EXPLORATION ET INTERROGATION DE DONNÉES RDF INTÉGRANT DE LA CONNAISSANCE MÉTIER» PAR HANANE OUKSILI

Présentée par : Hanane Ouksii Discipline : informatique Laboratoire : DAVID

Résumé :
Un nombre croissant de sources de données est publié sur le Web, décrites dans les langages proposés par le W3C tels que RDF, RDF(S) et OWL. Une quantité de données sans précédent est ainsi disponible pour les utilisateurs et les applications, mais l'exploitation pertinente de ces sources constitue encore un défi : l'interrogation des sources est en effet limitée d'abord car elle suppose la maîtrise d'un langage de requêtes tel que SPARQL, mais surtout car elle suppose une certaine connaissance de la source de données qui permet de cibler les ressources et les propriétés pertinentes pour les besoins spécifiques des applications. Le travail présenté ici s'intéresse à l'exploration de sources de données RDF, et ce selon deux axes complémentaires : découvrir d'une part les thèmes sur lesquels porte la source de données, fournir d'autre part un support pour l'interrogation d'une source sans l'utilisation de langage de requêtes, mais au moyen de mots clés. L'approche d'exploration proposée se compose ainsi de deux stratégies complémentaires : l'exploration thématique et la recherche par mots clés. La découverte

de thèmes dans une source de données RDF consiste à identifier un ensemble de sous-graphes, non nécessairement disjoints, chacun représentant un ensemble cohérent de ressources sémantiquement liées et définissant un thème selon le point de vue de l'utilisateur. Ces thèmes peuvent être utilisés pour permettre une exploration thématique de la source, où les utilisateurs pourront cibler les thèmes pertinents pour leurs besoins et limiter l'exploration aux seules ressources composant les thèmes sélectionnés. La recherche par mots clés est une façon simple et intuitive d'interroger les sources de données. Dans le cas des sources de données RDF, cette recherche pose un certain nombre de problèmes, comme l'indexation des éléments du graphe, l'identification des fragments du graphe pertinents pour une requête spécifique, l'agrégation de ces fragments pour former un résultat, et le classement des résultats obtenus. Nous abordons dans cette thèse ces différents problèmes, et nous proposons une approche qui permet, en réponse à une requête mots clés, de construire une liste de sous-graphes et de les classer, chaque sous-graphe correspondant à un résultat pertinent pour la requête. Pour chacune des deux stratégies d'exploration d'une source RDF, nous nous sommes intéressés à prendre en compte de la connaissance externe, permettant de mieux répondre aux besoins des utilisateurs. Cette connaissance externe peut représenter des connaissances du domaine, qui permettent de préciser le besoin exprimé dans le cas d'une requête, ou de prendre en compte des connaissances permettant d'affiner la définition des thèmes. Dans notre travail, nous nous sommes intéressés à formaliser cette connaissance externe et nous avons pour cela introduit la notion de pattern. Ces patterns représentent des équivalences de propriétés et de chemins dans le graphe représentant la source. Ils sont évalués et intégrés dans le processus d'exploration pour améliorer la qualité des résultats.

Abstract :

An increasing number of datasets is published on the Web, expressed in languages proposed by the W3C to describe Web data such as RDF, RDF(S) and OWL. The Web has become a unprecedented source of information available for users and applications, but the meaningful usage of this information source is still a challenge. Querying these data sources requires the knowledge of a formal query language such as SPARQL, but it mainly suffers from the lack of knowledge about the source itself, which is required in order to target the resources and properties relevant for the specific needs of the application.

The work described in this thesis addresses the exploration of RDF data sources. This exploration is done according to two complementary ways: discovering the themes or topics representing the content of the data source, and providing a support for an alternative way of querying the data sources by using keywords instead of a query

formulated in SPARQL. The proposed exploration approach combines two complementary strategies: thematic-based exploration and keyword search. Theme discovery from an RDF dataset consists in identifying a set of sub-graphs which are not necessarily disjoints, and such that each one represents a set of semantically related resources representing a theme according to the point of view of the user. These themes can be used to enable a thematic exploration of the data source where users can target the relevant theme and limit their exploration to the resources composing this theme. Keyword search is a simple and intuitive way of querying data sources. In the case of RDF datasets, this search raises several problems, such as indexing graph elements, identifying the relevant graph fragments for a specific query, aggregating these relevant fragments to build the query results, and the ranking of these results. In our work, we address these different problems and we propose an approach which takes as input a keyword query and provides a list of sub-graphs, each one representing a candidate result for the query. These sub-graphs are ordered according to their relevance to the query. For both keyword search and theme identification in RDF data sources, we have taken into account some external knowledge in order to capture the users needs, or to bridge the gap between the concepts invoked in a query and the ones of the data source. This external knowledge could be domain knowledge allowing to refine the user's need expressed by a query, or to refine the definition of themes. In our work, we have proposed a formalization to this external knowledge and we have introduced the notion of pattern to this end. These patterns represent equivalences between properties and paths in the dataset. They are evaluated and integrated in the exploration process to improve the quality of the result.

INFORMATIONS COMPLÉMENTAIRES

M. Mokrane BOUZEGHOUB, Professeur des Universités, Université de Versailles Saint-Quentin-en-Yvelines - Laboratoire DAVID - Directeur de these
Mme Isabelle COMYN-WATTIAU, Professeur des Universités, CNAM - Rapporteur
M. Omar BOUCELMA, Professeur des Universités, Université d'Aix-Marseille - Laboratoire LSIS - Rapporteur
Mme Yolaine BOURDA, Professeur des Universités, Centrale Supélec - Examinateur
Mme Daniela GRIGORI, Professeur des Universités, Université Paris Dauphine - Examinateur
Mme Zoubida KEDAD, Maître de conférences, Université de Versailles Saint-Quentin-en-Yvelines - Laboratoire DAVID - Co-encadrant de these

Contact : dredval service FED : theses@uvsq.fr