Ré-ordonnancement pour l'apprentissage de transformations de documents HTML

Guillaume Wisniewski, Patrick Gallinari

LIP6 — UMPC
8 rue du capitaine Scott
75015 Paris
{prénom.nom}@lip6.fr,
http://www-connex.lip6.fr

Résumé. Notre objectif est de transformer les documents Web vers un schéma médiateur XML défini a priori. C'est une étape nécessaire pour de nombreuses tâches de recherche d'information concernant le Web Sémantique, les documents semi-structurés, le traitement de sources hétérogènes, etc. Elle permet d'associer une structure sémantiquement riche à des documents dont le formats ne contient que des informations de présentation. Nous proposons de traiter ce problème comme un problème d'apprentissage structuré en le formalisant comme une transformation d'arbre en arbre.

Notre méthode de transformation comporte deux étapes. Dans une première étape, une grammaire hors-contexte probabiliste permet de générer un ensemble de solutions candidates. Dans une deuxième étape, ces solutaions candidates sont ordonnées grâce à un algorithme de ré-ordonnancement à base de perceptron à noyau. Cette étape d'ordonnancement nous permet d'utiliser de manière efficace des caractéristiques complexes définies à partir du document d'entrée et de la solution candidate.

1 Introduction

Le Web 2.0 a pour objectif de faciliter l'accès à l'information en représentant les documents Web par une structure sémantiquement riche et non par un traditionnel « sac de mots ». Cette structure est généralement définie par la représentation des documents sous forme d'arbres : des *éléments de contenu*, identifiés par la séquence étiquetée des feuilles de l'arbre, sont organisés selon une structure prédéfinie par un ensemble de *nœuds internes* représentant les relations entre éléments. Cette structure traduit les relations sémantiques ou logiques entre éléments de contenu. Les comparateurs de prix, le Web Sémantique sont des exemples de services fournis par le Web 2.0.

La plupart des documents du web utilisent des formats semi structurés comme le HTML, le XML, le PDF ou encore le WikiText. Ces formats permettent d'enrichir le texte à l'aide de balises et une interprétation directe de celles-ci permet de décrire les documents par un arbre, l'arbre DOM. Nous appellerons *structure syntaxique* cette structure directement liée à la manière dont l'information est codée. Les applications du Web 2.0 ne peuvent toutefois pas