Découverte de connaissances stratégiques interprétables dans le Web des données

Mots-clés Web sémantique, Linked Open Data (LOD), Wikidata, graphes de connaissances, ontologies, SPARQL, exploration de données, extraction de connaissances, analyse de données

Contexte et motivation

Le Web des données (Berners-Lee et al., 2001 ; Hogan et al., 2021) est déjà largement utilisé pour relier des jeux de données et en exploiter le sens, pour des tâches en recherche d'information à travers les agents personnels ou les infoboîtes. Ses graphes de connaissances devraient également soutenir les travailleurs du savoir tels que les journalistes, les analystes (entreprises et médias) ou les spécialistes en sciences humaines et sociales (Suchanek et Preda, 2014 ; Weikum, 2021). Ces utilisateurs avancés requièrent des outils d'analyse qui vont au-delà de la recherche d'entités ou de la recherche de leurs propriétés, par exemple ils souhaitent souvent comparer, agréger et classer les entités en fonction de quantités (Giacometti et al., 2021). La construction de ces indicateurs socio-économiques visant à quantifier des phénomènes et leur impact au sein d'un domaine s'impose comme un objectif majeur pour la découverte de connaissances stratégiques. Au sein du Web des données, la diversité des entités et de leurs liens rend complexe le développement de méthodes automatisées reflétant les spécificités de chaque discipline et dont les résultats sont pertinents au sein de chaque discipline. Par exemple, les modèles proposés en recherche d'information ignorent l'hétérogénéité des graphes de connaissances et le plus souvent, ils traitent tous les liens et toutes les entités indifféremment (Finin et al., 2005). Le même indicateur de classement peut comparer une entreprise avec un scientifique au prix de valeurs dénuées de sens aussi bien en économétrie qu'en scientométrie. Cette absence d'interprétabilité s'est accentuée ces dernières années avec le recours aux méthodes d'apprentissage opaques comme le plongement de graphes de connaissances (Wang et al., 2017). La tension entre la transdisciplinarité désirée des modèles pour analyser ces graphes de connaissances et l'interprétabilité disciplinaire attendue des indicateurs socio-économiques issus de ces modèles constitue un verrou scientifique.

Objectif de la thèse

L'objectif de ce travail de thèse est de proposer de nouveaux modèles transdisciplinaires et de les mettre en œuvre dans les graphes de connaissances du Web afin d'y découvrir automatiquement des indicateurs interprétables au sein de leur domaine socio-économique. Pour cela, nous envisageons de nous appuyer sur les travaux menés dans les différents champs de la science de la mesure comme l'infométrie (Egghe, 2005), mais aussi en Online Analytical Processing pour la construction d'indicateurs clés de performance (Lenz et Shoshani, 1997). Il sera aussi nécessaire d'exploiter les ontologies qui constituent un outil propre aux graphes de connaissances (Baader et al., 2004 ; Hogan et al., 2021). D'une part, des raisonnements automatiques sur les ontologies des graphes peuvent permettre de découvrir des indicateurs pertinents, et d'autre part une ontologie explicitant le sens de ces indicateurs pourra être un support à leur interprétation.

Les graphes de connaissances du Web sont nombreux, volumineux et distribués. Leur exploration automatique à la recherche d'indicateurs est d'autant plus ardue que les points d'accès publics à ces données sont bridés par des politiques d'usage juste empêchant l'exécution des requêtes d'analyse qui sont par nature coûteuses (Soulet et Suchanek, 2019). Ces contraintes soulèvent l'enjeu de

parcours parcimonieux en nombre de requêtes pour pouvoir générer des indicateurs à grande échelle (Giacometti et al., 2019 ; Giacometti et al., 2021).

Ce travail se situe donc à la croisée de l'extraction de connaissances et du Web sémantique. Les principales contributions attendues sont :

- Proposition de modèles transdisciplinaires et sensibles aux ontologies pour l'analyse de données
- Proposition d'algorithmes de génération d'indicateurs socio-économiques
- Développement d'un prototype en ligne pour mettre à disposition les résultats en suivant les principes FAIR

Contexte et encadrement

L'étudiant(e) sera accueilli(e) au sein de l'équipe BDTLN du Laboratoire d'Informatique Fondamentale et Appliquée de Tours (LIFAT) de l'Université de Tours. Les travaux seront dirigés par Arnaud Soulet (MdC HDR) et Béatrice Markhoff (MdC HDR).

Références

Baader, F., Horrocks, I., & Sattler, U. (2004). Description logics. In Handbook on ontologies (pp. 3-28). Springer, Berlin, Heidelberg.

Berners-Lee, T., J. Hendler, et O. Lassila (2001). The semantic web. Scientific american 284(5), 34–43.

Ding, L., T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. Doshi, et J. Sachs (2004).

Egghe, L. (2005). Power laws in the information production process : Lotkaian informetrics. Emerald.

Finin, T., Mayfield, J., Joshi, A., Cost, R. S., & Fink, C. (2005). Information retrieval and the semantic web. In Proc. of the 38th annual Hawaii international conference on system sciences (pp. 113a-113a).

Giacometti, A., Markhoff, B., & Soulet, A. (2019). Mining significant maximum cardinalities in knowledge bases. In International Semantic Web Conference (pp. 182-199). Springer, Cham.

Giacometti, A., Markhoff, B., & Soulet, A. (2021). Comparison Table Generation from Knowledge Bases. In European Semantic Web Conference (pp. 179-194). Springer, Cham.

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G. D., Gutierrez, C., ... & Zimmermann, A. (2021). Knowledge graphs. ACM Computing Surveys (CSUR), 54(4), 1-37.

Lenz, H.-J. & A. Shoshani (1997). Summarizability in OLAP and statistical data bases. In Proc. 9th International Conference on Scientific and Statistical Database Management, pp. 132–143. IEEE.

Soulet, A., & Suchanek, F. M. (2019). Anytime large-scale analytics of linked open data. In International Semantic Web Conference (pp. 576-592). Springer, Cham.

Suchanek, F. M., & Preda, N. (2014). Semantic culturomics. Proceedings of the VLDB Endowment, 7(12), 1215-1218.

Wang, Q., Mao, Z., Wang, B., & Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering, 29(12), 2724-2743.

Weikum, G. (2021). Knowledge Graphs 2021: A Data Odyssey. Proceedings of the VLDB Endowment, 14(12), 3233-3238.