CED : « Sciences et Techniques de l'Ingénieur »

AVIS DE SOUTENANCE

«AHMED RABHI»

Présentera ses travaux de recherche en vue de l'obtention du Doctorat en Sciences et Techniques

Intitulé de la thèse :

« Traitement optimisé des requêtes sur des sources distribuées et hétérogènes du web des données »

Date :	**Lundi 24 juillet 2023**
Heure :	**10 heures**
Lieu :	**« Salle de Conférences, ENSA de Tanger »**

Devant le jury :

Membres de jury

Pr. SBIHI Abderrahmane	FS de Kénitra	Président
Pr. Ladjel BELLATRECHE	LIAS -ENSMA Poitiers-France	Rapporteur
Pr. Abdelmounaime LACHKAR	ENSA de Tanger	Rapporteur
Pr. Aziz MABROUK	FFS de Tétouan	Rapporteur
Pr. Djamal BENSLIMANE	Université Claude Bernard Lyon 1 -France	Examinateur
Pr. Rachida FISSOUNE	ENSA de Tanger	Co-encadrante
Pr. Hassan BADIR	ENSA de Tanger	Directeur de Thèse

Université Abdelmalek Essaadi
Ecole Nationale des Sciences Appliquées
Tanger

RESUME

Le travail présenté dans cette thèse se focalise sur l'agrégation d'information à partir de sources de données distribuées et hétérogènes. En effet, il se peut qu'une information ne soit pas trouvée en interrogeant une seule source de données ; la réponse à la requête nécessite la collecte de ses fragments à partir de plusieurs sources de données. L'objectif général de ce travail est de concevoir et mettre en place un système de recherche agrégée capable de répondre à une requête de l'utilisateur en collectant les pièces de données nécessaires depuis des sources distribuées et indépendantes dans le web des données. On peut voir le web des données comme un espace hébergeant des datasets RDF accessibles via des services web appelés « endpoints SPARQL » supportant l'exécution des requêtes SPARQL. Notre système permet d'exécuter une requête SPARQL sur des sources de données distribuées et hétérogènes en décomposant la requête en sous-requêtes afin de chercher les données nécessaires pour chaque sous-requête, puis, retourne la réponse finale à l'utilisateur comme si l'exécution a eu lieux sur un seul large dataset.

Les problématiques et les objectifs traitées dans cette thèse se résument dans les trois points suivants : (i) Optimisation du temps d'exécution sans perte de données, (ii) traite- ment des parties manquantes de la requête, enfin, (iii) Adopter des métriques pour évaluer les sources de données et suivre les changements pouvant survenir aux sources de données.

(i) Optimisation du temps d'exécution : Plusieurs facteurs peuvent affecter le temps d'exécution dans les moteurs de recherche agrégée. Dans cette thèse, les facteurs sur lesquels on se focalise sont : le nombre de sources interrogées, le nombre de sous- requêtes exécutées et la taille des données transférées. Pour cela, nous proposons une stratégie qui vise à optimiser le temps d'exécution sans perte de données. Cette stratégie commence par filtrer les sous-requêtes candidates dans le but d'optimiser la taille des données transférées ainsi que le nombre de sous-requêtes exécutées, puis, le système effectue un plan d'exécutions basé sur un index local visant à sélectionner les sources pertinentes, ce plan d'exécution enrichit l'exécution parallélisée des sous-requêtes qui se fait sur une architecture de traitement parallèle basé sur le multithreading.

(ii) Traitement des parties manquantes de la requête : Il est possible que certaines parties de la requête soient introuvables. Pour cela, nous proposons un algorithme de réécriture de requêtes qui prend en considération les parties trouvées de la requête de l'utilisateur et la raffine pour retourner une réponse sémantiquement signifiante.

(iii) Mise en place d'une base de connaissance à jour sur les endpoints SPARQL : Les sources de données avec lesquelles notre système interagit sont gérés par des organisations externes. Ainsi, on propose une base de connaissance à jour pour suivre les changements pouvant survenir sur la qualité des endpoints SPARQL avec le temps. Cette base de connaissance permet d'évaluer les endpoints en se basant sur des métriques de performance, de rendement et de disponibilité.

Les résultats expérimentaux montrent que la stratégie proposée pour l'optimisation de temps d'exécution permet de minimiser le temps de traitement sans perte de données. En effet, la solution de filtrage de sous-requêtes candidates est efficace pour éviter les données dupliquées et pour raffiner la taille des données intermédiaires. Selon les résultats obtenus, notre solution de sélection de sources et de planification d'exécutions optimise le temps de d'exécution, puis, l'évaluation expérimentale confirme l'efficacité de l'architecture parallélisée puisqu'elle réduit le temps de traitement de manière considérable. Nous avons expérimenté l'algorithme proposé pour la réécriture de requêtes sous différentes conditions, et les résultats obtenus confirment son efficacité pour retourner la plus grande partie possible de l'information recherchée.