Extension d'un corpus d'articles scientifiques par recherche de similarités sémantiques : application à une problématique des sciences du sport

Fabrice MUHLENBACH

Univ. Lyon, UJM-Saint-Etienne, CNRS,
Laboratoire Hubert Curien UMR 5516, F-42023 Saint Etienne, France
fabrice.muhlenbach@univ-st-etienne.fr

La recherche scientifique menée dans les milieux académiques ou industriels est productrice de connaissances toujours plus nombreuses. Les chercheurs, pour apporter leurs contributions originales à ce savoir, doivent pouvoir accéder aux connaissances existantes (publications, brevets) et retrouver parmi elles les éléments pertinents (théories, démonstrations, méthodologies, résultats expérimentaux...) qui servent de socle à leurs travaux et permettent de délimiter les cadres de leurs propres apports scientifiques.

À l'heure actuelle, les publications scientifiques sont souvent accessibles soit directement sur des sites *web* de chercheurs, d'institutions, d'entreprises réalisant de la R & D ou d'éditeurs scientifiques, soit à travers des plates-formes spécifiques tels que les projets *Gallica*, *Google Livres* ou *ISTEX*¹. Ces bibliothèques numériques sont à la fois l'instrument et la matière première de la recherche et de l'innovation scientifique. Par conséquent, la maîtrise et l'utilisation efficace de ces sources constituent un enjeu stratégique pour le développement de la science, l'accroissement de la richesse économique et plus largement l'évolution de la société.

Néanmoins, l'exploration de ces gigantesques bibliothèques numériques n'est pas efficace en raison notamment des limitations cognitives humaines et du manque de temps. Il en résulte une exploitation faible de la richesse des bibliothèques numériques avec une focalisation à la fois sur les articles les plus récents (alors que des articles anciens pourraient pourtant s'avérer pertinents) et des articles limités à la communauté scientifique d'appartenance du chercheur (alors que des articles venant de disciplines complémentaires pourraient être intéressants).

Cette présentation propose plusieurs pistes menées par une équipe de recherche stéphano-lyonnaise afin de répondre à ce problème. À partir d'un corpus de base de quelques textes scientifiques de référence concernant un sujet donné, au moyen de techniques de fouille de textes, nous montrons qu'il est possible d'obtenir d'autres articles pertinents (dans notre cas, issus de la base ISTEX) liés au sujet du corpus d'entrée par des relations de similarités sémantiques. De plus, ces articles scientifiques nouveaux extraits des bibliothèques numériques viennent enrichir le corpus de départ en étant puisés dans des disciplines allant au-delà du seul domaine des articles fournis en exemple. Nous illustrons cette proposition par une problématique intéressant les sciences du sport, domaine par nature pluridisciplinaire puisque les pratiques sportives constituent un objet d'étude qui peut être abordé par différentes disciplines, en particulier sous les angles de la physiologie, de la psychologie ou de psycho-sociologie.

1. ISTEX : Initiative d'excellence de l'Information Scientifique et Technique, voir : <http://www.istex.fr/>