Sujet de Projet IMPACT 2013-2014

Option : Décision & Analyse de Données

Titre : Similarité structurelle générique entre documents XML

Mot-clés: similarité structurelle, arbres, distance d'édition, XML

Lieu : Université Lille 2 -IUT- Département Statistique et Informatique Décisionnelle
25-27 rue du Maréchal Foch 59100 ROUBAIX

Encadrants : Fatima BELKOUCH / Isabelle BIERMANN
Mail : fatima.belkouch@univ-lille2.fr

Sujet :

Comparer des documents XML est une clé majeure dans de nombreux problèmes de fouilles de données XML (Classification/Clustering, matching, frequent pattern….). La similarité structurelle a été largement étudiée dans la littérature et de nombreuse mesures ont été proposées chacune pour être efficace et pertinente dans l'application pour laquelle elle été destinée et pour un type de collections xml donné.

Le but du projet est d'arriver à identifier les besoins en similarité structurelle entre documents xml spécifiques à chaque application. De traduire cela en paramètres pour les intégrer dans une fonction de mesure de similarité générique, configurable pour chaque type d'application et chaque collection xml.

Le projet consiste à :

- Faire une recherche bibliographique pour étudier la similarité structurelle dans ses différents cadres applicatifs : clustering, data integration, querying, data Warehousing
- Identifier les paramètres importants dans la comparaison de deux documents XML
- Intégrer ces paramètres dans une mesure de similarité modulable selon le poids attribué à chaque paramètre pour chaque application
- Evaluer la mesure de similarité sur des données réelles et/ou synthétiques

Bibliographie

[1] "An overview on XML similarity: background, current trends and future directions" Joe Tekli, Richard Chbeir, Kokou Yetongnon. Computer Science Review, Volume 3, Issue 3, 2009, Pages 151–173, Elsevier

[2] "A Short Survey of Document Structure Similarity Algorithms" D. Buttler. In Proceedings of the 5th International Conference on internet Computing, USA, (2004) pp. 3-9.