Proposition de recherche doctorale

Recherche d'Objets Complexes dans le Web Structuré

Mots clés :

- **Directeur de thèse** : TALEL ABDESSALEM
- **Co-encadrant(s)** :
- **Unité de recherche** : Laboratoire Traitement et Communication de l'Information
- **Ecole doctorale** : École Doctorale Informatique, Télécommunications, Électronique de Paris
- **Domaine scientifique principal**: Divers

Résumé du projet de recherche (Langue 1)

Nous assistons aujourd'hui à un développement continu et rapide du {Web Structuré}, dans lequel les documents ne sont plus composés que de texte non structuré mais sont centrés sur les données, présentant des contenus structurés et des {objets complexes}. Les plates-formes de recherche d'information (IR) actuelles s'inspirent des procédures et des techniques utilisées dans les systèmes de recherche documentaire. Le passage à des contenus structurés, avec des schémas prédéfinis, nécessite des techniques d'interrogation plus précises et plus riches, et soulève de nouveaux défis auxquels nous essayons de fournir des réponses. En effet, la recherche par {mots-clés} n'est pas adaptée pour interroger le Web structuré. De nouveaux moyens de recherche sur le Web sont donc nécessaires, pour permettre à l'utilisateur de cibler des données complexes avec une sémantique précise. Nous étudions dans cette thèse les défis théoriques et pratiques qui sont soulevés dans l'interrogation des objets complexes. Nous proposons {ObjectRunner}, un système pour l'extraction et l'interrogation de données complexes du Web structuré, qui exploite la redondance du Web et la régularité des structures des pages pour mieux déterminer les données à extraire. Notre système fournit un résultat le plus complet possible à des requêtes plus riches que celles par simples mots-clés. Nous proposons une approche d'interrogation en deux étapes, qui permet à l'utilisateur de décrire de façon souple et précise le schéma des objets recherchés. Ensuite, pour chaque source Web (ensemble de pages {HTML}), le schéma cible et la structure des pages sont analysés pour : (1) sélectionner les sources répondant à la requête de l'utilisateur, (2) retrouver les données pouvant répondre à la requête de l'utilisateur, et (3) les extraire. La solution proposée par notre système est générique, dans le sens où elle n'est pas spécifique à un domaine d'application ou un type d'objets en particulier.