Restructuration automatique de documents dans les corpus semi-structurés hétérogènes

Guillaume Wisniewski*, Ludovic Denoyer*, Patrick Gallinari*
* Laboratoire d'Informatique de Paris 6
8 rue du Capitaine Scott, 75015 Paris
{guillaume.wisniewski, ludovic.denoyer, patrick.gallinari}@lip6.fr

Résumé. L'interrogation de grandes bases de documents semi-structurés (type XML) est un problème ouvert important. En effet, pour interroger un document dont le schéma est nouveau, un système doit pouvoir soit adapter la requête posée au document, soit adapter le document pour pouvoir lui appliquer la requête. Nous nous positionnons ici dans le cadre de la restructuration de documents qui consiste à transformer des documents semi-structurés issus de diverses sources dans un schéma de médiation connu. Nous proposons un cadre statistique général à la problématique de la restructuration de documents et détaillons une instance d'un modèle stochastique de documents structurés appliquée à cette problématique. Nous détaillons enfin un ensemble d'expériences effectuées sur les documents du corpus INEX afin de mesurer la capacité de notre modèle.

1 Introduction

Le développement du document électronique et du Web a vu émerger puis s'imposer des formats de données semi-structurées, tels le XML et le XHTML. Ces nouveaux formats, décrivant simultanément la structure logique des documents et le contenu de ceux-ci, permettent de représenter l'information sous une forme plus riche que le simple contenu et adaptée à des besoins spécifiques. Étant donné l'augmentation rapide du nombre de documents semi-structurés, il est devenu nécessaire d'adapter les méthodes de traitement de données existantes afin de tenir compte des spécificités de ces nouveaux formats ainsi que d'étudier les nouvelles problématiques que ces formats font émerger.

L'initiative INEX (Fuhr et al 2002) propose d'étudier la problématique de la recherche documentaire sur des documents semi-structurés. L'hétérogénéité des structures des données est rapidement apparue comme un obstacle à la conception de systèmes d'interrogation de données semi-structurées issues de différentes sources d'information. Bien que, dans le cadre d'INEX, cette problématique ait été ignorée jusqu'à présent, l'édition 2004 de la campagne d'évaluation propose une nouvelle tâche, la tâche *hétérogène,* qui y est consacrée. Deux solutions peuvent être imaginées pour résoudre ce problème : les systèmes peuvent soit adapter la requête posée au document, soit adapter le document pour pouvoir lui appliquer la requête. Nous adoptons ici la deuxième solution et proposons d'utiliser un *schéma de médiation* pour exprimer l'ensemble des documents considérés dans une structure commune. L'utilisateur n'interagira alors qu'avec ce schéma de médiation. Cette solution nécessite de pouvoir *restructurer* les documents afin d'adapter leur structure au schéma de médiation.

La problématique de restructuration des données est apparue depuis de nombreuses années dans de nombreux domaines tels les entrepôts de données, l'intégration de données, le web sémantique, ... Plus récemment, plusieurs travaux se sont intéressés à l'application de cette problématique aux données semi-structurées et plus particulièrement aux données