Introduction à la recherche
Équipe TALN

Méthodes de nettoyage de pages HTML/XHTML

Encadrants :
Fabien Poulard, Doctorant
Nicolas Hernandez, Maître de conférence

Nombre d'étudiants : 4

Contexte : Développement d'outils de collecte de données sur internet qui nécessitent une phase de nettoyage des pages Web afin d'en extraire le contenu pertinent.
Objet : Évaluation des méthodes de nettoyage existante et exploration d'une approche tenant compte de la structure arborescente des pages (DOM HTML).
Environnement informatique : HTML/XHTML/CSS ; Python ou Apache UIMA (Java)

Description :

Le domaine du traitement automatique des langues naturelles fait de plus en plus appel à des ressources extérieures collectées sur le Web. En ce sens, le Web constitue un corpus accessible à faible coût et d'une taille sans précédent (*cf.* *http://ngrams.googlelabs.com/info*). L'exploitation de cette ressource nécessite toutefois de lever un certain nombre de freins technologiques. Certains de ces freins relèvent directement de la linguistique informatique (identification de la langue des documents, de leur genre, de leur thématique...), d'autres sont transversaux à plusieurs domaines (parcours des sites, nettoyage des pages, interprétation de la structure visuelle...).

Le problème de nettoyage des pages HTML/XHTML est l'objet de plusieurs travaux, notamment [Freitag1998], [Cohen2002], [Gupta2003] et [Yang2003]. Une campagne d'évaluation internationale a notamment été lancée pour ce problème (http://cleaneval.sigwac.org.uk/). Certains robots de parcours du Web (*crawlers*) intègrent des outils permettant de filtrer le contenu des pages Web, notamment Web-Harvest (http://web-harvest.sourceforge.net/) ou WebSphinx (http://www-2.cs.cmu.edu/~rcm/websphinx/). Cependant, il n'existe pas à notre connaissance de système aboutit et librement disponible qui puisse être directement évalué pour la tâche d'extraction du contenu depuis des flux HTML.

L'objectif des étudiants sera dans un premier temps d'effectuer un état de l'art des techniques et des outils existants pour l'extraction de contenu depuis des arbres HTML/XTML bien formés. Ils expérimenteront et évalueront dans un second temps ces différentes techniques en implémentant un prototype de nettoyeur.

Bibliographie

Freitag1998: Freitag, D., Information extraction from HTML: Application of a general machine learning, 1998
Cohen2002: Cohen, William W. and Hurst, Matthew and Jensen, Lee S., A flexible learning system for wrapping tables and lists in HTML documents, 2002
Gupta2003: Gupta, S. and Kaiser, G. and Neistadt, D. and Grimm, P., DOM-based content extraction of HTML documents, 2003
Yang2003: Yang, Y. and Chen, Y. and Zhang, HJ, HTML page analysis based on visual cues, 2003