ANR-Programme MDCO – Projet C-MANTIC

Evelyne Bourion[1], Sylvain Loiseau[2], Egle Ramdani[1], Mathieu Valette[3]

[1]ERTIM, INaLCO, Paris
[2]LIMSI, CNRS, Orsay
[3]ATILF, CNRS, Nancy

Du point de vue de l'utilisateur, le reproche principal fait aux moteurs de recherche classiques porte sur la surabondance d'informations ramenées et leur faible adéquation aux besoins. Ceci tient au fait que les moteurs actuels restent incapables de discriminer les types de documents et donc d'évaluer leur pertinence sémantique. Le besoin d'outils plus efficaces a d'ailleurs convaincu Google de produire son nouveau moteur de recherche « Google Scholar ».

Or, aucun des moteurs existants, y compris Scholar, ne fournit de garantie qualitative sur les documents proposés : sont-ils bien des textes scientifiques ? Si non, quel est leur type, leur genre ? Relèvent-ils bien du domaine ?

En outre, ce ne sont pas les seules informations qui intéressent les utilisateurs. Les professionnels de la recherche d'information ont souvent des besoins très spécifiques en fonction des tâches à réaliser. Les techniques actuelles de référencement tiennent rarement compte de la pertinence des documents par rapport aux types d'utilisateurs et des modes d'utilisation de l'information.

Nous proposons d'ajouter aux critères de rappel et de précision la dimension de pertinence, accessible par une meilleure caractérisation des documents. Cet objectif s'obtient par des méthodes linguistiques, fournies en particulier par une approche « sémantique textuelle ».

La sémantique textuelle vise à extraire le sens du texte, sans le formaliser dans des ontologies (dont la généricité est remise en question par la nature culturelle de l'écrit). Son application a récemment permis, par exemple, de mettre en place un système informatique capable de repérer les textes racistes sur Internet (projet européen PRINCIP).

Les objectifs du projet C-MANTIC concernent deux axes :

1. D'une part, stabiliser et enrichir une méthodologie de caractérisation des documents adaptée à la recherche d'information (initié par le projet PRINCIP).
2. D'autre part, développer des outils qui permettraient d'appliquer la méthodologie – un moteur linguistique (LPU) et une interface utilisateur (CORPIST) pour la communication entre l'utilisateur final et LPU.

Le développement de la méthodologie fait l'appel à un corpus expérimental constitué essentiellement des documents disponibles sur l'internet et portant sur la thématique du tabagisme qui présente plusieurs intérêts : diversité de discours (médical, vulgarisé, publicitaire), enjeux sanitaires et économiques, caractère multilingue et multiculturel, résultats extensibles à de multiples domaines d'intérêt public. La nature exacte des documents a été définie par rapport aux besoins des utilisateurs tels que les organismes de prévention et le personnel médical.

Le corpus a deux fonctions:

– Évaluer et discriminer les documents pertinents : le corpus fonctionne comme un modèle implicite des différentes sortes de document cherchés et/ou rejetés;

– Permettre les contrastes: la stratégie différentielle adoptée (mises en relief statistique par des tests d'écart réduit) conduit à contraster un document sur un corpus qui relève uniquement de son genre (p.ex. thèses vs articles).

Les outils résultant de ce projet permettront de pérenniser la méthodologie et d'offrir à la communauté scientifique de nouvelles possibilités d'explorer les textes. Leur utilisation peut être très large et dépasse le seul domaine de la recherche d'information.