Animitex : Analyse d'Images fondée sur des Informations Textuelles

Mathieu Roche, Maguelonne Teisseire, Bruno Crémilleux, Pierre Gancarski, Christian Sallaberry, Hugo Alatrista-Salas, Nicolas Béchet, Delphine Bernhard, Sandra Bringay, Thierry Charnois, et al.

ANIMITEX

Analyse d'Images fondée sur des Informations Textuelles

Mathieu Roche[1, 2], Maguelonne Teisseire[1, 2], Bruno Crémilleux[3], Pierre Gancarski[4], Christian Sallaberry[5] *et al.*

(1) UMR TETIS, AgroParisTech, Cirad, Irstea, (2) LIRMM, CNRS, Univ. Montpellier 2, (3) GREYC, CNRS, Univ. Caen, (4) ICube, CNRS, Univ. Strasbourg, (5) LIUPPA, Université de Pau et des Pays de l'Adour

DOMAINE : Big data

MOTS-CLES : Extraction de connaissances et recherche d'information pour les données spatiales, temporelles et thématiques

KEYWORDS: Knowledge discovery and information retrieval for spatial, temporal, and topic data

1. Objectif du projet ANIMITEX

Le Web offre une quantité imposante de données de type textuel et nombreux sont les chercheurs de la communauté s'intéressant à la problématique d'extraction de connaissances dans de telles données. Plus récemment, l'accroissement du nombre d'images satellitaires très hautes résolutions mises à disposition auprès des utilisateurs finaux et des scientifiques permet une perception plus fine des phénomènes spatio-temporels mais pose le problème d'une analyse efficace et rapide d'une telle masse de données. L'objectif du projet CNRS ANIMITEX est d'exploiter l'ensemble des données ainsi mises à disposition : données textuelles massives et hétérogènes (blogs, rapports, articles de presse) et données satellitaires dans un cercle vertueux d'enrichissement des informations spatiales, temporelles et thématiques pouvant s'y trouver.

Ces très gros volumes de données sont associés à une répétitivité temporelle de plus en plus élevée, passant d'une dizaine d'images par an (satellites SPOT, Landsat, ...) à environ une image tous les 5 jours d'ici trois ans (satellites Sentinel-2). Le projet ANIMITEX a de nombreux domaines d'application. Par exemple, il permet d'aider l'annotation des images rendant ainsi possible une classification plus fine de ces données. Par ailleurs, l'appariement images-textes permettra d'enrichir les méthodes de recherche d'information et d'offrir à l'utilisateur un point de vue plus global des données. Ceci peut se révéler crucial pour le décideur dans le cadre de projets d'aménagement en exploitant les dires d'experts (gestionnaires, scientifiques, associations, entreprises spécialisées, etc.) relativement à un territoire. Dans ce cadre, nous souhaitons étudier avec plus de précision la construction d'une rocade au nord de Villeveyrac et d'une zone d'activité (projet Hinterland) dans la région de Thau (à proximité de Sète, France).

2. Données et processus mis en place

Les travaux actuels se concentrent sur l'utilisation et l'adaptation de techniques de Traitement Automatique du Langage Naturel (TALN) pour la reconnaissance des Entités Nommées Spatiales ainsi que des informations thématiques et temporelles. Pour cela, un jeu d'articles de presse (corpus textuel de 12000 textes) relatif à la région du Bassin de Thau entre les années 2010 et 2013 a été constitué pour réaliser les premiers tests de marquage. Une seconde partie du jeu de test est composée de fichiers raster (mosaïques d'images Pléiades - résolution spatiale 2x2 m - 4 bandes spectrales) couvrant l'ensemble du Bassin de Thau (cf. figure 1). Ces images satellitaires ont été mises à disposition à travers l'Equipex GEOSUD (http://www.equipex-geosud.fr/). Une classification détaillée de l'occupation du sol est actuellement en cours, elle aboutira à une couche numérique vectorielle où chaque entité spatiale (représentée sous forme de polygone) appartient à une classe

précise d'occupation du sol. La nomenclature de cette classification s'organise en 4 niveaux hiérarchiques (cf. figure 2). Nous nous sommes également préoccupés du problème multi-échelle associé aux différents niveaux de classification des images satellitaires mises à disposition.

Une fois la constitution des corpus effectuée, des méthodes de TALN ont permis d'identifier des descripteurs linguistiques véhiculant des informations spatiales, thématiques et temporelles dans les textes. L'utilisation combinée de lexiques et de règles dédiées (Gaio *et al.* 2012) permet d'identifier les entités spatiales absolues (par exemple, *Montpellier*) et relatives (par exemple, *au sud de Montpellier*) (Kergosien *et al.*, 2014). Une première chaîne de traitements fondée sur l'extraction de motifs séquentiels (Cellier *et al.,* 2010) a également été proposée afin de découvrir des relations entre entités spatiales (Alatrista Salas et Béchet, 2014). Les informations thématiques sont quant à elles identifiées par la reconnaissance de termes provenant de ressources sémantiques (thesaurus Agrovoc, nomenclature issue de classifications d'images, etc.) (Buscaldi et al. 2013).

Ces descripteurs linguistiques serviront d'éléments d'ancrage pour repérer les phénomènes décrits. Il s'agit ensuite de lier ces phénomènes annotés avec ceux identifiés dans les images interprétées traitant de la même zone et du même phénomène au même moment (Forestier *et al.*, 2012). L'objectif est d'utiliser les différences entre les deux modes d'expressions (texte vs. images interprétées) dans la spécification des objets géographiques et de leurs relations. L'appariement Texte-Image correspond à une forme de *Recherche d'Information* multicritère dans laquelle les descripteurs d'un couple texte-image à apparier ont différents degrés de recoupement. L'objectif est de permettre l'enrichissement de l'information véhiculée par un texte à l'aide d'images et inversement.

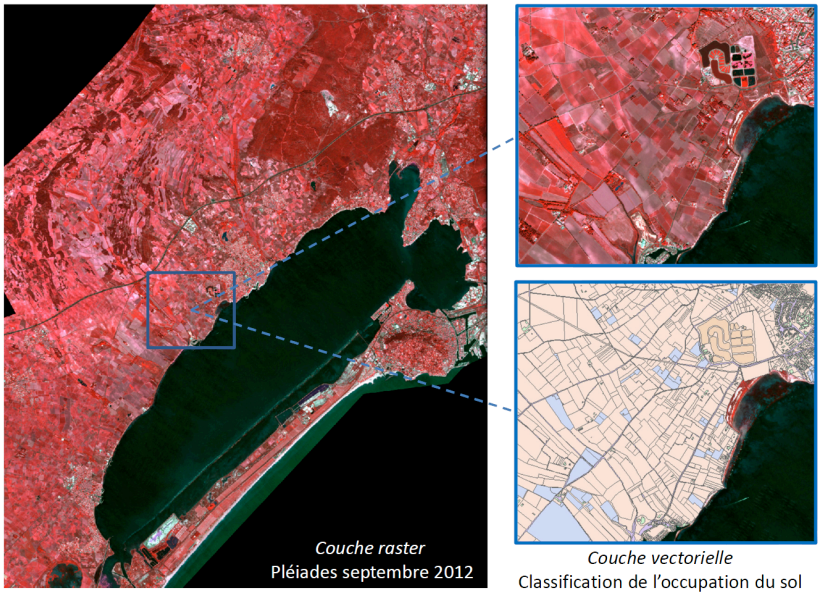

Figure 1 : Mosaïque images Pléiades autour du Bassin de Thau, les aperçus sur la droite représentent la superposition d'une classification vectorielle sur le fichier raster.

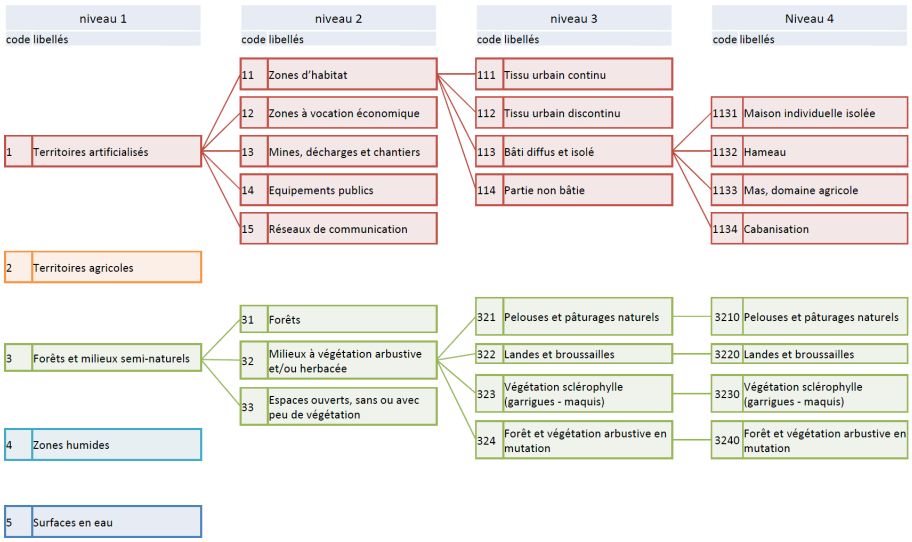

Figure 2 : Schéma illustrant la nomenclature de la classification du Bassin de Thau. Les niveaux de la hiérarchie sont détaillés pour quelques branches à titre d'exemple.

3. Partenariat et contact

Le consortium du projet (http://www.lirmm.fr/~mroche/ANIMITEX/) est composé des Universités de Caen, Montpellier, Pau et Strasbourg. Il fait appel aux trois disciplines impliquées : *informatique*, *géographie* et *télédétection*. Le LIRMM (Montpellier) s'intéresse aux problématiques d'extraction de connaissances à partir de gros volume de données. L'UMR ICube (Strasbourg) est spécialisée dans l'analyse d'images et la fouille de données complexes. ICube accueillera pour la durée du projet des spécialistes géographes du LIVE (Laboratoire Image, Ville et Environnement) et des spécialistes en analyse de textes du LiLPa (Linguistique, langues, parole) de l'Université de Strasbourg afin de former un pôle de compétences local sur tous les principaux aspects du projet. L'UMR TETIS «Territoires, Environnement, Télédétection et Information Spatiale» (Montpellier) a pour vocation de produire et diffuser des connaissances, des concepts, des méthodes et des outils permettant de

caractériser et de comprendre les dynamiques des espaces ruraux et des territoires, et de maîtriser l'information spatiale sur ces systèmes. Son expertise, en télédétection et fouille de données spatio-temporelles, complexes et hétérogènes, est un des socles du projet. Le partenaire LIUPPA (Pau), EA 3000, regroupe des chercheurs spécialisés dans le marquage, l'extraction et la recherche d'informations géographiques. L'UMR GREYC (Caen) regroupe des chercheurs spécialisés en fouille de données et traitement automatique des langues (TAL) qui apportent leurs compétences en fouille de séquences et fouille de textes.

Bibliographie

Buscaldi D., Bessagnet M.N., Royer A., Sallaberry C., Using the Semantics of Texts for Information Retrieval: A Concept and Domain Relation-Based Approach. ADBIS (2) : 257-266, 201

Gaio M., Nguyen V.T., Sallaberry C., Typage de noms toponymiques à des fins d'indexation géographique. *Traitement Automatique des Langues, Vol. 53(2),* pp.143-176, 2012

Cellier P., Charnois T., Plantevit M., Crémilleux B., Recursive Sequence Mining to Discover Named Entity Relations, *Symposium on Intelligent Data Analysis*, LNCS, pp. 30-41, 2010.

Forestier G., Puissant A., Wemmert C., Gançarski P. (2012), Knowledge-based Region Labeling for Remote Sensing Image Interpretation. Computers, Environment and Urban Systems, Vol. 36(5), pp. 470-480, 2012

Alatrista Salas H., Béchet N., Fouille de textes : une approche séquentielle pour découvrir des relations spatiales. *Atelier Cergeo - EGC*, 2014

Kergosien E., Laval B., Roche M., Teisseire M., Are opinions expressed in land-use planning documents? *International Journal of Geographical Information Science*, Vol. 28(4), pp.739-762, 2014

Auteurs additionnels : Hugo Alatrista-Salas, Nicolas Béchet, Delphine Bernhard, Sandra Bringay, Thierry Charnois, Mauro Gaio, Fábio N. Guttler, Dino Ienco, Eric Kergosien, Pierre Maurel, Pascal Poncelet, Arnaud Sallaberry, Christiane Weber