Restad : un logiciel d'indexation et de stockage relationnel de contenus XML

Yoann Moreau Eric SanJuan Patrice Bellot
LIA, 339, chemin des Meinajaries 84911 AVIGNON Cedex 9
{yoann.moreau,eric.sanjuan,patrice.bellot}@univ-avignon.fr

Restad [1] est un outil pour charger de grands nombres de documents XML dans une base de données PostgreSQL [2]. La structure XML ainsi que le contenu, y compris celui des attributs, est stocké sous forme relationnelle. Cela est fait sans aucun présupposé sur les DTDs des fichiers et permet de gérer tous les standards XML. Il est le format le plus courant pour des données semi-structurées (issues d'applications de bureautique, annotées par des analyseurs syntaxiques, enrichies de multiples annotations sémantiques...).

Restad reprend ainsi les fonctionnalités d'importation de TopX (Theobald *et al.*, 2005) et XRel (Yoshikawa *et al.*, 2001) en l'adaptant au SGBDR libre PostgreSQL et surtout, en intégrant la gestion des attributs de balises.
La structure hiérarchique des documents (balises et attributs) est représentée de manière relationnelle dans des tables. Le texte du document est enregistré en tant que bloc de texte nettoyé de toute balise XML. La table des balises conserve les positions de début et de fin de chaque balise, tandis que l'index plein-texte conserve pour chaque mot sa position.

Cette approche permet de stocker la forme et le contenu de documents XML, sans perte d'information. On peut ensuite utiliser les index de la base pour effectuer des recherches plein-texte en réduisant les requêtes à un ou plusieurs sous ensembles de l'arborescence des documents. L'utilisation d'un SGBD offre les performances du langage SQL pour effectuer des requêtes complexes.

La première version de Restad est écrite en Ruby et disponible sous licence GPL. Elle a été testée sur le corpus XML de la campagne INEX 2010 de 52Go comprenant l'ensemble du wikipedia en anglais enrichi de multiples annotations sémantiques (Schenkel *et al.*, 2007).
La base de données finale, après création de tous les index occupe un espace disque inférieur à 5 fois la taille du corpus. L'outil permet de re-générer le contenu XML d'un document très rapidement grâce aux index de la base. Différents tests sont prévus pour évaluer les performances avec des requêtes utilisant les balises XML. L'outil pourrait par la suite être facilement adapté à tout format de document arborescent et à tout autre SGBD.

Remerciements

Ces recherches ont bénéficié du soutien financier de l'Agence Nationale de la Recherche (ANR 2010 CORD 001 02) en faveur du projet CAAS.

Références

SCHENKEL R., SUCHANEK F. M. & KASNECI G. (2007). Yawn : A semantically annotated wikipedia xml corpus. In A. KEMPER, H. SCHÖNING, T. ROSE, M. JARKE, T. SEIDL, C. QUIX & C. BROCHHAUS, Eds., *BTW*, volume 103 of *LNI*, p. 277–291 : GI.

THEOBALD M., SCHENKEL R. & WEIKUM G. (2005). An efficient and versatile query engine for topx search. In K. BÖHM, C. S. JENSEN, L. M. HAAS, M. L. KERSTEN, P.-Å. LARSON & B. C. OOI, Eds., *VLDB*, p. 625–636 : ACM.

YOSHIKAWA M., AMAGASA T., SHIMURA T. & UEMURA S. (2001). XRel : A Path-Based Approach to Storage and Retrieval of XML.

1. Relational Storage for Tagged Documents (https ://github.com/ymoreau/Restad)
2. http ://www.postgresql.org