COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Nantes Université

THÈSE DE DOCTORAT

NANTES UNIVERSITÉ

ÉCOLE DOCTORALE N° 641
Mathématiques et Sciences et Technologies du numérique de l'Information et de la Communication
Spécialité : *Informatique*

Par
Minh-Hoang DANG

Querying the Web as a Knowledge Graph

Thèse présentée et soutenue à Nantes, le 28/11/2024
Unité de recherche : Laboratoire des sciences du numérique à Nantes (LS2N)

Rapporteurs avant soutenance :

Olivier CURÉ	Maître de conférence, Université Gustave Eiffel
Pierre-Antoine CHAMPIN	Maître de conférence, Université Claude Bernard Lyon 1

Composition du Jury :

Président :	TBD	TBD
Examinateurs :	Fatiha SAIS	Professeure des Universités, Univesité Paris Saclay
	Mounira HARZALLAH	Maître de Conférences (HDR), Nantes Université
Dir. de thèse :	Pascal MOLLI	Professeur des Universités, Nantes Université
Co-dir. de thèse :	Hala SKAF-MOLLI	Professeure des Universités, Nantes Université

COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Nantes Université

Titre : Requêter le Web en tant que Graphes de Connaissances

Mot clés : Web sémantique, annotation sémantique, Grands Modèles de Langue (LLMs), moteurs de requêtes fédérées

Résumé : Le Web sémantique vise à améliorer le Web en y ajoutant des données structurées, permettant ainsi des capacités de recherche plus avancées Malgré une croissance significative, d'importantes lacunes subsistent dans la couverture des informations du quotidien, et l'interrogation de grandes fédérations de points d'accès SPARQL présente des défis de performance. Cette thèse aborde deux questions clés : (1) Les modèles de Grands Modèles de Langue (LLM) peuvent-ils être fiables pour générer des annotations Schema.org précises ? (2) Comment optimiser les requêtes au sein d'une fédération croissante de points d'accès SPARQL ? Pour répondre à ces questions, cet ouvrage propose deux contributions majeures. D'abord, nous proposons LLM4Schema.org comme un outil pour valider les annotations générées par les LLMs. Ensuite, FedShop évalue l'évolutivité des fédérations SPARQL, tandis que FedUP améliore les performances des requêtes grâce à un algorithme tenant compte des résultats. Ensemble, ces innovations améliorent la couverture des données et l'efficacité des requêtes sur le Web sémantique.

Title: Querying the Web as Knowledge Graphs

Keywords: Semantic Web, Schema.org annotations, Large Language Models (LLMs), SPARQL federation engines

Abstract:

The Semantic Web aims to enhance the Web with structured data, enabling more advanced search capabilities. Despite significant growth, major gaps remain in everyday information coverage, and querying large federations of SPARQL endpoints presents performance challenges. This thesis addresses two key questions: (1) Can Large Language Models (LLMs) be trusted to generate accurate Schema.org annotations? (2) How can efficient querying be achieved across an expanding federation of SPARQL endpoints? To tackle these issues, this work introduces two contributions. First, we propose LLM4Schema.org as a tool validate machine-generated annotations. Second, FedShop benchmarks SPARQL federation scalability, and FedUP improves query performance with a result-aware algorithm. Together, these innovations enhance data coverage and query efficiency across the Semantic Web.