

Extraction de données sur Internet avec Retroweb

Fabrice Estiévenart*, Jean-Roch Meurisse**

*CETIC asbl, rue Clément Ader 8, 6041 Charleroi (Belgique)
fe@cetic.be,

**FUNDP, Institut d'Informatique, rue Grandgagnage 21, 5000 Namur (Belgique)
jrm@info.fundp.ac.be

Résumé. Ce document décrit *Retroweb*, une boîte à outils qui permet l'extraction de données structurées à partir de pages Web. Notre solution est semi-automatique car les données à extraire sont préalablement définies par l'utilisateur. L'intérêt de cette approche est qu'elle permet l'extraction de données ciblées et conformes aux besoins de l'application utilisatrice (migrateur, moteur de recherche, outil de veille). *Retroweb* se caractérise aussi par une grande facilité d'utilisation car il ne nécessite aucune connaissance de langage particulier, la définition des règles d'extraction se faisant directement de manière interactive dans le navigateur Internet. Ce document décrit les trois principaux processus de notre méthode.

1 Classification des pages

L'objectif de cette phase est d'identifier les principaux types de pages composant le site analysé. Un type de pages est un ensemble de pages relativement similaires tant sur le plan syntaxique (code HTML) que sémantique (concept représenté par la page).

Pour atteindre cet objectif, un taux de similarité est calculé entre les pages du site sur la base d'un ensemble de critères tels que ceux décrits dans Ricca et Tonella (2003).

2 Analyse sémantique des pages

Lors de cette étape, l'utilisateur définit les *composants* qu'il souhaite extraire à partir d'un échantillon représentatif de pages d'un même type. Un composant est un concept présent au sein des pages d'un même type. Il peut être absent de certaines pages et/ou y apparaître plusieurs fois. De plus, on lui associe une indication de format (i.e. texte simple ou balisé) et de localisation. Dans *Retroweb*, cette dernière propriété est exprimée sous la forme d'un chemin (XPath) dans l'arborescence formée par les balises HTML.

La figure 1 illustre le scénario de construction d'une règle d'extraction. (1) L'utilisateur sélectionne une instance du composant à définir et lui assigne un nom représentatif tandis que l'outil calcule son chemin d'accès XPath. (2) La règle est appliquée à chacune des pages de l'échantillon afin d'en vérifier la validité. (3) Si la valeur attendue pour chacune des pages n'a pu être extraite, la règle doit être raffinée. Pour ce faire plusieurs solutions sont proposées :