# Babouk – exploration orientée du web pour la constitution de corpus et de terminologies

Clément de Groc[1,2] Javier Couto[1,3] Helena Blancafort[1,4] Claude de Loupy[1]

(1) Syllabs, 15 rue Jean-Baptiste Berlier, 75013 Paris
(2) Univ. Paris Sud et LIMSI-CNRS, F-91405 Orsay
(3) MoDyCo, UMR 7114, CNRS-Université Paris Ouest Nanterre, La Défense
(4) Universitat Pompeu Fabra Roc Boronat,138, 08018 Barcelona, Spain
{cdegroc, jcouto, blancafort, loupy}@syllabs.com

Babouk est un crawler orienté (Chakrabarti et al., 1999) : son objectif est le rapatriement efficace de documents pertinents pour un domaine défini. Comparativement au crawling traditionnel, le crawling orienté permet un accès rapide à des données spécialisées tout en évitant le coût prohibitif d'un parcours en largeur du web. L'exploitation du web comme source de données linguistiques a permis de créer de nombreux corpus généralistes et spécialisés par le biais de requêtes à un moteur de recherche (Baroni, Bernardini, 2004) ou d'un crawl du web (Baroni & Ueyama, 2006). Babouk ne requiert qu'un petit ensemble de termes ou URLs amorces en entrée. Le reste de la procédure est automatique. L'utilisateur peut régler le crawler par un ensemble de paramètres et reprendre la main sur la procédure à tout moment.

Babouk doit trouver un maximum de documents pertinents en téléchargeant le minimum de pages. Le crawler s'appuie sur un catégoriseur qui filtre les documents non pertinents et ordonne par pertinence les pages à télécharger. Le catégoriseur est basé sur un lexique pondéré construit durant la première itération du crawling : une extension de l'entrée utilisateur est effectuée en utilisant la procédure BootCaT (Baroni, Bernardini, 2004). Le lexique est ensuite pondéré à l'aide d'une mesure de « représentativité » s'appuyant sur le web. Une phase de calibration automatique permet de déterminer un seuil pour la catégorisation. Pour guider le crawler en priorité vers les pages les plus pertinentes, le score fourni par le catégoriseur est utilisé de manière analogue au critère OPIC (Abiteboul et al., 2003).

Plusieurs critères d'arrêt on été implémentés tels qu'un nombre maximal de tokens ou de documents à télécharger, une profondeur ou une durée de crawl maximale. Plusieurs filtres sont appliqués dans le but d'améliorer la qualité des corpus constitués. L'utilisateur peut ainsi choisir de ne conserver que des pages d'une certaine taille ou appartenant à un certain format de fichier (parmi Microsoft Office, Adobe PDF, ou HTML). Il peut également limiter le crawl à certains domaines/sites ou, au contraire, les filtrer.

Babouk est basé sur Nutch et distribué sur une grappe de machines (optionnellement sur le « cloud »), ce qui assure un passage à l'échelle en termes de puissance de calcul nécessaire pour la réalisation de nombreux crawls simultanément. Enfin, les documents et méta-informations résultants du crawling peuvent être stockés dans une base de données distribuée assurant, encore une fois, la *scalabilité* du système. Les utilisateurs peuvent configurer et lancer leurs crawls à partir d'une interface web dynamique. Cette dernière offre également un suivi (logs) du crawl en temps réel.

ABITEBOUL M., PREDA M., COBENA G. (2003). Adaptive on-line page importance computation. Actes de *12th international conference on the World Wide Web – WWW*. 280-290.

BARONI M., BERNARDINI S. (2004). BootCaT : Bootstrapping Corpora and Terms from the Web. Actes de *4th international conference on language resources and evaluation – LREC*. 1313-1316.

BARONI M., UEYAMA M. (2006). Building general- and special-purpose corpora by Web crawling. Actes de *13th NIJL International Symposium, Language Corpora: Their Compilation and Application.* 31-40.

CHAKRABARTI S., DEN BERG M.V., DOM B. (1999). Focused crawling : a new approach to topic-specific Web resource discovery. Actes de *Computer Networks, vol. 31*. 1623-1640.