Sujet de Projet IMPACT 2013-2014

Option : Décision & Analyse de Données

Titre : Classification de documents XML basée sur la similarité structurelle

Mot-clés: XML, Clustering, structural similarity, tree edit distance, java

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 :

XML s'est imposé comme un métalangage permettant de représenter et d'échanger des données non seulement dans le web mais de façon générale en entreprise. Pour extraire des informations de nombreuses applications reposent sur la recherche par similarité de données. Evaluer la similarité entre documents XML reste un des problèmes cruciaux lors du processus de fouille de données.

Le projet s'inscrit dans le cadre du développement d'une plate-forme de fouille de données XML. Une mesure de similarité a été implémentée basée sur l'algorithme Edit Distance. Une classification des documents xml permettrait d'évaluer la pertinence de cette mesure de similarité.
Le travail consiste à :

- Faire un état de l'art des méthodes de classification de documents XML
- Déterminer la méthode adaptée au besoin
- Implémenter celle-ci en Java
- Créer une interface web pour les tests.

Bibliographie

[1] "XML Data Clustering: An Overview" Alsayed Algergawy, Marco Mesiti, Richi Nayak, Gunter Saake ACM Computing Surveys, Vol. 43, No. 4, Article 25, October 2011.
[2] "Efficient XML Structural Similarity Detection using Sub-tree Commonalities", Joe Tekli, Richard Chbeir, Kokou Yétongnon, The 22nd Brazilian Symposium on Databases (SBBD'07), pp. 116-130, Joao Pessoa, Brazil, ACM SIGMOD DiSC, October 2007