Inférence causale et sciences de données pour l'energie et l'industrie du futur

Porteur : Angela Bonifati (UCBL)
Partenaires : (si applicable)
Laboratoire : Liris
Composante : Informatique

Nature du financement demandé : Professeur Invité

Période : (de la venue du Professeur Invité)

Décembre 2025

Résumé :
Le raisonnement causal est un domaine émergent à l'intersection de la science des données, de l'inférence causale et de l'ingénierie des systèmes d'information. Il vise à répondre à des questions causales, en particulier en identifiant les variables et les liens entre causes et effets dans les données réelles.
La causalité se concentre sur l'identification et l'organisation des variables causales, souvent représentées sous forme de graphes acycliques dirigés (DAG), afin de modéliser les relations de cause à effet entre les attributs. Cette approche permet de surmonter les biais de confusion et de renforcer la validité des inférences causales.
Les travaux récents introduisent le problème de l'inférence causale dans la gestion et la science de données en stipulant un seul modèle causal fourni par l'utilisateur. L'enjeux consiste à analyser des données complexes sous la forme de graphes de propriétés. Elle constitue une avancée significative vers une compréhension plus profonde des mécanismes sous-jacents aux phénomènes étudiés.
Avec l'invitation de M.me Sudeepa Roy, professeur à Duke University aux USA et experte d'inférence causale et gestion de données dans des domaines pluridisciplinaires, nous intendons introduire ces thématiques dans les cours du Master International DISS (Data and Intelligence for Smart Systems) à UCBL.

Sujet développé :

La causalité constitue une abstraction fondamentale pour décrire la genèse des phénomènes réels. Le raisonnement causal, moteur de l'intelligence humaine, gagne en importance dans l'Intelligence Artificielle, car il éclaire les mécanismes de prise de décision. Les relations cause-effet et probabilités conditionnelles sont les pierres angulaires des modèles causaux structurels (SCM), qui représentent le processus générateur de données de façon concise à l'aide de graphes causaux [1]. Ces graphes permettent d'estimer des probabilités d'intervention (P(y|do(x))) à partir de données purement observationnelles, en s'appuyant sur le do-calcul de Judea Pearl (récipiendiare du Prix Turing 2011 pour ses travaux sur la théorie de la causalité) pour transformer ces probabilités en expressions exploitables en evitant de runner des experiméntations et simulations couteuses.

Parallèlement, les systèmes de gestion de données NoSQL exploitent massivement les graphes [2], notamment les property graphs : des multigraphes dirigés munis de propriétés (paires clé-valeur) pour sommets et arêtes . Toutefois, malgré leur expressivité, ces modèles ne supportent actuellement ni les relations causales ni les outils pour l'analyse causale profonde. En effet, les DAGs, qui représentent les relations causales via des arêtes dirigées sans cycles, modélisent explicitement les distributions conditionnelles et les dépendances causales . Or ces structures sont aujourd'hui crées par des experts, souvent hors des systèmes de bases de données.

L'inférence causale vise à aller au-delà de la corrélation pour identifier les causes et prédire les effets d'interventions (par ex. : « Une campagne anti-tabac réduirait-elle l'incidence du SARS ? ») ou raisonner sur des scénarios contrefactuels (« Qu'aurait-il été si le prix des cigarettes n'avait pas augmenté ? »). Ces questions demeurent hors de portée des requêtes en graphe classiques.

Pour pallier ce déficit, il est crucial d'intégrer les DAGs dans les property graphs [6], afin de les faire devenir des artefacts pleinement gérés : associables aux données observationnelles, nettoyés, versionnés, et utilisables dans les requêtes. Cela implique la formalisation de vues causales (GAV, LAV ou GLAV) reliant chaque variable causalement pertinente à un nœud ou attribut du graphe, la définition d'opérations causales (par exemple, suppression d'arête) et l'extension des langages de requête (openCypher, GQL, SQL/PGQ) avec des primitives dédiées à la causalité et au do-calcul.

Actuellement, les modèles probabilistes sur graphes (graphiques incertains, bases relationnelles probabilistes) ne traitent que les corrélations, sans distinguer causalité . Il devient donc urgent de proposer un modèle unifié, un graphe de propriété causal, hébergeant à la fois probabilités conditionnelles et relations causales, compatible avec les infrastructures existantes, et évolutif via un langage de requête enrichi.

Ce chantier s'inscrit dans un mouvement global : intégrer la causalité à la donnée pour créer des systèmes intelligents, capables d'engendrer des analyses personnalisées et de guider des décisions scientifiques de manière transparente, interprétable et rigoureuse [3, 4, 5]. Ces systèmes deviendront incontournables dans les domaines de l'energie et industrie du futur. Ils sont une des activités de recherche centrale du Liris.

C'est la raison pour laquelle, nous souhaitons inviter pour 1 mois environ un des spécialistes de renom aux USA dans le domaine de la causalité et science de données : Sudeepa Roy. Elle est professeur à Université de Duke aux USA et membre du laboratoire Duke DB Group. Elle a publié plusieurs articles sur la causalité, science de données et IA [3,4,5].

Elle interviendra dans une ou deux conférences générales pour les étudiants du Master Internaitonal M2 DISS (Data and Intelligence for Smart Systems) à UCBL sur les modèles causales pour la science de données, pour la réalisation d'enseignements auprès des étudiants de M2 dans les modules dédiés en collaboration avec les responsables de ces UE et également dans la participation dans l'encadrement de stages de ce Master. Par ailleurs, pendant son séjour sur Lyon, nous souhaitons également le faire intervenir dans un de nos séminaires du Liris sur des sujets très en pointe sur les structures composites. Enfin, sa venue sur Lyon permettra des réunions de travail quotidiennes sur un sujet Recherche en cours de développement entre nos 2 laboratoires ce qui boostera son avancée. La rédaction de publications est également visée.

Bibliographie

[1] J. Pearl. Causality: Models, Reasoning, and Inference. 2000.

[2] A. Bonifati, G. H. L. Fletcher, H. Voigt, and N. Yakovets. Querying Graphs. Synthesis Lectures on Data Management.Morgan & Claypool Publishers, 2018.

[3] Brit Youngmann, Michael J. Cafarella, Amir Gilad, Sudeepa Roy:
Summarized Causal Explanations For Aggregate Views. Proc. ACM Manag. Data 2(1): 71:1-71:27 (2024)

[4] Babak Salimi, Harsh Parikh, Moe Kayali, Lise Getoor, Sudeepa Roy, Dan Suciu:
Causal Relational Learning. SIGMOD Conference 2020: 241-256

[5] Travis Seale-Carlisle, Saksham Jain, Courtney Lee, Caroline Levenson, Swathi Ramprasad, Brandon Garrett, Sudeepa Roy, Cynthia Rudin, Alexander Volfovsky:
Evaluating Pre-trial Programs Using Interpretable Machine Learning Matching Algorithms for Causal Inference. AAAI 2024: 22331-22340

[6] Amedeo Pachera, Mattia Palmiotto, Angela Bonifati and Andrea Mauri: What If: Causal Analysis with Graph Databases. Proceedings of the VLDB Endow. Vol. 18, 2025. *(to appear)*