SemCaDo: a serendipitous causal discovery algorithm for ontology evolution

Montassar Ben Messaoud
LARODEC, Institut Supérieur de Gestion,
LINA, Ecole Polytechnique
de l'Université de Nantes.
benmessaoud.montassar@hotmail.fr

Philippe Leray
LINA, UMR 6241, France,
Ecole Polytechnique de l'Université de Nantes.
philippe.leray@univ-nantes.fr

Nahla Ben Amor
LARODEC, 41, Avenue de la liberté,
2000 Le Bardo, Tunisie.
nahla.benamor@gmx.fr.

Abstract

With the rising need to reuse the existing knowledge when learning Causal Bayesian Networks (CBNs), the ontologies can supply valuable semantic information to make further interesting discoveries with the minimum expected cost and effort. In this paper, we propose a cyclic approach in which we make use of the ontology in an interchangeable way. The first direction involves the integration of semantic knowledge to anticipate the optimal choice of experimentations via a serendipitous causal discovery strategy. The second complementary direction concerns an enrichment process by which it will be possible to reuse these causal discoveries, support the evolving character of the semantic background and make an ontology evolution.

1 Introduction

Bayesian networks (BNs), first introduced by Pearl [Pearl, 1988], are compact graphical probabilistic models that are capable of modeling domains comprising uncertainty. Due to the Markov equivalence property, when learning the Completed Partially Directed Acyclic Graph (CPDAG) from observational data and randomly choosing one possible complete instantiation in the equivalence space, we are not so sure if that graph reflects the true causal structure. For this reason, an extension of traditional BNs is introduced, where each edge is interpreted as a direct causal influence between a parent node and a child node, relative to the other nodes in the network [Pearl, 2000].

This paper provides a substantially extended version of our previous work [Ben Messaoud *et al.*, 2009] in which we introduce the preliminary findings for integrating a semantic distance calculus to choose the appropriate interventions. Further developments along this direction have been made in order to deploy more efficient strategies to integrate the semantic prior knowledge, improve the causal discovery process and reuse the new discovered information to support the ontology evolution.

The remainder of this paper is arranged as follows: Section 2 gives the necessary background for both CBNs and ontologies. Section 3 sets out how to use the ontological knowledge to enhance the causal discovery and vice versa. In Section 4, we show simulation results to evaluate the performances of the proposed algorithm. Concluding remarks and future works are given in Section 5.

2 Basic concepts & background

2.1 Ontologies

There are different definitions in the literature of what should be an ontology. The most notorious was given by Tom Gruber [Gruber, 1995], stipulating that an ontology is an *explicit* specification of a *conceptualization*. The "conceptualization", here, refers to an abstract model of some phenomenon having real by identifying its relevant concepts. The word "explicit" means that all concepts used and the constraints on their use are explicitly defined.

In the simplest case, an ontology describes a hierarchy of concepts (i.e. classes) related by taxonomic relationships (is-a, part-of). In more sophisticated cases, an ontology describes domain classes, properties (or attributes) for each class, class instances (or individuals) and also the relationships that hold between class instances. It is also possible to add some logical axioms to constrain concept interpretation and express complex relationships between concepts.

Hence, more formally, an ontology can be defined as a set of labeled classes C=$\{C_1, ..., C_n\}$, hierarchically ordered by the subclass relations (i.e. is-a, part-of relations). For each concept C_i we identify k meaningful properties p_j, where j $\in$ [1, k]. We use H_i to denote the finite domain of instance candidates with each concept C_i and c_i to denote any instance of C_i. We also use R to represent the set of semantical (i.e non-hierarchical) relations between concepts and R_c to represent the subset of causal ones. Finally, formal axioms or structural assertions $<c_i, c_j, s>$ can be included, where $s \in S$ is a constraint-relationship like "must, must not, should, should not, etc".

Practically speaking, the ontologies are often a very large and complex structure, requiring a great deal of effort and expertise to maintain and upgrade the existing knowledge. Such proposals can take several different forms such as a change in the domain, the diffusion of new discoveries or just an information received by some external source [Flouris *et al.*, 2008].

There are many ways to change the ontology in response

to the fast-changing environment. One possible direction is the ontology evolution which consists in taking the ontology from one consistent state to another by updating (adding or modifying) the concepts, their properties and the associated relations [Khattak *et al.*, 2009].

The ontology evolution can be of two types [Khattak *et al.*, 2009]:

- Ontology population: When new concept instances are added, the ontology is said to be populated.
- Ontology enrichment: Which consists in updating (adding or modifying) concepts, properties and relations in a given ontology.

In order to establish the context in which the ontology evolution takes place, the principle of ontology continuity should be fulfilled [Xuan *et al.*, 2006]. It supposes that the ontology evolution should not make false an axiom that was previously true. When changes do not fulfill the requirement of ontological continuity, it is not any more an evolution, it is rather an ontology revolution.

2.2 Causal Bayesian Networks

A Causal Bayesian Network (CBN) is a Directed Acyclic Graph (DAG) where the set of nodes V represents discrete random variables X=$\{X_1, X_2,.., X_n\}$ and the set of edges E represents causal direct dependencies over V. We use D_i to denote the finite domain associated with each variable X_i and x_i to denote any instance of X_i. We denote by $Pa(X_i)$ the set of parents nodes for X_i and $Nei(X_i)$ the set of its neighboring variables.

To learn CBNs, the observational data don't contain enough information to discover the true structure of the graph, which will be restricted to the Completed Partially Directed Acyclic Graph (CPDAG). Thus we have to collect further information on causality via interventions (i.e. actions tentatively adopted without being sure of the outcome). Here, we should note that intervening on a system may be very expensive, time-consuming or even impossible to perform. For this reason, the choice of variables to experiment on can be vital when the number of interventions is restricted.

3 SemCaDo: a serendipitous causal discovery algorithm for ontology evolution

Generally, in the research area, scientific discoveries represent a payoff for years of well-planned works with clear objectives.

This affirmation did not exclude the case of other important discoveries that are made while researchers were conducting their works in totally unrelated fields and the examples are abundant from Nobel's flash of inspiration while testing the effect of dynamite to Pasteur brainstorm when he accidentally discovered the role of attenuated microbes in immunization.

In this way, we propose a new causal discovery algorithm which stimulates serendipitous discoveries when performing the experimentations using the following CBN-Ontology correspondences.

3.1 CBNs vs Ontologies

One of the main motivations when realizing this work is the similarities between CBNs and ontologies. This is particularly true when comparing the structure of the two models as shown in the following correspondences:

1. Nodes (V_i) $\leftrightarrow$ Concepts (C_i): The ontology concepts, which are relevant to the considered domain are represented by the nodes of the CBN.
2. Random variables (X_i) $\leftrightarrow$ Concept attributes ($C_i.p_j$): All random variables in the CBN are represented as specific concept attributes in the ontology.
3. Causal dependencies (E) $\leftrightarrow$ Semantic causal relations (R_c): The correspondence between the two models in term of causality will be as follows:
   - A causal relation between two concepts in the ontology will be represented by a directed link between the corresponding CBN nodes. It is read as $c_Y.p_j$ is the direct consequence of $c_X.p_j$, where p_j is the concept attribute used to make the correspondence.
   - A causal dependency represented by a directed link in the CBN will be represented by a specific causal relation between the appropriate concepts in the ontology.
4. Observational or experimental data ($D_{obs,int}$) $\leftrightarrow$ Concept-attribute instances ($c_i.p_j$): We make a correspondence between the observational (resp. interventional) data at our disposal and the instances of the domain ontology. Each observation (resp. intervention) can be viewed, in the ontological context, as a state instantiation of a given concept attribute.

3.2 SemCaDo Sketch

Our approach relies on extending the MyCaDo algorithm [Meganck *et al.*, 2006] in order to incorporate available knowledge from domain ontologies. The original character of the SemCaDo (Semantic Causal Discovery) algorithm is essentially its ability to make impressive discoveries and reuse the capitalized knowledge in CBNs.

The correspondences between CBNs and ontologies in SemCaDo must respect the following constraints:

- Only a single ontology should be specified for each causal discovery task.
- Each causal graph node must be modeled by a corresponding concept in the domain ontology. The concepts which are candidates to be a member of such correspondence have to share the same studied attribute p_j.
- The causal discoveries concern concepts sharing the same semantic type (e.g. direct transcriptional regulation between genes). This means that all concepts C_i modeled in the CBN must belong to the same superconcept SC and the causal relationship under study R_c should be defined for any element of SC to any other one.
- The ontology evolution should be realized without introducing inconsistencies or admitting axiom violations.

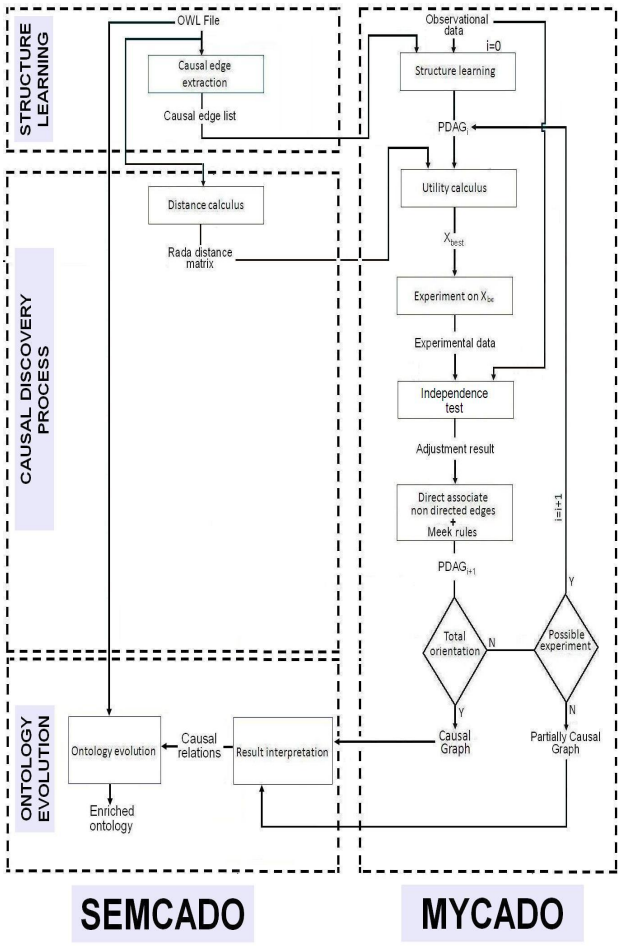

Figure 1: The SemCaDo Algorithm

The general overview of the SemCaDo algorithm is given in Figure 1. As inputs, SemCaDo needs an observational dataset and a corresponding domain ontology. Then it will proceed through three main phases:

1) Learning the initial structure using causal prior knowledge:

The ontology in input may contain some causal relations in addition to hierarchical and semantic relations. Those causal relations should be integrated from the beginning in the structure learning process in order to reduce the task complexity and better the final output. Therefore, such direct cause to effect relations will be incorporated as constraints when using structure learning algorithms. Our main objective is to narrow the corresponding search space by introducing some restrictions that all elements in this space must satisfy.

In our context, the only constraint that will be defined is edge existence. But we could also imagine in future work that some axioms in the ontology also give us some information about forbidden edges. All these edge constraints can easily be incorporated in usual BN structure learning algorithm [de Campos and Castellano, 2007]. Under some condition of consistency, these existence restrictions shall be fulfilled, in the sense that they are assumed to be true for the CBN representing the domain knowledge, and therefore all potential Partially Directed Acyclic Graph (PDAG) must necessarily satisfy them.

Definition 1 *Given a domain ontology $\mathcal{O}$, let G=(C, R_c) be the DAG where R_c: $C_i \times C_j$ represents the subset of semantic causal relations extracted from $\mathcal{O}$. This subset included both direct and logically derivable semantic causal relations. Let H=(X, E_h) be a PDAG, where X is the set of the corresponding random variables and E_h corresponds to the causal dependencies between them. H is consistent with the existence restrictions in G if and only if:*

$\forall\, C_i, C_j \in C$, *if* $C_i \rightarrow C_j \in R_c$ *then* $X_i \rightarrow X_j \in E_h$.

When we are specifying the set of existence restrictions to be used, it is necessary to make sure that these restrictions can indeed be satisfied. In fact, such causal integration may lead to possible conflicts between the two models. When this occurs, we have to maintain the initial causal information in the PDAG since we are supposed to use perfect observational data. On the other hand, we should ensure the consistency of the existence restrictions in such a way that no directed cycles are created in G.

2) Causal discovery process:

Before delving into the details of our approach, we first review the principal idea of the causal discovery process. When performing an experimentation on X_i, we have to measure all neighboring variables and accordingly to the result direct all edges connecting X_i and Nei(X_i). This edge orientation represents one instantiation among all possible instantiations. It is then possible to continue the edge orientation by using the Meek rules [Meek, 1995] to infer new causal relations.

The aim of this phase is to decide which experiment will be performed and hence also which variables will be altered and measured. For this purpose, the strategy we propose in our approach makes use of a semantic distance calculus (e.g. Rada distance [Rada *et al.*, 1989]) provided by the ontology structure. So, for each node in the graph, SemCaDo calculates its semantic inertia, denoted by $SemIn(X_i)$ and expressed as follows:

$$SemIn(X_i) = \frac{\sum_{X_j \in Nei(X_i) \cup X_i} dist_{Rada}(mscs(Nei(X_i) \cup X_i), X_j)}{Card(Nei(X_i) \cup X_i)} \quad (1)$$

where:

- $mscs(c_i, c_j)$: the most specific common subsumer of the two concepts c_i and c_j, where $i \neq j$.
- $dist_{Rada}(c_i, c_j)$: the shortest path between c_i and c_j, where $i \neq j$,
- $Card(M)$: the cardinality of any set.

Moreover, the semantic inertia presents the following properties:

- When the experimented variable and all its neighbors lie at the same level in the concept hierarchy, the semantic inertia will be equal to the number of hierarchical levels needed to reach the mscs.

- It essentially depends on semantic distance between the studied concepts. This means that the more this distance is important, the more the SemIn will be maximized.

Further to these, we also integrate a semantic cumulus relative to inferred edges [Meek, 1995] denoted by $Inferred\_Gain$ in our utility function. For this purpose, we use I(X_i) to denote the set of nodes attached by inferred edges after performing an experimentation on X_i. So, the $Inferred\_Gain$ formula is expressed as follows:

$$Inferred\_Gain(X_i) = \frac{\sum_{X_j \in I(X_i)} dist_{Rada}(mscs(I(X_i)), X_j)}{Card(I(X_i))} \quad (2)$$

Note that we don't use here all the information provided by the ontology. We should also consider the axioms to check if any new relation could be inferred from the semantic point of view. Better interacting with the axioms is one of our perspectives for future work.

When using the two proposed terms weighted by the cost of experiment (i.e. $cost(A_{X_i})$) and measurement (i.e. $cost(M_{X_i})$), our utility function will be as follows:

$$U(X_i) = \frac{SemIn(X_i) + Inferred\_Gain(X_i)}{\alpha cost(A_{X_i}) + \beta cost(M_{X_i})} \quad (3)$$

where measures of importance α, $\beta \in [0,1]$.

This utility function will be of great importance to highlight the serendipitous character of SemCaDo algorithm by guiding the causal discovery process to investigate unexplored areas and conduct more informative experiments.

3) Edge orientation & ontology evolution:

Once the specified intervention performed, the meek rules [Meek, 1995] will be applied to infer new arcs until no more edges can be oriented. We note that these orientation rules are proven to be correct and complete subject to any additional background knowledge.

Since certain experimentation can not be performed, either because of ethical reasons or simply because it is impossible to do it, the final causal graph can be either a CBN or a partially causal graph. In both cases, the causal knowledge will be extracted and interpreted for an eventual ontology evolution.

In this way, the causal relations will be translated as semantic causal relations between the corresponding ontology concepts. For this purpose, SemCaDo algorithm uses a six-phases evolution process [Stojanovic *et al.*, 2002]:

- Change capturing: The aim of this initial step in the ontology evolution process is to capture the new discovered causal relations on the current causal graph which are not actually modeled. It starts after finishing the structure learning step in order to treat all changes in a consistent and unified manner.
- Change representation: In order to be correctly implemented, we have to represent these causal changes formally, explicitly and in a suitable format. In the context of SemCaDo algorithm, we only handle elementary changes (i.e. restricted to adding semantic causal relations) that cannot be decomposed into simpler ones.
- Semantics of change: The semantics of change is the phase that enables the resolution of ontology changes in a systematic manner by ensuring the consistency of the ontology. In our case, conflicting knowledge is highly possible to occur when deducing causal conclusions from the ontology axioms. Such inconsistencies should be handled by automated reasoning. This step also prevents the creation of new cycles in the ontology when integrating the causal discoveries. This consistency rule is maintained since the causal discovery step in SemCaDo avoid the creation of cycles during the structure learning.
- Implementation: In order to avoid performing unwanted changes, a list of all consequences in the ontology and dependent artifacts should be generated and presented to the ontology engineer, who should then be able to accept the change or reject it. If the implementer agrees to add the new causal relationships, all actions to apply the change have to be performed.
- Propagation: Pursuing and adopting the new causal discoveries can generate additional changes in the other parts of the ontology. These changes are called derived changes. That is why, during this step, it is necessary to determine the direct and indirect types of changes to be applied. In case of ambiguity, the ontology expert decide on the action to occur. A human intervention at this level is essential to remove the ambiguity and to make the final decision.
- Validation: Change validation enables justification of performed changes and undoing them at user's request. If the output of SemCaDo causal discovery step is a partially directed graph, it is possible to restart the cycle when there sufficient budget to make further discoveries.

4 Experimental study

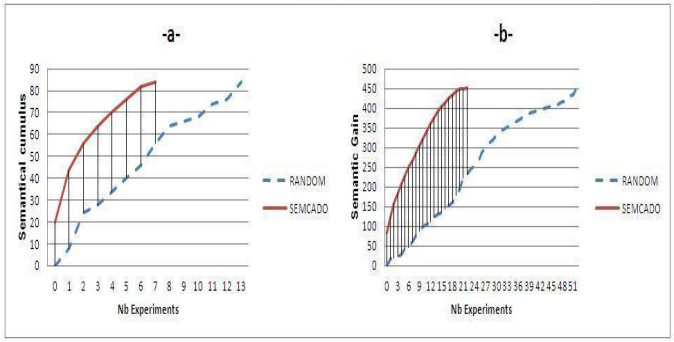

Figure 2: The semantic gain given the number of experiments using SemCaDo and random approach on relatively small graphs (a) and bigger ones (b).

In the experimental evaluation, we will compare our approach to a random causal discovery strategy.

For this purpose, we randomly create a set of syntectic 50 and 200 node graphs and simulate the result of a structure

learning algorithm working with a perfect infinite dataset. For each simulated graph, we automatically generate a corresponding concept hierarchy in which we integrate a varying percentage (10% to 40%) of the initial causal relations. As we do not dispose of a real system to intervene upon, we decide to simulate the experimentations directly in the previously generated CBNs.

Another point to consider in our experimental study concerns the calculation of the semantic gain. In fact, after each SemCaDo (resp. Random) iteration, we measure the sum of semantic distances ([Rada *et al.*, 1989] in these experiments) relative to the new directed edges in the graph and update a semantic cumulus. In both strategies, the two corresponding curves are increasing in, meaning that the higher is the number of experimented variables, the higher is the value of the semantic gain. Nevertheless the more the curve is increasing faster, the more the approach is converging to the best and most impressive experiments.

Figure 2 shows that, during the experimentation process, SemCaDo comfortably outperforms the random approach in term of semantic gain. This is essentially due to the initial causal knowledge integration and the causal discovery strategy when performing the experimentations. But if the two curves reach the same maxima when obtaining a fully directed graph, where is the evolutionary contribution of SemCaDo? Let us remember that we are approaching a decision problem which is subject to the experimentation costs and the budget allocation. Taking into account this constraint, the semantic domination of SemCaDo will be extremely beneficial when the number of experiments is limited.

All these experimental results show how the SemCaDo algorithm can adopt a serendipitous attitude with the minimum expected cost and effort. This is indeed a new avenue of causal investigation, moving far away from traditional techniques.

5 Conclusions and Future Works

In this paper, we outlined our serendipitous and cyclic approach which aims to: i) Integrate the causal prior knowledge contained in the domain ontology when learning the structure of the partially directed graph from observational data. ii) Use the semantic distance calculus to guide the iterative causal discovery process and investigate unexpected causal relations. iii) Capture the required causal discoveries to be applied to ontology evolution.

The SemCaDo algorithm is an initial attempt towards a more ambitious framework exploiting the power of CBNs and ontologies. Future works will detail how to revolutionize the ontologies by adding new axioms and ignoring others when incorporating causal discoveries.

References

[Ben Messaoud *et al.*, 2009] Montassar Ben Messaoud, Philippe Leray, and Nahla Ben Amor. Integrating ontological knowledge for iterative causal discovery and visualization. In *ECSQARU*, pages 168–179, 2009.

[de Campos and Castellano, 2007] Luis M. de Campos and Javier G. Castellano. Bayesian network learning algorithms using structural restrictions. *International Journal of Approximate Reasoning*, pages 233–254, 2007. Eighth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 2005).

[Flouris *et al.*, 2008] Giorgos Flouris, Dimitris Manakanatas, Haridimos Kondylakis, Dimitris Plexousakis, and Grigoris Antoniou. Ontology change: classification and survey. In *Knowledge Engineering Review*, volume 23, pages 117–152, 2008.

[Gruber, 1995] T. R. Gruber. Towards Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal Human-Computer Studies Vol. 43, Issues 5-6*, pages 907–928, 1995.

[Khattak *et al.*, 2009] Asad M. Khattak, Khalid Latif, Songyoung Lee, and Young-Koo Lee. Ontology Evolution: A Survey and Future Challenges. In Dominik Ślzak, Tai-hoon Kim, Jianhua Ma, Wai-Chi Fang, Frode E. Sandnes, Byeong-Ho Kang, and Bongen Gu, editors, *U-and E-Service, Science and Technology*, volume 62, pages 68–75. Springer Berlin Heidelberg, 2009.

[Meek, 1995] Christopher Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the Proceedings of the Eleventh Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 403–41, San Francisco, CA, 1995. Morgan Kaufmann.

[Meganck *et al.*, 2006] Stijn Meganck, Philippe Leray, and Bernard Manderick. Learning causal bayesian networks from observations and experiments: A decision theoretic approach. In Vicen Torra, Yasuo Narukawa, Ada Valls, and Josep Domingo-Ferrer, editors, *MDAI*, volume 3885 of *Lecture Notes in Computer Science*, pages 58–69. Springer, 2006.

[Pearl, 1988] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[Pearl, 2000] Judea Pearl. *Causality: models, reasoning, and inference*. 2000.

[Rada *et al.*, 1989] Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19(1):17–30, 1989.

[Stojanovic *et al.*, 2002] Ljiljana Stojanovic, Alexander Maedche, Boris Motik, and Nenad Stojanovic. User-driven ontology evolution management. In *EKAW '02: Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management*, pages 285–300, London, UK, 2002.

[Xuan *et al.*, 2006] Dung Nguyen Xuan, Ladjel Bellatreche, and Guy Pierra. A versioning management model for ontology-based data warehouses. In A. Min Tjoa and Juan Trujillo, editors, *DaWaK*, volume 4081 of *Lecture Notes in Computer Science*, pages 195–206. Springer, 2006.