Sujet de thèse 2022-2025 présélectionné par l'institut SCAI

GraphologIA : Actionner les méga-graphes de données dans les dataflows d'apprentissage automatique

Encadrants : Hubert.Naacke@sorbonne-universite.fr
Camelia.Constantin@sorbonne-universite.fr

Date limite pour candidater : 26 avril 2022

Contexte et Motivation

L'omniprésence des méga-graphes. Dans de nombreux domaines applicatifs les objets étudiés sont structurés en graphe car ce modèle apporte de la flexibilité et de l'extensibilité dans la représentation des données et leur gestion: graphe de connaissances, d'interaction entre molécules, réseaux sociaux, relations entre lieux et événements. D'après Gartner, ces graphes seront au cœur de 80% des analyses de données dès 2025 [1]. Dans un contexte big data, la capacité d'acquisition de données ne cesse de croître et entraîne une croissance des besoins pour manipuler des graphes toujours plus grands, complexes et variables dans le temps. Ces derniers sont qualifiés de **méga-graphes de données (big data graphs)**.

En science des données, l'apprentissage automatique subit actuellement une mutation importante avec la création de nouveaux modèles dédiés aux graphes, tels que les GNN et leurs extensions [8]. Ces modèles permettent de transformer des **graphes statistiques** (vecteurs et matrices d'adjacences) vers une représentation qui s'avère plus efficace en termes de qualité pour des tâches générales de classification et de prédiction de liens. Toutefois, mettre en action ces modèles sur des graphes statistiques extraits à partir de méga-graphes de données réels est d'autant plus difficile que les données à transformer sont volumineuses, variées et dynamiques.

Nous constatons que les méga-graphes de données réelles ne sont pas directement injectables dans la plupart des modèles d'apprentissage pour les raisons suivantes :

- Hétérogénéité des graphes : les nœuds représentent des concepts variés (personne, lieu, objet d'étude, événement), possèdent des propriétés dont la sémantique n'est pas alignée sur un référentiel commun et des valeurs de types divers (textuel, numérique). Les liens entre nœuds sont eux-mêmes très hétérogènes.

- Volume de données : la taille des méga-graphes de données n'est pas uniquement définie par le nombre de nœuds et d'arcs qu'ils contiennent (matrices d'adjacence). Il faut prendre en compte des informations plus riches sur les nœuds et les arcs qui peuvent être volumineuses. L'analyse de ces graphes nécessite le développement d'algorithmes distribués qui exploitent au mieux les infrastructures big data pour passer à l'échelle.

- Dynamicité des graphes : Dans de nombreux cas d'usage, les graphes représentent de l'information captée en continu (comme par exemple les actions des utilisateurs dans les réseaux sociaux, les flux provenant de capteurs et autres appareils d'observation). Il s'agit de tenir compte de la modification des propriétés d'un nœud ou d'un lien, et des modifications structurelles du graphe lors de l'ajout ou la suppression de nœuds ou de liens.

- Évolution des modèles d'apprentissage automatique : le graphe attendu à l'entrée d'un modèle d'apprentissage automatique n'a pas une définition figée. Les modèles d'apprentissage les plus récents prennent en compte des graphes statistiques décrits non seulement par leur matrice d'adjacence mais aussi par des propriétés plus complexes telles que le score de page rank ou le degré de centralité des nœuds.

Objectif scientifique

L'**objectif** de cette thèse est de concevoir un framework efficace pour construire et analyser de manière automatique des méga-graphes à partir de données hétérogènes et dynamiques. Cela permettra d'exécuter plus efficacement à la fois la préparation des données d'apprentissage et l'entraînement des modèles d'apprentissage.

Ce framework devra permettre d'accomplir les tâches suivantes dans l'analyse et la préparation des données pour les tâches d'apprentissage :

- Compréhension fine des données initiales et de leur dynamique : caractériser les propriétés des graphes initiaux ainsi que la façon dont ces propriétés évoluent dans le temps. Une propriété pourra être définie comme une agrégation d'informations issues de plusieurs nœuds du graphe, sur une plage de temps donnée. Par exemple, dans un réseau social, une propriété peut être le nombre de connexions (i.e. de "followers") qu'un utilisateur possède, et la dynamique de cette propriété peut être décrite par la moyenne hebdomadaire du nombre de connexions.

- Unification et alignement de données : déterminer les liens logiques non explicités entre les différentes données réelles. Identifier les référentiels permettant d'aligner les concepts utilisés dans différents graphes. L'unification de données pourra être formulée en combinant des requêtes de jointure et d'agrégation sur les données initiales avec des tâches d'analyse plus complexes. Par exemple, aligner les mots-clés d'une documentation technique avec un référentiel métier.

- Exécution incrémentale efficace du processus d'alignement du graphe : concevoir des algorithmes performants pour exécuter les opérations d'alignement et compléter le graphe obtenu à partir des données nouvellement arrivées. Les calculs nécessaires à l'alignement sont généralement complexes et nécessitent d'agréger des informations issues de sous-graphes déterminés par des fermetures transitives, ce qui s'avère très coûteux à calculer. Les résultats de ces calculs devront être mis à jour de manière incrémentale en fonction de l'arrivée des nouvelles données. L'objectif est ici de proposer des solutions qui s'appliquent à des graphes contenant des centaines de milliards d'arcs, ainsi que des nœuds avec un degré très déséquilibré.

Justification de l'approche scientifique

Les résultats de cette thèse apporteront plus d'agilité dans le cycle dit d'*ingénierie IA* et comportant les quatre étapes suivantes : préparation des données, définition du modèle, entrainement, validation. La clé de l'approche est un langage déclaratif de haut niveau pour définir chaque étape du cycle. Ainsi, le cycle devient un objet manipulable pouvant être optimisé et exécuté. Lorsque de nouvelles données sont disponibles, il sera possible de re-exécuter automatiquement et incrémentalement les étapes du cycle pour, *in fine*, améliorer la qualité de la tâche d'apprentissage visée.

- Définition d'un langage déclaratif de haut niveau pour décrire de manière logique et déclarative le processus qui transforme les données initiales vers un graphe. Ce langage sera suffisamment expressif pour supporter les opérations d'unification et d'alignement de données décrites ci-dessus. Proposer une méthode pour traduire le processus décrit dans ce langage de haut niveau en un ensemble de tâches qui coopèrent pour préparer les données et produire un méga-graphe.

- Conception de nouvelles méthodes d'indexation pour accéder de manière aléatoire à diverses zones du graphe, tout en minimisant la latence, quelle que soit la taille du graphe. Ces index permettront de naviguer rapidement à travers de multiples portions du graphe afin de calculer les indicateurs temporels capturant la dynamique de graphe.

- Validation expérimentale du bénéfice du framework. Comparer des chaînes de préparation de méga-graphe implémentées avec les méthodes de l'état de l'art [9], avec leur équivalent conçu à l'aide du framework et mesurer le bénéfice en termes de performance. Évaluer le temps gagné pour déployer un dataflow de préparation de méga-graphes.

Les impacts sociétaux de cette thèse sont multiples du fait de la variété des grands graphes existants. Les retombées peuvent concerner par exemple la détection d'attaques sur Internet (motifs dans les réseaux de communication) ou de bulles de filtrage dans les réseaux sociaux [4].

Adéquation à l'Institut

Cette thèse s'inscrit dans le domaine des sciences de données et propose des nouvelles solutions essentielles pour le déploiement de dataflows de machine learning. Certains résultats pourraient être diffusés sous la forme d'ateliers organisés par SCAI ou ses partenaires (Datacraft), ce qui permettrait, entre autres, de valider l'applicabilité de la méthode à travers différents cas d'usage.

Rôle et compétences scientifiques des encadrants

Les méthodes envisagées pour concevoir ce framework reposent sur l'expertise des encadrants en optimisation de requêtes big data et en analyse de grands graphes. En particulier, les travaux sur les métriques d'agrégation temporelle dans les réseaux sociaux [2] et ceux sur les motifs de voisinage [3] serviront de socle pour définir les métriques capturant la dynamqiue temporelle des propriétés du graphe. Les travaux sur l'analyse et la visualisation de réseaux complexes [6] [7], menés dans le cadre du projet ANR EPIQUE [5], seront mis à profit pour définir le langage déclaratif de dataflow d'IA et son exécution efficace avec des très larges jeux de données.

Profil de l'étudiant recherché

Nous recherchons un ou une candidate motivé/e avec des bonnes compétences en bases de données (SQL, indexation), algorithmique et programmation (Python, Java). Des connaissances en optimisation de requêtes, en algorithmique sur les graphes et en apprentissage automatique sont un plus.

Publications en lien avec le projet

[1] Gartner. Top 10 Data and Analytics Technology Trends for 2021, Trend 8: Graph Relates Everything https://www.gartner.com/en/newsroom/press-releases/2021-03-16-gartner-identifies-top-10-data-and-analytics-technologies-trends-for-2021

[2] J. Debure, S. Brunessaux, C. Constantin, C. du Mouza, A pattern-based Approach for an Early Detection of Popular Twitter Accounts, International Database Engineering and Applications Symposium (IDEAS), 2020.

[3] Q. Grossetti, C. Constantin, C. du Mouza, N. Travers, An Homophily-based Approach for Fast Post Recommendation on Twitter, EDBT 2018

[4] Uthsav Chitra and Christopher Musco. 2020. Analyzing the Impact of Filter Bubbles on Social Network Polarization. In Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM '20). Association for Computing Machinery, New York, NY, USA, 115–123.

[5] EPIQUE Projet ANR http://www-bd.lip6.fr/wiki/site/recherche/projets/epique/start

[6] Li KE, Bernd Amann, Hubert Naacke, Exploring the Evolution of Science with Pivot Topic Graphs: 3rd International Workshop on Big Data Visual Exploration and Analytics EDBT/ICDT 2020

[7] Ke Li, Hubert Naacke et Bernd Amann. « An Analytic Graph Data Model and Query Language for Exploring the Evolution of Science». In :Big Data Research 26 (2021), 18 pages.

[8] William L. Hamilton: Graph Representation Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan & Claypool Publishers 2020

[9] E. Hussein, A. Ghanem et al: Graph Data Mining with Arabesque. SIGMOD 2017. 1647-1650