

Modélisation d'un entrepôt de documents XML

Fatma Abdelhedi*, Landry Ntsama*,**
Gilles Zurfluh*

*IRIT-SIG, Université de Toulouse 1 Capitole
2 rue du Doyen Gabriel Marti, 31042 Toulouse, France
prenom.nom@irit.fr,
<http://www.irit.fr/-Equipe-SIG->
**landry.ntsama@ut-capitole.fr

Résumé. Le format XML est aujourd'hui omniprésent dans les organisations et sur le Web. Il facilite le transport et l'échange de données complexes et hétérogènes représentant une information précieuse très peu, voire pas du tout exploitée. Les technologies OLAP et les systèmes d'entrepôts de données actuels permettent l'analyse et le stockage des données transactionnelles issues des bases données relationnelles. Cependant, ces outils ne sont pas adaptés à l'analyse des documents XML du fait de leur structure hiérarchique, ou de leur contenu étant souvent textuel. Nous proposons dans cet article une approche permettant de construire un entrepôt de document XML « centré document » dont le schéma conceptuel est modélisé en utilisant le formalisme UML ; nous présentons aussi une architecture pour l'intégration physique de ces documents dans un environnement XML natif.

1 Introduction

Le format XML facilite la description et l'échange des données complexes provenant de sources hétérogènes. De plus sa structure balisée et auto-descriptive (Fankhauser et Klement, 2003) permet une représentation aisée de différents types de données (structurés et semi-structurés). C'est aujourd'hui un standard de communication tant au sein des organisations que sur le Web. Les systèmes OLAP et les entrepôts de données permettent respectivement l'analyse multidimensionnelle et le stockage orienté sujet des données transactionnelles. Les données stockées sont décrites par un schéma multidimensionnel (Golfarelli et al., 1998), qui permet à un décideur de réaliser des analyses OLAP sur les cubes de données définis dans l'entrepôt. Les systèmes OLAP permettent ainsi une analyse interactive en représentant les données suivant plusieurs perspectives ou dimensions. L'analyse OLAP sur des données relationnelles (structurées) est un processus aujourd'hui parfaitement maîtrisé. Mais ce processus s'avère inadapté pour des documents XML qui présentent une structure hiérarchique profonde et différent des données plates tant au niveau de leur modèle de données que de leur modèles de requêtes (Bordawekar et Lang, 2005).

Nous proposons une nouvelle approche permettant de construire un entrepôt sur des documents XML « centré document » (Pérez et al., 2008) en les intégrant au niveau physique dans