Stage Recherche - M2

Mise en relation des opinions et informations géographiques extraites de corpus texte liés à l'aménagement du territoire

E. Kergosien et M. Roche

Le travail proposé s'inscrit dans un projet plus large Senterritoire (http://www.msh-m.fr/programmes/senterritoire) qui a pour objectif de proposer un environnement décisionnel fondé sur une analyse automatique des textes liés à l'aménagement du territoire. Il est porté par l'équipe Texte du LIRMM dans le cadre d'un fort partenariat scientifique pluridisciplinaire avec l'équipe Tatoo et le laboratoire TETIS.

CONTEXTE GLOBAL :

De nombreux travaux ont été réalisés par la communauté Extraction de connaissances sur la fouille d'opinion et de sentiments dans des données, qu'elles soient structurées ou non. L'intérêt de ces travaux n'est plus à démontrer dans des applications comme les critiques de films, matériels hifi, restaurants, ... Dans le contexte de l'aménagement du territoire, où se posent les questions de l'appropriation de l'information géographique par les acteurs, la problématique devient plus complexe. En effet, elle nécessite de mettre en relation la notion d'opinion (ou sentiment) à celle d'information géographique caractérisant un territoire. Dans le cadre de nos travaux, une information géographique est composée au moins d'une entité thématique et d'une entité spatiale.

Des premiers travaux menés dans le cadre du projet Senterritoire ont permis de générer une liste d'entités spatiales (Tahrat et al, 2013), une liste d'opinions spécialisée (Kergosien et al., 2013) et une liste de thèmes extraits des textes en utilisant le thésaurus Agrovoc. Aussi, il existe de nombreux travaux utilisant les textes ou des connaissances du domaine comme support à des méthodes d'extraction de relations sémantiques. Cependant, les approches se concentrent sur la mise en relation de l'un des éléments du triptyque *opinion*, *thème* et *entité spatiale*, et non de l'ensemble. Aussi, dans la plupart de ces approches dites supervisées, il est nécessaire de disposer d'un jeu d'entrainement afin d'apprendre le modèle de classification. Or, les relations ne sont pas toujours déterminées a priori et leurs découvertes conduisent à une valeur ajoutée dans les applications réelles. Dans le cadre des recherches proposées pour ce stage, nous sommes dans un contexte exploratoire et non supervisé afin de réaliser la mise en correspondance de connaissances par la découverte des liens spatiaux, thématiques et d'opinions existant, qui pris individuellement ne permettent pas de répondre aux enjeux sociétaux.

OBJECTIF :

Dans le cadre de ce stage de master, les recherches à mener par l'étudiant s'inscrivent en complément des étapes initiées dans le projet et concernent plus précisément la mise en place d'une méthode automatisée pour détecter des relations sémantiques entre des éléments de type *opinion*, *thème* et *entité spatiale* extraits du corpus constitué de documents texte décrivant l'aménagement du territoire du bassin de Thau. Un travail intermédiaire consiste à identifier l'échelle (syntagme nominal, phrase, paragraphe, document, etc.) la plus adaptée pour extraire ces relations.

Méthodologie :

1. Réalisation d'un état de l'art sur :
    a. La modélisation des informations à structurer : le triptyque *opinion-thème-entité spatiale* ;
    b. L'identification de relations entre les éléments du triptyque. Il faudra réaliser une veille exhaustive regroupant les approches linguistiques et fouille de textes, ainsi que les approches hybrides.
2. Définition et mise en oeuvre d'une méthode automatisée pour identifier des relations entre éléments du triptyque *opinion*, *thème* et *entité spatiale* selon différentes échelles (mots, groupe de mots, phrase, paragraphe, document). Pour cela, l'étudiant s'appuiera sur les premiers résultats des travaux menés dans l'équipe, à savoir une liste d'entités spatiales, une liste d'opinions de domaine et une liste de thèmes extraits des textes en utilisant le thésaurus Agrovoc.
3. Proposition d'une approche permettant de généraliser les relations identifiées en relations sémantiques en exploitant les structures des différentes ressources (thésaurus Agrovoc pour le thème, gazetteers pour les informations spatiales, dictionnaires de sentiments).

MOTS-CLES :

Traitement Automatique du Langage Naturel, Fouille de textes, thésaurus, Information géographique, Aménagement du territoire.

Durée : 5 mois

Gratification : 436,05 € mensuel

Références :

- S. Tahrat, E. Kergosien, S. Bringay, M. Roche, M. Teisseire: Text2Geo : des données textuelles aux informations géospatiales. EGC 2013: 407-412
- E. Kergosien, P. Maurel, M. Roche, M. Teisseire. OPITER : Fouille de données d'opinion pour les territoires, In Spatial Analysis and GEOmatics (Sagéo'13), Brest, 2013.