Extraction de relations dans les documents Web

Rémi Gilleron * , Patrick Marty * , Marc Tommasi * , Fabien Torre*

* Projet Mostrare Inria Futurs & Université de Charles de Gaulle - Lille III
59653 Villeneuve d'Ascq CEDEX FRANCE
prenom.nom@univ-lille3.fr

Résumé. Nous présentons un système pour l'inférence de programmes d'extraction de relations dans les documents Web. Il utilise les vues textuelle et structurelle sur les documents. L'extraction des relations est incrémentale et utilise des méthodes de composition et d'enrichissement. Nous montrons que notre système est capable d'extraire des relations pour les organisations existantes dans les documents Web (listes, tables, tables tournées, tables croisées).

1 Introduction

Le développement d'Internet comme source d'informations a conduit à l'élaboration de programmes nommés *wrappers* pour collecter de l'information sur les sites Web. Ces programmes sont difficiles à concevoir et à maintenir. Deux approches sont alors envisageables : la première consiste à assister l'utilisateur, c'est le cas du système Lixto (Baumgartner et al., 2001) dans lequel on spécifie le wrapper dans un langage logique avec l'aide d'un environnement visuel ; la seconde consiste à générer automatiquement le wrapper en limitant l'intervention de l'utilisateur à l'annotation des informations à extraire sur quelques documents. Cette approche est fondée sur le fait que la plupart des documents sur Internet sont générés par programme et présentent des régularités exploitables par les méthodes d'apprentissage automatique.

Les premiers systèmes d'induction de wrappers n'utilisaient que l'aspect textuel des documents (Hsu et Dung, 1998; Kushmerick, 1997). Avec l'apparition de XML, ces approches textuelles sont devenues insuffisantes. Les systèmes actuels utilisent la structure arborescente des documents du Web (Carme et al., 2005; Cohen et al., 2003; Kosala et al., 2002; Muslea et al., 2003; Thomas, 2003). Nous nous inscrivons dans cette démarche en proposant un système d'induction qui utilise à la fois les vues textuelle et arborescente. Beaucoup de systèmes d'induction de wrappers sont conçus pour des tâches unaires. Un wrapper unaire extrait un ensemble de valeurs, par exemple l'ensemble des noms de produits disponibles sur un site marchand. Un wrapper n-aire extrait les instances d'une relation n-aire, par exemple les couples (nom du produit, prix). Il existe deux approches pour induire un wrapper n-aire : soit combiner n wrappers unaires, soit apprendre directement le wrapper n-aire. La première approche nécessite l'obtention d'un modèle pour la combinaison, ou une intervention de la part de l'utilisateur (Jensen et Cohen, 2001; Muslea et al., 2003), ou encore l'utilisation d'heuristiques. La seconde approche est illustrée par les systèmes WIEN (Kushmerick, 1997) et SOFT MEALY (Hsu et Dung, 1998) utilisant des délimiteurs textuels pour repérer les composantes des tuples et le système LIPX (Thomas, 2003) basé sur la logique du premier ordre.