Proposition de TER pour la Maîtrise d'Informatique : recherche Bayesienne sur le Web

Encadrants :

Roberto Di Cosmo Université de Paris 7
`http://www.dicosmo.org`, E-mail : `roberto@dicosmo.org`
Tel : 01 44 27 86 55.

Laboratoires d'accueil : PPS (Université de Paris 7).

Cadre de la recherche : Ces dernières années ont vu un essor spectaculaire des technologies liées au Web, et à l'accroissement exponentiel de l'information disponible en ligne. Nous nous intéressons ici en particulier au problème de la qualité des résultats fournis par la recherche d'information à l'aide de moteurs de recherche tels Google. L'expérience courante montre que, avec le temps, la qualité des résultats qui sont retournés en réponse à une requête donnée posée par un internaute se dégrade inexorablement : les bons résultats sont inévitablement noyés dans une marée de résultats qui ne nous intéressent pas. Ceci est une conséquence inévitable de l'ambiguïté du langage naturel, et de la difficulté, non seulement pour les non experts, de reformuler la requête initiale avec un choix de mots clefs suffisamment précis pour écarter de la masse des résultats ceux qui ne sont pas pertinents.

Ce problème n'est pas nouveau, et on a pu voir plusieurs solutions qui ont été proposées depuis la généralisation de l'usage du Web :

classification humaine comme dans le cas de Yahoo : des centaines d'employés passent leur temps à arpenter e Web et classifient les sites selon des catégories fixées à l'avance

extraction de réseaux sémantiques comme dans le cas de LiveTopics : chaque document se voit associer un réseau de concepts ("topics") et l'utilisateur peut alors choisir d'affiner sa recherche en incluant ou excluant certains des "topics" qu'il a rencontré

exploitation des réseaux endogènes comme dans Google[1] : on considère comme document plus significatif celui vers lequel pointe un maximum d'autres documents repérés par la requête

réseaux d'utilisateurs comme dans le projet Galilei : chaque utilisateur peut créer plusieurs "profils" pour ses recherches (ex : profil "cinéfil d'auteur", profil "acheteur de voiture ancienne", etc.), et on peut fédérer les utilisateurs en mettant en commun les requêtes correspondants à un même profil

Et pourtant, il y a une approche intéressante, et facile à mettre en oeuvre, qui n'a jusque là pas été mise en pratique, du moins à ma connaissance : la recherche

[1] Google ne fait pas que ça, bien sûr.

par filtrage Bayesien. Le filtrage Bayesien a été popularisé par Paul Graham pour éliminer automatiquement un maximum de spam[2] des messages électroniques que nous recevons quotidiennement. Il utilise un théorème élémentaire de théorie de probabilités, connu sous le nom de formule de Bayes, pour extraire des messages bons et mauvais déjà reçus une distribution de probabilités qui a des bonnes propriétés de prédiction sur les nouveaux messages qui arrivent.

Le but de ce TER est de tester l'approche Bayesien dans le cas des recherches Web, en développant une fine couche logicielle qui utilise Google comme moteur de recherche, mais filtre ensuite les pages retrouvées, en utilisant les indications de l'utilisateur pour les séparer en bonnes et mauvaises, de la même façon qu'un filtre anti-spam Bayesien. Cet approche peut être utilement composé avec l'approche des réseaux d'utilisateurs, et la notion de profil de recherche.

La construction du prototype est un travail pratique d'implantation, mais la mise à point du modèle statistique nécessite quelques notions de théorie des probabilités.

[2] Courrier électronique non sollicité.