Exploitation de l'hétérogénéité dans les données textuelles

Utilisation de données produites à Madagascar

Harness the hetorogeneity in textual data

Use of data produced in Madagascar

Jacques Fize[1,2], Mathieu Roche[1,2], Maguelonne Teisseire[2,3]

[1] CIRAD, Montpellier, France
{prénom.nom}@cirad.fr
[2] TETIS, Irstea, AgroParisTech, CIRAD, CNRS, Univ Montpellier, Montpellier, France
[3] IRSTEA, Montpellier, France
maguelonne.teisseire@irstea.fr

RÉSUMÉ. Depuis plusieurs décennies, on observe une utilisation croissante des systèmes d'information, ce qui provoque une augmentation exponentielle des données textuelles. Bien que l'aspect volumétrique de ces données textuelles soit résolu, sa dimension hétérogène reste un défi pour la communauté scientifique. La maitrise de ces données hétérogènes offre de nombreuses opportunités par un accès à une information plus riche. Dans nos travaux, nous concevons un processus de mise en correspondance de données textuelles hétérogènes, basé sur leur spatialité. Dans cet article, nous présentons les résultats retournés par ce processus sur des données produites à Madagascar dans le cadre du projet BVLAC, dirigé par le CIRAD. En se basant sur un ensemble de 4 critères de qualité, nous obtenons de bonnes correspondances spatiales entre ces documents.

ABSTRACT. Over the last decades, there has been an increasing use of information systems, resulting in an exponential increase in textual data. Although the volumetric dimension of these textual data has been resolved, its heterogeneous dimension remains a challenge for the scientific community. The management of the heterogeneity in data offers many opportunities through an access to a richer information. In our work, we design a process for mapping heterogeneous textual data, based on their spatiality. In this article, we present the results returned by this process on data produced in Madagascar as part of the BVLAC project, led by CIRAD. Based on a set of 4 quality criteria, we obtain good spatial correspondence between these documents.

MOTS-CLÉS. Fouille de texte, similarité spatiale, représentation spatiale.

KEYWORDS. text-mining, spatial similarity, spatial representation.

Introduction

Depuis plusieurs décennies, on observe une utilisation croissante des systèmes d'information, qui provoque une augmentation exponentielle des données textuelles accessibles. Ces données sont produites par des acteurs scientifiques, industriels, gouvernementaux, et citoyens à travers l'utilisation de services tels que les réseaux sociaux, les sites d'e-commerce, ou encore de crowdsourcing. Bien que l'aspect volumétrique de ces données textuelles soit globalement résolu, sa dimension hétérogène reste un défi pour la communauté scientifique. La maîtrise de l'hétérogénéité des données offre de nombreuses opportunités comme l'accès à une information plus complète et représentée selon différents angles (statistique, sociologique, politique, etc.). Dans le but de rapprocher les différentes données associées à une information, de nouvelles méthodes pour mesurer la similarité sont nécessaires.

Dans cet article, nous présentons un processus de mise en correspondance de données hétérogènes, sur leur dimension spatiale. Ce processus est fondé sur une représentation de la spatialité dans les textes : la Spatial Textual Representation. La STR est une structure graphe composée des entités spatiales présentes dans les données et les

relations spatiales (voisinage, inclusion) qu'elles entretiennent. Dans nos travaux, chaque étape du processus – i.e. génération de la STR et mise en correspondance – est évaluée sur différents corpus de documents. Dans cette article, nous présentons les résultats obtenus sur le corpus *AgroMada*, composé de documents hétérogènes produits par le CIRAD [1], sur la thématique de l'agroécologie à Madagascar.

Cet article est organisé de la manière suivante. Dans la section 2, nous présentons le processus de création et de comparaison de données basé sur la spatialité. Dans la section 3, nous présentons les données et les protocoles d'évaluation mis en place. Enfin, nous concluons dans la section 4.

1. Travaux Connexes

Il existe de nombreuses méthodes (Pang *et al.*, 2015; Demisse *et al.*, 2017; Ma *et al.*, 2018) pour faire de l'appariement de données hétérogènes structurées (méta-données). Toutefois, peu couvrent également le problème des données hétérogènes non-structurées et de leur similarité spatiale (Patanè & Spagnuolo, 2016). Dans cet article, nous abordons l'appariement de données textuelles hétérogènes (structurées et non structurées) à travers la dimension spatiale.

2. Mise en correspondance de données textuelles hétérogènes

Le processus de mise en correspondance est divisé en deux étapes : (i) la génération des représentations de la spatialité dans chaque document, puis (ii) puis la mesure de la similarité entre celles-ci. Les sous-sections suivantes présentent chacune de ces étapes.

2.1. Représentation de la spatialité : STR

Pour représenter la spatialité dans les textes, nous utilisons la STR, ou Spatial Textual Representation (voir Figure 2.1). La STR (Fize *et al.*, 2017) est une représentation graphe, où les sommets correspondent aux entités spatiales, connectées selon leurs relations spatiales (adjacence, inclusion) sous forme d'arêtes. La génération de la STR pour un document se déroule en deux étapes : (i) l'extraction d'entités spatiales, ou *Geoparsing* (ii) la liaison et la transformation de celles-ci, ou *Geocompletion*.

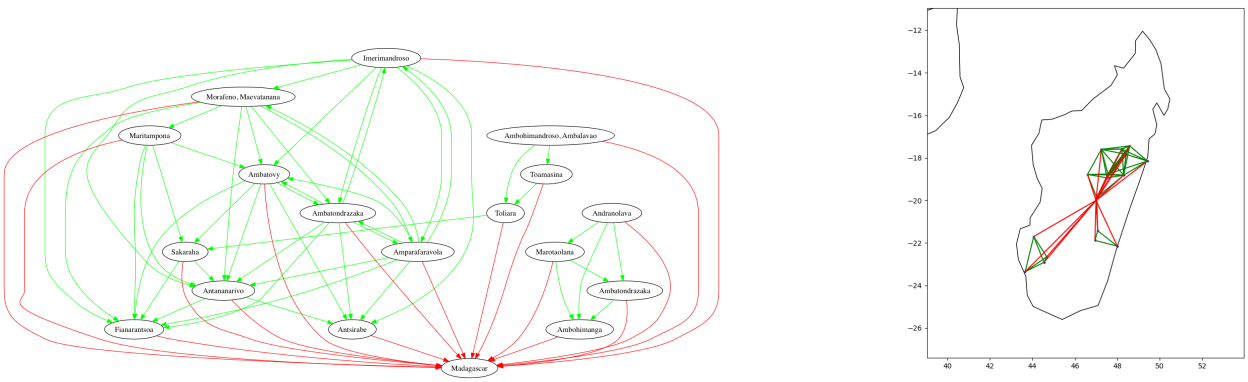
2.1.1. Geoparsing

La construction de la STR est fondée sur l'extraction d'indicateurs spatiaux dans un document, ici les lieux mentionnés. Pour cela, nous utilisons des méthodes de **geoparsing** divisées en deux phases : le *geotagging* puis le *geocoding*.

La première phase, ou **geotagging**, consiste à identifier les noms de lieux (toponymes) dans le texte. Afin d'identifier les noms de lieux, on utilise des méthodes de reconnaissance d'entités nommées (NER). Puis, une fois les toponymes identifiés, la deuxième phase (ou **geocoding**) consiste à associer chaque toponyme avec une entrée dans un géoréférentiel. Un géoréférentiel est un jeu de données géographiques où chaque entrée est associée à un nom, des coordonnées et d'autres informations (classe, alias,...). Pour répondre à nos besoins, nous avons créé un géoréférentiel, appelé Geodict (Fize & Shrivastava, 2017), construit à partir de Wikidata, OpenStreetMap et Geonames.

La difficulté du processus de geoparsing réside dans l'ambiguïté possible d'un toponyme. Par exemple, il existe plus de 60 entités spatiales ayant le nom "Paris". Pour résoudre ces ambiguïtés, nous utilisons une stratégie de désambiguïsation. Cette stratégie de désambiguïsation associe l'entité spatiale la plus communément utilisée pour un toponyme. Par exemple, *Paris* sera associé à *Paris, France*. Cette association se base sur la popularité d'une entité spatiale calculée à l'aide de l'algorithme de PageRank sur le corpus de Wikipedia.

1. Centre de coopération internationale en recherche agronomique pour le développement



(a) *(b)*

FIGURE 2.1. *Une STR centrée autour de Madagascar (a) et sa projection(b) sur une carte*

2.1.2. Geocompletion

Une fois les entités spatiales identifiées dans un document, la seconde étape consiste à relier celles-ci par leurs **relations spatiales**. Nous avons sélectionné deux types de relations spatiales. Premièrement, la relation d'inclusion donne une indication sur la position dans la hiérarchie spatiale d'une entité. Deuxièmement, la relation d'adjacence indique la proximité de certaines entités entre elles. Pour extraire les différentes relations, nous utilisons les informations spatiales provenant d'un géoreférentiel (Fize & Shrivastava, 2017). Puis, une fois les entités spatiales reliées, on obtient une première version de la STR comme illustré dans la Figure 2.1.

Toutefois, l'information spatiale intégrée dans la STR peut être incomplète. Premièrement, la présence de bruits peut nuire à l'extraction de toponymes, particulièrement dans les données hétérogènes. Deuxièmement, certaines entités sont implicites ou connues uniquement du producteur et de l'utilisateur de la donnée. Pour compléter l'information contenue dans la STR, nous proposons deux **transformations de la STR** : la *généralisation* et l'*extension*.

La **généralisation** est un processus qui transforme chaque entité de la STR par son entité supérieure dans la hiérarchie spatiale. Par exemple, *Paris* deviendra *Ile-de-France*. La généralisation d'une entité est limitée par un niveau de granularité (Pays, région, etc.). L'**extension** est un processus qui étend les entités avec un faible impact, i.e. rarement mentionnée, avec une entité proche avec un impact plus fort. Concrètement, pour chaque entité avec un faible impact, on ajoute une entité populaire existante dans un radius défini. Chaque transformation est illustrée dans la Figure 2.2.

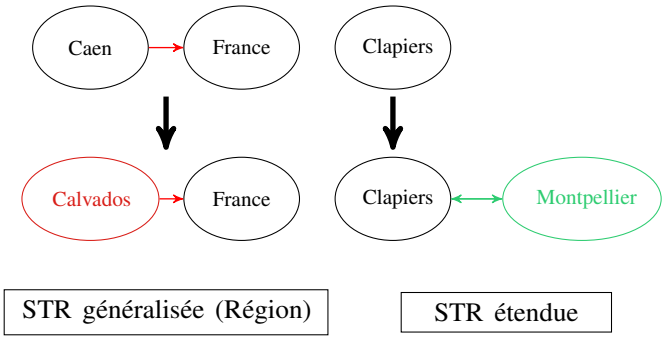


FIGURE 2.2. *Différentes transformations de la STR*

2.2. Mesurer la similarité entre les STRs

Pour mesurer la similarité entre les STRs, nous utilisons des algorithmes de *graph matching* (Fize *et al.*, 2018). On distingue deux grandes catégories d'algorithmes : *structure-based* et *pattern-based*. La première catégorie regroupe les algorithmes comparant uniquement la structure des graphes (sommets, arêtes). La deuxième catégorie regroupe les algorithmes utilisant les motifs présents dans les graphes pour faire la comparaison.

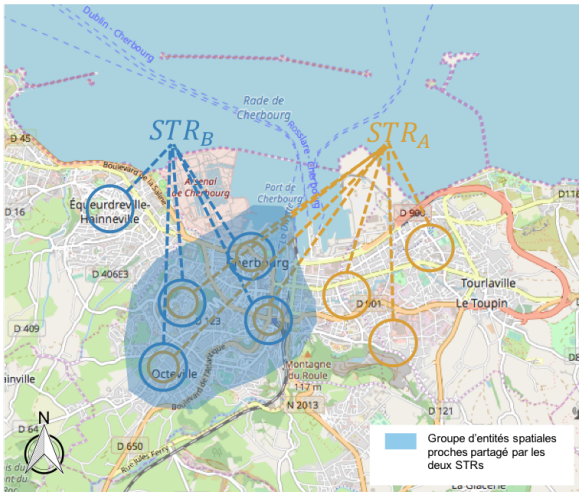


FIGURE 3.3. *Illustration d'un groupe d'entités spatiales respectant le troisième critère*

3. Expérimentations

3.1. AgroMada : données produites à Madagascar

Dans la dernière décennie, le CIRAD s'est engagé dans le projet BVLAC, dont l'objectif était de promouvoir l'utilisation de techniques d'agriculture durable sur le territoire de Madagascar. À l'issue de ce projet, un ensemble de données variées a été produit : *thèses, mémoires, rapport, relevés, fiche technique, etc*. Le corpus AgroMada est composé de 5552 documents hétérogènes en français, en anglais et contenant des termes appartenant à un lexique sur l'agroécologie. Tous les jeux de données utilisés sont mis à dispositions sur le Dataverse du CIRAD (Fize, 2018a,b).

3.2. Évaluation de la mise en correspondance

Pour évaluer la qualité de la mise en correspondance entre les STRs, nous proposons d'utiliser 4 critères :

1. **Entités Spatiales Communes (ESC).** Ce critère est validé si les deux STR possèdent des entités spatiales communes.
2. **Entités Spatiales Proches (ESP).** Ce critère est validé si certaines entités propres à un graphe possèdent une proximité avec d'autres entités d'un autre graphe. Cette proximité se traduit par l'existence de relations spatiales (inclusion, voisinage) ou d'une faible distance géodésique [2].
3. **L'emprise spatiale caractéristique (ESCC).** Ce critère est validé si des groupes significatifs d'entités spatiales proches sont présents dans les deux STRs. Un groupe d'entités spatiales est significatif s'il contient un nombre d'entités supérieur à la moyenne. Un exemple de groupes d'entitiés spatiales validant le critère est illustré dans la Figure 3.3.
4. **L'emprise spatiale stricte (ESS).** Ce critère est validé quand les deux STR possèdent une disposition géographique commune de leurs entités spatiales.

Pour cette évaluation, nous avons sélectionné un échantillon de 100 documents du corpus AgroMada. Puis, pour chaque document, on associe les n documents les plus similaires selon différentes combinaisons (type STR, mesure de similarité). Une fois les correspondances annotées selon les 4 critères, la valeur moyenne de la **précision@n** est calculée. Elle indique la proportion moyenne des n-plus similaires STRs pertinentes pour une STR. Ici, la pertinence se traduit par le respect d'un critère. Dans la Table 3.1, nous indiquons les combinaisons de mesure et de type de STR avec des valeurs de précision dominantes sur l'ensemble des critères.

Résultats Au travers des résultats de la Table 3.1, on observe des valeurs de précision élevées concernant la mise en correspondance selon les critères de partage d'entités spatiales (ESC) ou de groupement d'entités (ESCC).

2. Distance entre deux points sur la surface d'un ellipsoïde (ici la Terre).

Tableau 3.1. *Couple de mesure (Fize et al., 2018) et de type de STR formant la frontière de Pareto*

Mesure	Transformation	*ESC*	*ESP*	*ESCC*	*ESS*
VertexEdgeOverlap	$\emptyset$	**0,97**	0,31	**0.97**	0,36
MCS	Extension (n=2)	0,96	**0.43**	0,96	**0,38**
MCS	Extension (n=1)	**0,97**	0,42	0,96	0,37

Cependant, on obtient une valeur de précision moyenne sur les critères ESP ou ESCC dû à leur rareté dans les correspondances. Parmi les combinaisons dominantes, on souligne que l'utilisation de versions transformées de la STR, ici l'extension, améliore globalement la précision du système. Enfin, la mise en correspondance utilisant des algorithmes *structured-based* (Fize *et al.*, 2018) permet d'obtenir de meilleures valeurs de précision en particulier pour le critère *ESP*.

4. Conclusion

Dans cet article, nous présentons un processus de mise en correspondance de données textuelles appliqué à des données hétérogènes produites à Madagascar. Ce processus repose sur une représentation spatiale des données, la STR et un ensemble d'algorithmes (graph matching) pour les mesures de similarité. La génération de cette représentation est effectuée en deux phases : l'extraction d'entités spatiales (geoparsing) et leur connexion basée sur leurs relations spatiales (geocompletion). Nous proposons différentes transformations pour obtenir une meilleure approximation de la spatialité d'un document. À travers une évaluation fondée sur 4 critères, nous obtenons de bons résultats en matière de qualité de mise en correspondance.

Remerciements : Ces travaux sont financés dans le cadre du projet **SONGES** (Occitanie et FEDER). Nous remercions Jean-Philippe Tonneau pour son expertise pour la contribution et l'analyse des données du CIRAD.

Références

Demisse G., Tadesse T., Atnafu S., Hill S., Wardlow B., Bayissa Y. & Shiferaw A. (2017). Information Mining from Heterogeneous Data Sources : A Case Study on Drought Predictions. *Information*, 8(3), 79.

Fize J. (2018a). Bvlac corpus - extracted data. *CIRAD Dataverse*. http://dx.doi.org/10.18167/DVN1/8LIG1D.

Fize J. (2018b). Spatial representation of the bvlac and padi-web corpora. *CIRAD Dataverse*. http://dx.doi.org/10.18167/DVN1/JLXBLA.

Fize J., Roche M. & Teisseire M. (2017). Spatial Textual Representation (STR) ou comment représenter la spatialité des donnnées textuelles. In *Spatial Analysis and GEOmatics 2017*, Rouen, France : INSA de rouen.

Fize J. & Shrivastava G. (2017). GeoDict : an integrated gazetteer. In *Proceedings of Language, Ontology, Terminology and Knowledge Structures Workshop (LOTKS 2017)*, Montpellier, France : Association for Computational Linguistics.

Fize J., Teisseire M. & Roche M. (2018). Matching Heterogeneous Textual Data Using Spatial Features. In *SSTDM 2018 : International Workshop on Spatial and Spatiotemporal Data Mining*, p. 1389–1396.

Ma Z., Zhao Z. & Yan L. (2018). Heterogeneous fuzzy XML data integration based on structural and semantic similarities. *Fuzzy Sets and Systems*, 351, 64–89.

Pang L. Y., Zhong R. Y., Fang J. & Huang G. Q. (2015). Data-source interoperability service for heterogeneous information integration in ubiquitous enterprises. *Advanced Engineering Informatics*, 29(3), 549–561.

Patanè G. & Spagnuolo M. (2016). Heterogenous Spatial Data : Fusion, Modeling, and Analysis for GIS Applications. *Synthesis Lectures on Visual Computing*, 8(2), 1–155.