"Extraction, Exploitation and Evaluation of Document-based Knowledge."

Mémoire d'Habilitation à Diriger des Recherches (HDR) d'Antoine Doucet, soutenu le 30 avril 2012 à l'Université de Caen Basse-Normandie, laboratoire GREYC (CNRS UMR 6072).

Mots-clés : fouille de données séquentielles, unités multi-mots, recherche d'information, évaluation des systèmes d'information, méthodes multilingues, passage à l'échelle.

Jury :

Mme Isabelle Tellier (Pr, Paris 3 Sorbonne Nouvelle, présidente)
M. Massih-Reza Amini (MC-HDR, Université Pierre et Marie Curie, Paris 6, rapporteur)
M. Pavel Brazdil (Pr, Université de Porto, Portugal, rapporteur)
M. Manuel Vilares Ferro (Pr, Université de Vigo, Espagne, rapporteur)
M. Bruno Crémilleux (Pr, Université de Caen Basse-Normandie, examinateur)
Mme Mounia Lalmas (Pr, Yahoo! Research Barcelone, examinatrice)
M. Gaël Dias (Pr, Université de Caen Basse-Normandie, directeur)

Résumé :

Les travaux présentés dans ce mémoire gravitent autour du document numérique : Extraction de connaissances, utilisation de connaissances et évaluation des connaissances extraites, d'un point de vue théorique aussi bien qu'expérimental.

Le fil directeur de mes travaux de recherche est la généricité des méthodes produites, avec une attention particulière apportée à la question du passage à l'échelle. Ceci implique que les algorithmes, principalement appliqués au texte dans ce mémoire, fonctionnent en réalité pour tout type de donnée séquentielle.

Sur le matériau textuel, la généricité et la robustesse algorithmique des méthodes permettent d'obtenir des approches endogènes, fonctionnant pour toute langue, pour tout genre et pour tout type de document (et de collection de documents). Le matériau expérimental couvre ainsi des langues utilisant différents alphabets, et des langues appartenant à différentes familles linguistiques. Les traitements peuvent d'ailleurs être appliqués de la même manière au grain phrase, mot, ou même caractère.
Les collections traitées vont des dépêches d'agence de presse aux ouvrages numérisés, en passant par les articles scientifiques.

Ce mémoire présente mes travaux en fonction des différentes étapes du *pipeline* de traitement des documents, de leur appréhension à l'évaluation applicative. Le document est ainsi organisé en trois parties décrivant des contributions en :

1) extraction de connaissances (fouille de données séquentielle et veille multilingue) ;
2) exploitation des connaissances acquises, par des applications en recherche d'information, classification et détection de synonymes via un algorithme efficace d'alignement de paraphrases ;
3) méthodologie d'évaluation des systèmes d'information dans un contexte de données massives, notamment l'évaluation des performances des systèmes de recherche d'information sur des bibliothèques numérisées.