La fouille de textes *au service de la documentation*

Les masses de données textuelles aujourd'hui disponibles engendrent un problème spécifique lié à leur traitement automatique. Des méthodes de fouille de textes et de traitement automatique du langage peuvent en partie répondre à cette difficulté. Approche des procédés et des nouveaux défis à relever présentés par deux chercheurs du Cirad, centre de recherche français qui répond, avec les pays du Sud, aux enjeux internationaux de l'agriculture et du développement.

Les processus de fouille de textes sont souvent composés de deux phases successives. Dans un premier temps, ces méthodes consistent à extraire les *descripteurs linguistiques* les plus significatifs à partir de documents (résumés, articles scientifiques, etc.). Les descripteurs linguistiques peuvent être des mots simples (par exemple, « culture »), mais aussi des termes composés (par exemple, « agriculture familiale »). Nous appellerons de tels descripteurs linguistiques des *termes*. Ces derniers représentent le matériau de base qui associe une certaine sémantique aux documents. Par exemple, les termes « riz », « irrigation » et « Madagascar » présents dans un document mettent en lumière une thématique générale liée à l'*agronomie*. La deuxième phase du processus consiste à exploiter ces termes pour, par exemple, classer automatiquement les documents dans des catégories (*agronomie*, *élevage*, etc.). Cette classification repose sur le postulat suivant : si des documents possèdent de nombreux termes en commun alors ils peuvent être regroupés et reliés à une même thématique. Sur la base de cette hypothèse et de mesures de similarité associées, des méthodes d'*apprentissage automatique* permettent, à partir de textes annotés manuellement par les professionnels de la documentation, de construire automatiquement un modèle et un algorithme de classification. Ce dernier pourra alors être appliqué avec succès sur de nouveaux documents, permettant ainsi de traiter des milliers de textes de manière automatique. Ces approches sont bien sûr très dépendantes de la qualité des termes issus des données textuelles. Ce point est discuté dans la section suivante.

L'EXTRACTION DE LA TERMINOLOGIE

Les méthodes d'extraction de la terminologie combinent harmonieusement les informations syntaxiques et statistiques. Les *informations syntaxiques* permettent d'identifier des termes ayant des structures syntagmatiques complexes (patrons de type nom-adjectif, nom-préposition-nom, etc.). Pour extraire des termes respectant de telles structures, il est nécessaire d'apposer, au préalable, une étiquette grammaticale à chacun des mots du texte avec des étiqueteurs dédiés (étiqueteur de Brill, Tree Tagger, etc.).

Par ailleurs, les *informations statistiques* apportent une pondération des termes candidats extraits afin de proposer les plus pertinents aux experts. En effet, la fréquence d'un terme n'est pas nécessairement un critère de sélection adapté. À titre d'exemple, le mot « agriculture » présent dans de très nombreuses publications du Cirad se révèle en fait très général et pas suffisamment *discriminant*. Ainsi, des mesures de discriminance et d'autres méthodes de pondérations qui calculent, par exemple, la dépendance des mots composant les termes complexes peuvent être appliquées. Elles ont notamment été utilisées dans le cadre d'une étude menée en 2014 qui s'est portée sur un corpus de résumés bilingues (français et anglais) extraits d'Agritrop[1], la base de données de publications scientifiques du Cirad. Une synthèse des résultats obtenus est présentée dans la section suivante.

LA FOUILLE DE PUBLICATIONS SCIENTIFIQUES

Les méthodes de fouille de textes décrites précédemment ont été appliquées et adaptées à l'aide de divers pré- et post-traitements. Elles ont permis d'extraire automatiquement des termes en français tout à fait pertinents tels que « développement durable », « développement rural », « sécurité alimentaire », « croissance démographique », « aménagement du territoire », « gouvernance territoriale », etc. Ainsi, 28 % des termes simples et 12 % des termes composés sélectionnés avec nos systèmes sont présents dans un thésaurus du domaine de l'agriculture, Agrovoc[2] (vocabulaire contrôlé issu de la FAO – Food and Agriculture Organization – des Nations unies).

Ceci met en exergue que, d'une part, les termes fournis par les approches de fouille de textes sont pertinents et que, d'autre part, des termes nouveaux qui sont souvent des termes composés sont également mis en avant. Ils peuvent enrichir des thésaurus spécialisés. Une étude qualitative avec

[1] http://agritrop.cirad.fr

[2] http://aims.fao.org/fr/standards/agrovoc

Le terme « Madagascar » peut signifier une entité autant spatiale que politique. Des approches de désambiguïsation sémantique s'avèrent donc nécessaire dans le cadre de la fouille de publications scientifiques.

Luc Legay / Flickr (CC BY-SA 2.0)

un professionnel des ressources documentaires a confirmé ces premières constatations.

Par ailleurs, en partenariat avec les pays du Sud, l'intérêt pour un organisme de recherche comme le Cirad est de disposer d'une représentation spatiale de son activité. Une piste d'exploration a été amorcée en combinant ces techniques de fouille de textes et la prise en compte d'informations spatiales extraites dans les articles. Ces entités sont alors mises en relation avec des référentiels géographiques (par exemple, GeoName[3]). Notons que des approches de désambiguïsation sémantique peuvent, au préalable, distinguer une entité spatiale d'une organisation (par exemple, selon le contexte, le mot « Madagascar » présent dans un texte peut être une entité spatiale ou une entité politique). Ainsi, l'ensemble de ce processus permet de mettre en relief les articles scientifiques (représentés par un identifiant) traitant de sujets précis (termes extraits) en lien avec des entités spatiales (par exemple, un pays ou une ville). Les résultats peuvent alors être diffusés sur des plateformes dédiées à la publication de données en *open data* qui offrent certains modules de visualisation.

VERS DE NOUVEAUX DÉFIS...

Pour traiter les masses de données aujourd'hui disponibles (c'est-à-dire l'infobésité), la problématique de recherche du *big data* est classiquement mise en avant avec les 3 V qui la caractérisent : volume, variété et vélocité. Mais d'autres caractéristiques ne doivent pas être négligées comme la versatilité, la véracité, la visualisation ou la valorisation des données et informations. Toutes ces problématiques ouvrent de nouvelles disciplines de recherche comme la *science des données* qui mêle mathématiques, statistiques, informatique et visualisation.

Dans ce contexte, une voie de recherche tout à fait prometteuse concerne le croisement des données hétérogènes qui met en perspective des connaissances nouvelles. Dans le cadre de nos travaux, ceci revient à mettre en relation les publications scientifiques du Cirad avec d'autres types de données disponibles (enquêtes, images, bases de données, etc.). Aussi, l'extraction et l'exploitation de mots-clés disciplinaires et thématiques sur de la donnée, au sens large du terme, est envisageable dès lors qu'il existe des référentiels grâce auxquels un couplage avec les processus de fouille de textes et de données restitue des résultats qualifiés. Par itération de ce processus sur des données hétérogènes en associant des mesures de similarité, la capitalisation de l'ensemble pourrait constituer demain un résultat majeur.

Mathieu Roche et Sophie Fortuno
Cirad, UMR TETIS, Montpellier
mathieu.roche@cirad.fr
sophie.fortuno@cirad.fr

Remerciements
Nous remercions les documentalistes du Cirad, les étudiants du master IPS (Université Montpellier 2) et de la licence professionnelle Management des ressources numériques (Université Paul Valéry) et Juan Antonio Lossio (doctorant au Lirmm) qui ont activement participé à ces travaux de recherche liés à la fouille de textes.