affect <*outsideAir/temperature*> as well as changing <*insideAir/temperature*> will not affect <*heater/state*>. Also, we add a self-cycle on <*insideAir/temperature*> since this observable depends on earlier observations in the time series. The resulting summary causal graph is referenced in Figure 1b. The according rolled-up, time-dependent causal graph is shown in Figure 1c.

3.1 Regression Agent

The *Regression Agent* retrieves the causal dependencies and the observations from the RDF graph to fit regression models as causal mechanisms for each *CausalNode*. The regression model's inputs are the parent *CausalNode* instances. The temperature measurement within the *Incubator* is a time-series observation. Therefore, the agent fits an linear auto-regressive model for the *CausalNode* <*insideAir/temperature*> with inputs <*outsideAir/temperature*> and <*heater/state*>. The auto-regressive approach will assume a time lag of 1. In the future also more sophisticated methods could be used to determine the time lag automatically. The time lag considers that a new value is predicted based on the value one time instance before the prediction value in the time series (see time t in Figure 1c). The agent compresses and adds the regression model with a specific URI into the RDF graph.

3.2 Outlier Explanation Agent

The *Outlier Explanation Agent* uses a method described by Janzing et al. [3]. This approach utilizes an Additive Noise Model (ANM) and Shapley values [6]. ANMs take into account that each observation contains an amount of noise. To calculate the amount of noise of one observation, we compute the difference between the observation we made and the estimation of our causal mechanism – the prediction of the regression model. High noise, therefore, means that a high amount of the observation is not explainable by the parent nodes, which are influencing the observed value. With the *Incubator* example, we use the regression model of <*insideAir/temperature*> with the parent variable's observations as input and compute an estimation. With the computed estimation, we calculate the noise as the difference between the estimation and the actual observation for <*insideAir/temperature*>. For all variables with no parent variables and therefore no known causal mechanism to explain the observation, we use the observation's value as the noise feature (e.g., for <*outsideAir/temperature*>).

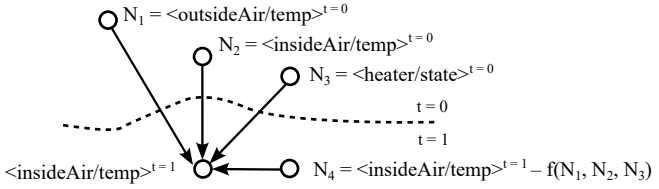

Fig. 2. Computing noise terms using the causal mechanism f

To compute the contribution of each noise feature to our outlying value, Janzing et al. [3] use Shapley values. To do so, we need to describe our outlying observation of <*insideAir/temperature*> as a function of all associated noise

features. We implemented the noise-dependent function in a recursive and localized manner using the RDF graph (see Algorithm 1). For the *Incubator* this function is $<insideAir/temp>^{t=1} = f(N_{parents}) + N_4$ (see Figure 2) with causal mechanism f. Noise terms with high contributions are assumed to be the root causes of the outlier. As visualized in Figure 2), four noise terms are potential root causes for the outlier on $<insideAir/temperature>$ at the investigated time $t = 1$.

Algorithm 1: Localized, Recursive Noise-dependent Function

Data: Target Node T, Noise Samples N, RDF Graph G
Result: Prediction d for Target Node T based on Noise Samples
Function `NoiseDependentFunction(`T`):`
  $parents = G.queryParents(T)$;
  if $len(parents) == 0$ **then**
    return $N[T]$;
  else
    $model = G.queryModel(T)$;
    $input =$ [`NoiseDependentFunction(`p`)` for p in $parents$];
    return $N[T] + model.estimate(input)$;
  end

4 Conclusion

We showed causal inference on causal annotated SOSA observations. The limitations of this approach include that the causal structure needs to be known. A wrongful causal structure, therefore, could lead to wrong conclusions. Our future research includes studying how semantically annotated observations under interventions may help to create causal insights.

Acknowledgements The authors would like to thank Daniel Henselmann for his helpful feedback.

References

1. Feng, H., Gomes, C., Thule, C., Lausdahl, K., Sandberg, M., Larsen, P.G.: The incubator case study for digital twin engineering. arXiv preprint arXiv:2102.10390 (2021)
2. Janowicz, K., Haller, A., Cox, S.J., Le Phuoc, D., Lefrançois, M.: Sosa: A lightweight ontology for sensors, observations, samples, and actuators. Journal of Web Semantics **56**, 1–10 (2019)
3. Janzing, D., Budhathoki, K., Minorics, L., Blöbaum, P.: Causal structure based root cause analysis of outliers. arXiv preprint arXiv:1912.02724 (2019)
4. Pearl, J., Verma, T.S.: A theory of inferred causation. In: Studies in Logic and the Foundations of Mathematics, vol. 134, pp. 789–811. Elsevier (1995)
5. Russell, S.J., Norvig, P.: Artificial intelligence a modern approach. London (2010)
6. Shapley, L.S., et al.: A value for n-person games (1953)