Ladjel BELLATRECHE
École Nationale Supérieure de Mécanique et d'Aérotechnique
LIAS – UR 20299
1 avenue Clément Ader BP 40109
86961 Chasseneuil Cedex – France
Tél : +33 (0)5 49 49 80 72
bellatreche@ensma.fr

Sujet de Stage Master 2/Ingénieur

Titre de Stage : *Etude des l'efficacité énergétique des triple Stores*
Responsables scientifiques :
Ladjel BELLATRECHE, LIAS, ISAE-ENSMA, Poitiers, bellatreche@ensma.fr
Gayo DIALLO, BPH INSERM 1219, Université de Bordeaux, gayo.diallo@u-bordeaux.fr

Ce projet est financé par par le réseau R3 TESNA (Réseau Régional de Recherche : Transition Énergétique (`https://www.r3-tesna.com/`)

Lieu de Stage : ISAE-ENSMA, Poitiers.

1 Contexte Général

Le développement du Web Sémantique, initié par Tim Berners-Lee, a joué un rôle fondamental dans l'essor des graphes de connaissances (GdC). Un GdC est une collection structurée et sémantiquement organisée, lisible par des machines, d'entités interconnectées (classes, relations, instances) exprimant des faits simples sous le format (sujet, prédicat, objet), appelés triplets. Ces derniers peuvent représenter des informations d'intérêt universel ou spécifique à un domaine. Les GdC peuvent également inclure des dimensions spatiales et temporelles, des attributs, des règles de bon sens, ainsi que des contextes enrichis d'entités et de faits sous forme de données textuelles, visuelles, de descripteurs et de statistiques.

Parmi les GdC les plus connus, on trouve, Google Knowledge Graph, Amazon Product Knowledge Graph, DBpedia, et YAGO. Ces graphes ont largement contribué à l'amélioration de nombreux systèmes, notamment les systèmes de recommandation (ex. Netflix), les systèmes de question-réponse, et plus récemment les Large Language Models.

Les GdC contiennent un volume massif de données, sous forme de millions, voire de milliards de triples RDF. Par exemple, DBpedia contient plus de 5 milliards de triples, tandis que le Google Knowledge Graph est estimé à des dizaines de milliards d'entités et des centaines de milliards de relations.

Pour stocker, interroger et gérer ces GdC, des triple stores ont été développés. Ces stores se divisent en deux grandes catégories : (1) les triple stores non natifs, construits sur des systèmes de gestion de bases de données relationnelles, et (2) les triple stores natifs, développés spécifiquement pour prendre en charge les caractéristiques des triplets RDF. Notons que l'exploration de ces GdC se fait à l'aide du langage de requêtes SPARQL, conçu pour permettre des recherches complexes et sémantiques sur des ensembles de données vastes et hétérogènes. Initialement, les processeurs de requêtes SPARQL ont été optimisés pour réduire les temps de réponse. Cependant, dans une ère où la sobriété énergétique devient une préoccupation centrale, l'étude de leur **efficacité énergétique** apparaît comme une question cruciale et urgente. Développer un moteur SPARQL capable d'offrir un compromis entre la performance des requêtes et leur consommation énergétique devient un enjeu crucial pour la communauté de recherche.

Ladjel BELLATRECHE
École Nationale Supérieure de Mécanique et d'Aérotechnique
LIAS – UR 20299
1 avenue Clément Ader BP 40109
86961 Chasseneuil Cedex – France
Tél : +33 (0)5 49 49 80 72
bellatreche@ensma.fr

2 Objectif de Stage

L'objectif de ce stage est de transposer notre expérience sur l'étude de l'efficacité énergétique des systèmes de stockage relationnels dans le contexte des triple stores. Pour mener à bien cette étude, nous réaliserons une analyse empirique visant à évaluer la consommation énergétique des triple stores (libres et commerciaux). Cette évaluation nous permettra d'identifier les composants et opérations les plus énergivores. Une fois ces éléments identifiés, nous proposerons des modèles basés sur l'apprentissage automatique et l'apprentissage profond pour estimer l'énergie consommée. Ces modèles seront ensuite validés et déployés sur des triple stores libres tels que Virtuoso et Jena.

3 Bibliographie

1. Simon Pierre Dembele : Auditer l'énergie. Avant de déployer ses modèles : Vers des optimiseurs verts de requêtes analytiques. (Auditing Energy - Before Deploying Models : Towards Green Optimizers of Analytical Queries). École Nationale Supérieure de Mécanique et d'Aérotechique, Poitiers, France, 2021
2. Simon Pierre Dembele, Ladjel Bellatreche, Carlos Ordonez, Amine Roukh : Think big, start small : a good initiative to design green query optimizers. Clust. Comput. 23(3) : 2323-2345 (2020)
3. Ladjel Bellatreche, Fouad Djellali, Wojciech Macyna, Carlos Ordonez : Energy-Aware Query Processing : A Case Study on Join Reordering. IEEE Big Data 2023 : 3743-3752
4. Amine Roukh, Ladjel Bellatreche, Selma Bouarar, Ahcène Boukorca : Eco-Physic : Eco-Physical design initiative for very large databases. Inf. Syst. 68 : 44-63 (2017)
5. Jorge Galicia : Revisiting Data Partitioning for Scalable RDF Graph Processing. (Revisiter le partitionnement des données pour le traitement scalable des graphes RDF). École Nationale Supérieure de Mécanique et d'Aérotechique, Poitiers, France, 2021
6. Julien Aimonier-Davat, Hala Skaf-Molli, Pascal Molli, Minh Hoang Dang, Brice Nédelec : Join Ordering of SPARQL Property Path Queries. ESWC 2023 : 38-54

4 Profil du candidat

La candidate ou le candidat devra être inscrit(e) en Master 2 ou en dernière année d'école d'ingénieur et posséder des connaissances en bases de données/connaissances, en programmation, et en apprentissage automatique. Des notions en optimisation de requêtes relationnelles sont un atout pour ce stage.
Un bon niveau en français et en anglais est requis.

5 Documents à fournir

- Curriculum Vitae ;
- Lettre de motivation ;

Ladjel BELLATRECHE
École Nationale Supérieure de Mécanique et d'Aérotechnique
LIAS – UR 20299
1 avenue Clément Ader BP 40109
86961 Chasseneuil Cedex – France
Tél : +33 (0)5 49 49 80 72
bellatreche@ensma.fr

- Notes de tout le parcours universitaire
- Tout autre document jugé nécessaire par le candidat pouvant enrichir le dossier de candidature (ex. lettres de recommandation)