Résumé

Le sujet de cette thèse s'inscrit dans le cadre général de la Recherche d'Information et la gestion des données distribuées. Notre problématique concerne l'évaluation et l'optimisation de requêtes agrégatives (*Aggregated Search*). La Recherche d'Information Agrégative (RIA) est un nouveau paradigme permettant l'accès à l'information massivement distribuée. Elle a pour but de retourner à l'utilisateur d'un système de recherche d'information des objets résultats qui sont riches et porteurs de connaissances. Ces objets n'existent pas en tant que tels dans les sources. Ils sont construits par assemblage (ou configuration ou agrégation) de fragments issus de diffèrentes sources. Les sources peuvent être non spécifiées dans l'expression de la requête mais découvertes dynamiquement lors de la recherche. Nous nous intéressons particulièrement à l'exploitation des dépendances de données pour optimiser les accès aux sources distribuées. Dans ce cadre, nous proposons une approche pour l'un des sous processus de systèmes de RIA, principalement le processus d'indexation/organisation des documents. Nous considérons dans cette thèse, les systèmes de recherche d'information orientés graphes (graphes RDF). Utilisant les relations dans les graphes, notre travail s'inscrit dans le cadre de la recherche d'information agrégative relationnelle (*Relational Aggregated Search*) où les relations sont exploitées pour agréger des fragments d'information. Nous proposons d'optimiser l'accès aux sources d'information dans un système de recherche d'information agrégative. Ces sources contiennent des fragments d'information répondant partiellement à la requête. L'objectif est de minimiser le nombre de sources interrogées pour chaque fragment de la requête, ainsi que de maximiser les opérations d'agrégations de fragments dans une même source. Nous proposons d'effectuer cela en réorganisant la/les base(s) de graphes dans plusieurs clusters d'information dédiés aux requêtes agrégatives. Ces clusters sont obtenus à partir d'une approche de clustering sémantique ou structurel des prédicats des graphes RDF. Pour le clustering structurel, nous utilisons les algorithmes d'extraction de sous-graphes fréquents et dans ce cadre nous élaborons une étude comparative des performances de ces algorithmes. Pour le clustering sémantique, nous utilisons les métadonnées descriptives des prédicats dont nous appliquons des outils de similarité textuelle sémantique. Nous définissons une approche de décomposition de requêtes basée essentiellement sur le clustering choisi.

Mots clés

Recherche d'information agrégative, Partitionnement sémantique, Fouille de graphes, Extraction de sous-graphes fréquents, Décomposition de requêtes, Recherche distribuée

Abstract

In this research, we are interested in investigating issues related to query evaluation and optimization in the framework of aggregated search. Aggregated search is a new paradigm to access massively distributed information. It aims to produce answers to queries by combining fragments of information from different sources. The queries search for objects (documents) that do not exist as such in the targeted sources, but are built from fragments extracted from the different sources. The sources might not be specified in the query expression, they are dynamically discovered at runtime. In our work, we consider data dependencies to propose a framework for optimizing query evaluation over distributed graph-oriented data sources. For this purpose, we propose an approach for the document indexing/orgranizing process of aggregated search systems. We consider information retrieval systems that are graph oriented (RDF graphs). Using graph relationships, our work is within relational aggregated search where relationships are used to aggregate fragments of information. Our goal is to optimize the access to source of information in a aggregated search system. These sources contain fragments of information that are relevant partially for the query. We aim at minimizing the number of sources to ask, also at maximizing the aggregation operations within a same source. For this, we propose to reorganize the graph database(s) in clusters, dedicated to aggregated queries. We use a semantic or strucutral clustering of RDF predicates. For structural clustering, we propose to use frequent subgraph mining algorithms, we performed for this a comparative study of their performances. For semantic clustering, we use the descriptive metadata of RDF predicates and apply semantic textual similarity methods to calculate their relatedness. Following the clustering, we define query decomposing rules based on the semantic/structural aspects of RDF predicates.

Keywords

Relational aggregated search, Semantic partitioning, Clustering, Graph mining, Frequent subgraph mining, Query decomposition, Distributed search