جامعة محمد الخامس بالرباط
Université Mohammed V de Rabat

École Nationale Supérieure d'Informatique et d'Analyse des Systèmes
Centre d'Études Doctorales en Sciences des Technologies de l'Information et de l'Ingénieur

AVIS DE SOUTENANCE DE THÈSE DE DOCTORAT

Madame Ouafa AJARROUD

Soutiendra publiquement sa thèse de Doctorat en Informatique

Le Samedi 20 janvier 2024 à 10h00 au Grand Amphi à l'ENSIAS

Intitulé de la thèse

A New Semantic Coverage-based Approach For Efficient Cache Management In Mediation Systems

Président :

Pr. Laila CHEIKHI, PES, ENSIAS, Université Mohammed V de Rabat

Directeur de thèse :

Pr. Ahmed ZELLOU, PES, ENSIAS, Université Mohammed V de Rabat

Rapporteurs :

Pr. El Habib BEN LAHMAR, PES, Faculté des Sciences Ben M'Sick, Université Hassan II, Casablanca

Pr. Soumia ZITI, PES, Faculté des Sciences, Université Mohammed V de Rabat

Pr. Imane HILAL, PH, École des Sciences de l'Information de Rabat

Examinateur :

Pr. Mohamed RADOUANE, PH, ENSIAS, Université Mohammed V de Rabat

Abstract: Information mediation has emerged as a solution to address the problems of data heterogeneity and source diversity. In fact, the exponential increase in data volume, together with the diverse and autonomous nature of different data sources, has made integration extremely difficult. A mediation system provides users with transparent and unique access to information from distributed and heterogeneous sources. Like many information systems, a mediator can also include a cache to improve its operational efficiency and performance. This cache mechanism consists of a set of previously queried data regions, which minimizes the frequency of interrogating remote sources. Therefore, the use of a cache during mediation could reduce response times. In addition, it provides a smaller data set to be considered as a backup in case one or more sources become unavailable.

It is true that caching can be useful to address performance problems in information mediation. However, the administration of a cache, especially if several regions are involved, introduces complexities in terms of configuration and maintenance. Furthermore, it can lead to additional latency in data retrieval, due to the extensive application of the trimming process. Complex formulas with many disjoint segments may be sent to sources. Therefore, efficient cache management and efficient query processing are key factors for optimal results.

This thesis presents three contributions aimed at optimizing cache performance in a mediator. Our first contribution is an approach based on computing query coverage rates in cached regions. It aims to prioritize specific regions and filter out others during the rewriting process to ensure optimal performance. It is based on thresholds we have defined, taking into account the state of remote sources. We then proposed the ontology-based "OntoCoverage" approach as a second contribution. It enables us to calculate semantic coverages and make better use of the cache, while maintaining efficient response times. Although effective, this approach requires the creation and management of ontologies, which can be a challenge for some mediation systems. In this context, we presented our third contribution, "WEQP". It allows the semantic processing of cache formulas without the need for regularly updated ontologies. WEQP represents the cache as a set of hypercubes based on word embedding techniques, providing results comparable to OntoCoverage but with broader applicability across different mediation systems.

Keywords: Ontologies, Query Coverage, Semantic Cache, Virtual Integration.

Résumé: L'augmentation exponentielle du volume de données, ainsi que la nature autonome des sources, ont rendu l'intégration extrêmement difficile. La médiation d'information a été proposée à cet égard. Un système de médiation fournit aux utilisateurs un accès transparent et unique à des informations provenant de sources distribuées et hétérogènes. Comme de nombreux systèmes, un médiateur peut également inclure un cache afin d'améliorer son efficacité et ses performances. Un cache consiste en un ensemble de régions de données précédemment sollicitées

minimisant la fréquence d'interrogation des sources distantes. L'utilisation d'un cache pourrait donc réduire les temps de réponse. En outre, il fournit un dépôt de données à utiliser si des sources sont indisponibles.
Certes, la mise en cache est utile pour résoudre les problèmes de performance d'un médiateur. Toutefois, l'administration d'un cache, en particulier si plusieurs régions sont concernées, introduit des complexités. En effet, elle peut entraîner des temps de latence supplémentaires, en raison de l'application extensive du processus de découpage. Des formules complexes comportant de nombreux segments disjoints peuvent aussi être envoyées aux sources. Par conséquent, une gestion et un traitement des requêtes efficaces sont des facteurs clés pour obtenir des résultats optimaux.

Cette thèse présente trois contributions visant à optimiser les performances du cache dans un médiateur. Notre première contribution est basée sur le calcul des taux de couverture des requêtes dans les régions mises en cache. Elle vise à prioriser des régions spécifiques et à filtrer les autres lors du processus de réécriture. Elle se base sur des seuils que nous avons définis, en tenant compte de l'état des sources distantes. Nous avons ensuite proposé "OntoCoverage" comme deuxième contribution. En se basant sur les ontologies, elle permet de calculer les couvertures sémantiques et de mieux utiliser le cache, tout en maintenant des temps de réponse réduits. Bien qu'efficace, cette approche nécessite la gestion d'ontologies, ce qui peut constituer un défi pour certains médiateurs. Dans ce contexte, nous avons présenté notre troisième contribution, "WEQP". Elle permet le traitement sémantique des formules du cache sans nécessiter d'ontologies régulièrement mises à jour. WEQP représente le cache comme un ensemble d'hypercubes basés sur les techniques word embedding, fournissant des résultats comparables à OntoCoverage mais avec une applicabilité plus large à travers différents systèmes de médiation.

Mots clés: Cache sémantique, Couverture, Intégration virtuelle, Ontologies.