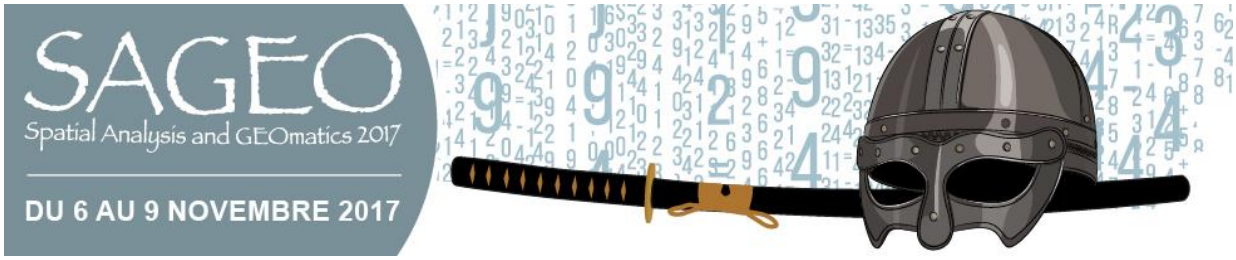

Actes de l'atelier

EXCES - **EXtraction de Connaissances à partir de donnéEs Spatialisées**

Eric Kergosien (GERiiCO, Université Lille 3)
Christian Sallaberry (LUIPPA, Université Pau, Pays de l'Adour)
Maguelonne Teisseire (IRSTEA, UMR TETIS, Montpellier)

https://sageo2017.sciencesconf.org/resource/page/id/20

Lundi 6 novembre 2017, Rouen

PRÉFACE

La production et les usages des ressources numériques géoréférencées sont en pleine expansion. Afin d'exploiter leurs contenus, les documents sont annotés, indexés et analysés conformément à des modèles de données dédiés à la description de domaines particuliers.
Ces dernières années une variété de travaux a mis en exergue le potentiel de l'extraction, de l'analyse et de la recherche de l'information géographique dans des corpus composés de documents textuels, d'images, de cartes, ... Un certain nombre de moteurs ou de services spécifiques à la recherche d'informations géographiques ont été proposés : ils visent les informations spatiales pour leur grande majorité, mais également les informations spatio-temporelles et thématiques, pour d'autres.
Le but de cet atelier est de rassembler la communauté croissante des professionnels et chercheurs du domaine de l'extraction, de la recherche et de l'analyse d'information géographique et des applications qui en découlent. Il se situe à la croisée de plusieurs disciplines : bien entendu en géomatique, mais également en sciences cognitives, en traitement automatique des langages naturels, en fouille de données, en extraction d'information et en recherche d'information.
Comment exploiter de façon efficace les dimensions thématiques, spatiales et temporelles de l'information disponible sur le Web ? Comment utiliser la complémentarité de sources externes de connaissances ?

Ces actes regroupent les soumissions acceptées à l'atelier EXCES en 2017 dans le cadre de la conférence Spatial Analysis and GEOmatics (SAGEO). Cette année, l'atelier a été organisé en trois temps : une présentation invitée de Davide Buscaldi (Université de Paris 13, LIPN) sur le thème « *Information Géographique, textes et média sociaux* », puis un ensemble de présentations orales des articles retenus pour l'atelier et finalement un temps dédié à la discussion avec l'ensemble des participants. Nous espérons que le lecteur qui n'a pu y assister trouvera toutes les informations dans les articles de ce volume.

Nous tenons à remercier tous les auteurs pour leurs propositions d'articles. Nous remercions également chaleureusement Davide Buscaldi, maître de conférences à l'université Paris 13, pour son intervention en tant que conférencier invité à la journée EXCES'2017. En espérant que ces articles vous apporteront de nouvelles perspectives sur la gestion et l'analyse de données spatiales et temporelles, nous vous souhaitons une bonne lecture.

Eric Kergosien
Université Lille-3/GERiiCO

Christian Sallaberry
Université UPPA/LIUPPA

Maguelonne Teisseire
UMR TETIS/IRSTEA

TABLE DES MATIÈRES

Session 1 : Analyse spatiale et médias sociaux

Session 2 : Exploitation automatisée de textes – toponymie, dynamiques spatiales

Gemedoc : Un outil pour annoter les correspondances entre les documents

Jacques Fize[1], **Maguelonne Teisseire**[1], **Mathieu Roche**[1]

** UMR 9000 TETIS, Cirad, Irstea, CNRS, AgroparisTech, Univ. Montpellier Maison de la Télédétection, Montpellier, France*
{firstname}.{lastname}@teledetection.fr

RÉSUMÉ. *Nous présentons* Gemedoc, *une plateforme pour annoter la similarité inter-document pour un corpus sur différentes dimensions : thématique et spatiale. Pour évaluer la similarité, nous avons conçu un protocole d'annotation divisé en deux étapes : (1) l'identification de descripteurs pour chaque dimension; (2) l'annotation de la similarité sur une échelle de 4 degrés. À terme, les annotations récoltées doivent permettre de construire un corpus destiné à évaluer les méthodes et les représentations dans des applications de mise en correspondance de documents.*

ABSTRACT. *We present* Gemedoc *a platform for text similarity annotation for a corpus on different dimensions: spatial and thematic. In order to annotate the similarity between two documents, we designed an annotation protocol divided in two steps: (1) identification of dimension features; (2) similarity annotation on a 4-degree scale. Ultimately, gathered annotations will permit to build a corpus aimed to evaluate methods and representations in text matching applications.*

MOTS-CLÉS : *fouille de textes, mise en correspondance de textes, plateforme d'annotation*

KEYWORDS: *text mining, text matching, annotation platform*

Introduction

En Recherche d'Information, l'identification de correspondances ou l'alignement entre données textuelles est essentiel. De manière générale, cette recherche se concentre sur la mise en relation de deux objets, la requête et le(s) document(s) affilié(s). Ces deux objets comparés sont souvent de tailles différentes et pour surmonter cette différence, diverses méthodes sont mises en place pour extraire le plus d'informations possibles. Parmi celles-ci, nous pouvons mentionner les travaux d'extension de requêtes (Xu, Croft, 1996 ; Dalton *et al.*, 2014), qui ont pour objectif d'étendre l'ensemble des descripteurs associés à la requête.

D'autres travaux utilisent des méthodes d'alignement de documents, notamment, dans le domaine de questions-réponses (Voorhees *et al.*, 1999 ; Voorhees, 2001 ; Dang *et al.*, 2007), la détection de plagiat (Potthast *et al.*, 2010) ou encore la traduction utilisant l'alignement bilingue de documents (Zou *et al.*, 2013).

Dans nos recherches, nous nous intéressons à la mise en correspondance de documents hétérogènes. Il s'agit de développer des modèles de représentation et des méthodes dédiées à la recherche de similarité entre les documents selon différentes dimensions : la thématique, la spatialité et la temporalité. Les contributions sont nombreuses tels que : la découverte de connaissances, la mise en relation entre des producteurs de données, la cartographie de corpus, etc.

Dans cet article, nous présentons Gemedoc, un outil permettant d'annoter la similarité inter-document pour un corpus. Il est accompagné d'un protocole d'annotation fondé sur la similarité entre documents selon deux dimensions : la thématique et la spatialité. À termes, les résultats récoltés à travers diverses annotations doivent permettre de construire un corpus destiné à l'évaluation des méthodes de mise en correspondance de documents.

1. Évaluer la similarité entre deux documents

L'établissement d'un protocole d'annotation de la similarité entre deux documents selon une dimension, est difficile à définir. Il s'agit de trouver un équilibre entre deux extrêmes. L'un, qui est de définir strictement le processus d'annotation au risque de biaiser les résultats. L'autre, qui est de laisser l'affect de l'utilisateur inférer sur l'annotation et la rendre inutilisable.

Par conséquent, nous avons choisi de définir un protocole d'annotation simple, en laissant l'utilisateur libre sur la base de comparaison des documents, tout en lui donnant peu d'indices.

1.1. Modalités d'annotation

Pour annoter la similarité entre deux documents, nous avons choisi une échelle avec 4 degrés de similarités :

– **Ne sais pas.** L'annotateur ne sait pas évaluer la similarité entre les deux documents.

– **Différent.** L'annotateur indique que les documents n'ont (ou presque) rien en commun.

– **Similaire.** L'annotateur indique que les documents partagent quelques similarités.

– **Très similaire.** L'annotateur indique que les documents sont presque identiques.

1.2. Procédure d'annotation

Dans cette partie, nous illustrons la procédure d'annotation à l'aide des deux textes [1] ci-dessous. Ces deux documents traitent de la situation des migrants à Idomeni, en Grèce : l'un est un résumé, l'autre est un témoignage d'une infirmière présente sur les lieux.

Texte 1 The winding road across the wheat fields near the Greek village of Idomeni is full of people carrying large bags on their shoulders, babies in their arms and putting one step in front of the other. The stream of humanity continues day and night but not an average of 150 a day, (and only Syrians and the Iraqis who are lucky enough to have a passport or ID card from their home country) can continue the journey out of this place and across the border into the Former Yugoslav Republic of Macedonia (FYROM) and onwards to western and northern Europe. Few are leaving but more, many more keep coming, only to end up getting stranded in what is becoming unsustainable humanitarian situation. Today, in a transit camp that has the capacity to host 1,500 people, there are more than 11,000 crammed in trapped without information, in a mix of anxiety and delusion.

1. Source : Médecins Sans Frontières 2017

Texte 2 Daniela, an MSF nurse in Idomeni sums it up "there is confusion, stress. Lack of reliable information. There is a growing feeling of anger. Many refugees have been waiting here for over ten days. People are extremely exhausted." In the clinic that MSF operates in Idomeni, whole families, pregnant woman and kids arrive in a constant stream, as do many disabled people and elderly people suffering from chronic diseases. People, including babies and the elderly, are forced to sleep out in the cold, with just with a sleeping bag to keep them warm. The big tents made available by MSF have been full for days, and hundreds of small tents, are spread everywhere, even next to the train track. Omar, 24 years old, a Palestinian refugee from Homs camp in Syria is exhausted "This is making me very nervous, I don't know what is coming next. This waiting is killing me. We feel ignored here."

Comme énoncé précédemment, la similarité entre ces deux documents est évaluée selon deux dimensions : la thématique et la spatialité.

Similarité thématique
La similarité thématique est souvent perçu à travers le vocabulaire utilisé. Dans les deux exemples ci-dessus, les deux textes partagent de nombreuses thématiques communes telles que : l'aide humanitaire, la migration, les épreuves subites et la famille. Par conséquent, thématiquement, on considère que ces deux textes sont très similaires.

Similarité spatiale
Contrairement à la similarité thématique, la similarité spatiale peut dépendre de plusieurs facteurs qui varient selon ce qu'on cherche. Une première approche serait de comparer le contexte générale des deux textes, ici Idomeni. Une deuxième approche, consisterait à comparer les entités spatiales identifiées (Syrie, Idomenie, Macédoine, etc.). Enfin, dans l'étude de textes migratoires, une troisième approche comparerait les deux documents en se focalisant sur la similarité des parcours des individus.

Si l'on fonde notre raisonnement selon la première approche, les deux textes sont très similaires. Tandis que la deuxième approche fait ressortir quelques différences. Par conséquent, dans leur spatialité, ces deux textes sont partiellement similaires.

Gemedoc permet de capturer ces différences de similarités entre deux documents, selon la dimension étudiée.

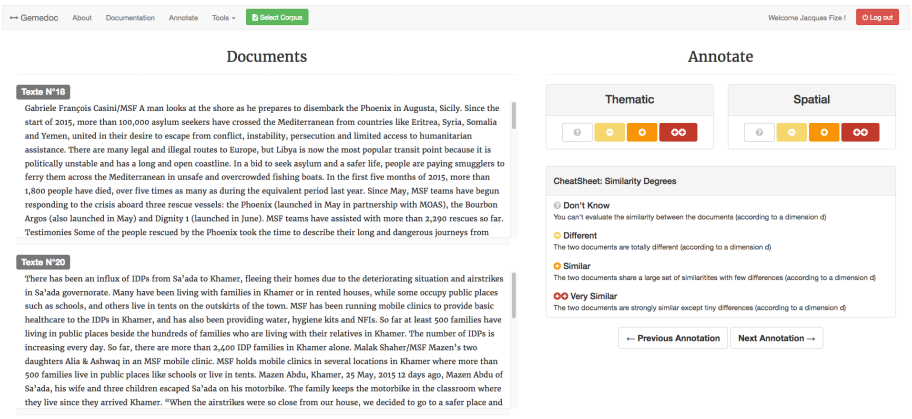

FIGURE *1. Aperçu de l'interface de* GEMEDOC

1.3. Corpus utilisés

Notre objectif étant de réaliser une mise en correspondance selon différentes dimensions (thématique, spatialité), nous récoltons divers corpus où ces deux dimensions sont exploitées. Pour cause, si la thématique est intrinsèque à tous documents, ce n'est pas forcément le cas de la spatialité.

2. Un outil dédié : Gemedoc

Afin d'annoter la similarité entre les documents d'un même corpus, nous avons décidé de développer un outil dédié : GEMEDOC. GEMEDOC est une application Web fonctionnant à l'aide d'un programme Python utilisant le module Flask [2]. Nous avons choisi ce format pour faciliter la mise en place de l'outil que cela soit :

– dans sa conception : interface en HTML5/CSS ;

– dans sa publication : hébergé donc aucune installation nécessaire pour l'annotateur.

La Figure 1 montre l'interface principale de GEMEDOC et ses différentes composantes.

2. http://flask.pocoo.org/

3. Conclusion

À travers l'atelier EXCES, nous souhaitons effectuer une première campagne d'annotation, et ainsi profiter de la présence d'un public expert dans le traitement de la spatialité. Nous souhaitons mettre à disposition différents corpus de textes à plusieurs groupes. À l'issue de l'atelier, nous collecterons les résultats de chaque groupe, à partir desquelles nous identifierons les convergences sur la similarité entre les documents. Puis, une fois les résultats obtenus, nous évaluerons la pertinence de notre modèle d'évaluation et les possibles améliorations.

Une fois les différents corpus annotés et le modèle d'annotation fixé, nous envisageons de fusionner les résultats au sein d'un unique corpus. Une fois le corpus généré, ce dernier nous permettra d'évaluer les représentations et les mesures de similarité destinées à la mise en correspondance de documents hétérogènes.

Bibliographie

Dalton J., Dietz L., Allan J. (2014). Entity query feature expansion using knowledge base links. In *Proceedings of the 37th international acm sigir conference on research & development in information retrieval*, p. 365–374.

Dang H. T., Kelly D., Lin J. J. (2007). Overview of the trec 2007 question answering track. In *Trec*, vol. 7, p. 63.

Potthast M., Stein B., Barrón-Cedeño A., Rosso P. (2010). An evaluation framework for plagiarism detection. In *Proceedings of the 23rd international conference on computational linguistics: Posters*, p. 997–1005.

Voorhees E. M. (2001). The trec question answering track. *Natural Language Engineering*, vol. 7, n° 4, p. 361–378.

Voorhees E. M. *et al.* (1999). The trec-8 question answering track report. In *Trec*, vol. 99, p. 77–82.

Xu J., Croft W. B. (1996). Quary expansion using local and global document analysis. In *Acm sigir forum*, vol. 51, p. 168–175.

Zou W. Y., Socher R., Cer D., Manning C. D. (2013). Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, p. 1393–1398.

Comment les hôtes et clients d'Airbnb parlent-ils des lieux ? Une analyse exploratoire à partir du cas parisien

Marianne Guérois[1] et Malika Madelin[2]

1. CIST Axe Information territoriale locale, Université Paris Diderot, UMR CNRS 8504 Géographie-cités
8 pl. Paul Ricœur, 75013 Paris, France
marianne.guerois@univ-paris-diderot.fr

2. CIST Axe Information territoriale locale, Université Paris Diderot, UMR CNRS 8586 PRODIG
8 pl. Paul Ricœur, 75013 Paris, France
malika.madelin@univ-paris-diderot.fr

RESUME. Cette présentation, proposée dans le cadre de l'atelier EXtraction de Connaissances à partir de donnéEs Spatialisées (EXCES), porte sur l'analyse de l'information géographique des données textuelles dans les annonces et commentaires des locations Airbnb (corpus sur Paris, avril 2017). La démarche repose sur l'identification et l'extraction des mots relatifs aux localisations dans le titre, le descriptif et les commentaires des biens, mais aussi sur le croisement de cette information avec les coordonnées géographiques des biens.
MC : analyse textuelle, information géographique, Airbnb

MOTS-CLES : analyse textuelle, information géographique, locations Airbnb, Paris.

ABSTRACT. This paper proposed for the EXCES workshop (EXtraction de Connaissances à partir de donnéEs Spatialisées) is about the analysis of the geographical information lying in textual data related to Airbnb rentals listings and comments (Paris corpus, April 2017). The methodology presented relies on the identification and the extraction of the words related to location, either in the title or in the description and the comments of the rentals. It is also based on the intersection of such information and the spatial coordinates of the rentals.

KEYWORDS: textual analysis, geographical information, Airbnb rentals, Paris.

Les grandes villes sont un terrain privilégié de l'analyse de l'extraction et du traitement de ressources numériques géoréférencées. L'information territoriale locale y est particulièrement riche et les évolutions de la production de cette information soulèvent de nombreuses questions méthodologiques liées à la représentation et au croisement de données très hétérogènes de par leurs sources, leurs formats comme leurs types d'implantation géographique. Le projet Grandes métropoles du CIST (http://www.gis-cist.fr/axes-scientifiques/projet-grandes-metropoles/) a été initié afin d'explorer les enjeux liés à la manipulation de données locales de diverses natures (données médiatiques, données issues des plates-formes du web, données de mobilité…) et d'en extraire des analyses comparables d'une métropole à l'autre. Il vise ainsi à constituer une plate-forme d'échanges autour des défis théoriques et méthodologiques soulevés par le croisement de données locales de plus en plus foisonnantes (ouverture des données publiques, multiplication de données localisées *via* les plates-formes Web et les médias sociaux, diffusion de capteurs individuels – de pollution, de température…).

Parmi les données explorées dans le cadre de ce projet, les données Airbnb issues des plates-formes internet de location touristique entre particuliers offrent un exemple riche de traces numériques du Web 2.0 détournées en sources d'information géographique pour l'analyse des dynamiques métropolitaines. Les enjeux scientifiques, voire politiques, liés à ces informations sont en effet nombreux et ont déjà suscité plusieurs études portant principalement sur la pression foncière et le tourisme (par exemple Gutierrez *et al*, 2017 ; Quattrone *et al.*, 2016 ; travaux du groupe de recherche Net(h)no-graphies[1]). D'un point de vue méthodologique, l'intérêt suscité par ces données tient tout d'abord à leur exhaustivité (a priori tous les biens de ce marché sont renseignés et localisés), à leur haute résolution spatiale (même si la localisation est « floutée » à 150 m près) et temporelle (exploitable pour peu qu'un archivage de *web scraping* permette d'en garder la mémoire), à leur richesse et diversité sémantique (description du bien, prix, commentaires…). De plus, dans un contexte où l'accès aux données immobilières localisées à l'adresse s'avère en général très coûteux, les données Airbnb sont facilement accessibles et peuvent être extraites soit directement *via* l'extraction des données de la plate-forme, soit par l'intermédiaire du site indépendant Inside Airbnb (insideairbnb.com) qui facilite le téléchargement d'un grand nombre d'informations pour les centres-villes d'une quarantaine de métropoles dans le monde, avec parfois plusieurs enregistrements temporels.

Dans le prolongement des travaux menés dans le cadre du projet Grandes métropoles[2] (Giraud *et al.*, 2017), on s'intéressera dans cet atelier à l'analyse de l'information géographique des données textuelles relatives aux biens loués. Les données textuelles de localisation sont présentes à plusieurs niveaux de l'information extraite du site Airbnb, aussi bien dans les descriptifs rédigés par les hôtes que dans les commentaires laissés par les clients. L'analyse de ces données,

1. http://netnographies.com/
2. Plus spécifiquement sur l'analyse des données Airbnb : http://www.gis-cist.fr/wp-content/uploads/2016/07/cist-gdes_metropoles-seminaire-20160517-airbnb-ppt.pdf et http://www.gis-cist.fr/analysis-of-geographical-information-in-textual-data/

croisées avec les attributs X,Y de localisation des biens, présente plusieurs intérêts. On peut se demander d'une part si la localisation des biens décrite par les hôtes est en cohérence avec la localisation spatiale de l'hébergement et avec les commentaires des clients. Au-delà de la mention de noms de lieux touristiques, cette description peut témoigner de relations spatiales, telles que la situation au centre de Paris, l'appartenance à un quartier, la proximité à des commerces, stations de métro, etc. Peut-on détecter, à partir de ces commentaires, quelles sont les propriétés des localisations recherchées ou évitées par les clients ? Sont-ils conformes à la manière dont les hôtes les décrivent ? D'autres questionnements peuvent porter sur les langues utilisées dans les commentaires, ce qu'elles reflétent de l'internationalisation du service et de l'émergence de certaines communautés de langue autour de cette activité.

La démarche adoptée est empirique et s'appuie sur un large corpus de données quantitatives et textuelles décrivant les biens sur Paris (avril 2017), données issues du site InsideAirbnb. On s'intéressera dans cet atelier spécifiquement à quatre champs distincts de la base : les coordonnées géographiques du bien, le titre (court) de l'annonce, la description détaillée de la location (ces deux dernières informations étant données par les hôtes), ainsi que les commentaires (rédigés par les clients). En volume, les descriptions renseignent plus de 52 700 locations et les hôtes utilisent globalement la place accordée pour ce champ (soit 1 000 caractères ; médiane 937). Près des trois quarts de ces locations sont ensuite décrits par plus de 660 000 commentaires (de 1 à 8 575 caractères). Dans cet atelier, on décrira la démarche ici adoptée, qui part d'une critique des données, puis d'une extraction des mots relatifs aux localisations pour s'interroger dans un second temps sur les possibilités d'automatisation de l'analyse des discours sur les lieux dans la base Airbnb. Ces informations peuvent être croisées avec les coordonnées en XY du bien mis en location, d'où la possibilité de cartographier précisément les occurrences des termes qui décrivent les localisations dans les discours. La présentation de la démarche adoptée s'accompagnera de la présentation de la chaîne de traitements construite sous R pour pouvoir transposer la démarche à d'autres villes. Les enjeux méthodologiques concernent aussi bien l'identification de l'information géographique contenue dans ces corpus de texte que l'intérêt du croisement de cette information avec les coordonnées des biens.

Une première piste d'analyse sera montrée, elle porte sur l'identification même des lieux cités dans les textes.

Après avoir préparé le corpus (minuscule ; suppression de la ponctuation, des mots vides et de certains mots communs), une analyse textuelle élémentaire peut tout d'abord être conduite à partir du découpage en mots du texte recensé dans les titres d'annonce : si cela n'apparaît pas de manière systématique, il est fréquent que le titre mentionne un nom de ville, de quartier, de rue, de site touristique ou d'un autre lieu remarquable. Au-delà du champ lexical relatif aux caractéristiques propres de la location (notamment au type de logement), l'analyse de la fréquence des termes rencontrés met en exergue l'importance des mots liés à la localisation des biens : par exemple, « Paris » ressort sans surprise du classement, en représentant près de 25% des mots contenus dans l'ensemble des titres, soit la plus forte

occurrence rencontrée. Mais d'autres mots relatifs aux lieux arrivent en tête de ce classement (montmartre, eiffel, cœur, …).

Cette analyse exploratoire permet d'extraire l'information qu'on qualifiera de géographique et de constituer un premier index des lieux géographiques et des termes relatifs à la localisation. Lorsqu'on isole et qu'on extrait ces termes dans nos corpus, la cartographie des occurrences par arrondissement souligne la hiérarchie des lieux dans la ville (notamment des lieux touristiques). Cette hiérarchie peut être appréciée non seulement en fonction du nombre d'occurrences d'un terme ou d'un groupe de termes, mais aussi en fonction de son « rayonnement » spatial, au-delà du quartier en question. Cette comparaison du rayonnement peut s'appuyer sur une cartographie ponctuelle des locations décrites à l'aide du terme, complétées par une analyse statistique de la variation des occurrences en fonction de la distance à ce lieu remarquable (par exemple en prenant « Eiffel », on calcule la proportion des locations avec « Eiffel » en fonction de la distance à la tour Eiffel). Deux difficultés méthodologiques principales accompagnent cette étape : la première vient de la définition des lieux et termes géographiques ; une autre difficulté tient à la manière dont la situation géographique est formulée à partir d'un nom de lieu : le terme « Eiffel » peut correspondre aussi bien à une localisation à proximité de la tour ou d'une vue sur la tour, depuis la location, ce qui n'a pas la même signification en termes de valorisation du bien.

Enfin, on peut se demander dans quelle mesure les termes choisis pour nommer les lieux correspondent à la localisation « exacte » des biens. Dans le même ordre d'idées, on peut faire l'hypothèse que l'analyse comparée des mots géographiques présents dans le contenu du descriptif et des commentaires (par exemple à travers des indicateurs tels que le nombre de fois où un terme « géographique » est repris dans le commentaire) permet de faire ressortir les textes où la description de la localisation des biens a été le plus discutée.

Une deuxième approche s'intéresse aux attributs des localisations, à la manière dont ces lieux sont qualifiés dans les textes. Cela peut renvoyer d'une part à une appréciation globale, *via* par exemple l'analyse de sentiments ou encore l'analyse de la localisation des mots mentionnant l'argument de centralité ou de situation idéale. Il peut être intéressant par ailleurs d'identifier les attributs les plus fréquents (calme, animé, pittoresque …) et là encore de confronter le contenu des descriptifs et des commentaires.

Bibliographie

Giraud T., Grasland C., Guérois M., Madelin M., Severo M. (2017). Données massives et information géographique, *La lettre de l'InSHS*, 45, p. 25-28.

Gutierrez J., Garcia-Palomares J. C., Romanillos G., Salas Olmedo M. H. (2017). The eruption of Airbnb in tourist cities: Comparing spatial patterns of hotels and peer-to-peer accommodation in Barcelona, *Tourism Management*, 62, p. 278-291.

Quattrone G., Proserpio D., Quercia D., Capra L., Musolesi M. (2016). Who benefits from the « sharing » economy of AirBnb ?, *Proceedings of the 25th International Conference on World Wide Web 2016*, Montréal, Canada.

Ouellet P. (2001). Sémiotique de l'empathie. L'expérience esthétique de l'autre, *Actes du colloque Sémio 2001*, Pulim, Limoges.

Twitter comme corpus numérique d'analyse des représentations territoriales Application au Parc national des Calanques de Marseille Cassis La Ciotat

Siqi Fan[1], Philippe Deboudt[2], Amel Fraisse[3], Eric Kergosien[3]

1. Université d'Orléans

2. Laboratoire TVES
EA 4477
Université de Lille
philippe.deboudt@univ-lille1.fr

3. Laboratoire GERiiCO
EA 4073
Université de Lille
prenom.nom@univ-lille3.fr

RESUME. À partir du terrain constitué par le Parc national des Calanques, cette communication présente les objectifs, la méthodologie et les premiers résultats d'un projet de recherche interdisciplinaire (géographie, sciences de l'information et de la communication, informatique) soutenu par le LABEX DRIIHM-CNRS et OHM Littoral Méditerranéen. Deux objectifs sont recherchés : (1) constituer une base de données de contenus numériques concernant les relations entre le Parc national et les populations ; (2) analyser la contribution des réseaux numériques au rapprochement entre les enjeux écologiques de préservation de la biodiversité, du patrimoine naturel et les enjeux sociaux d'accessibilité et de développement des usages. Dans cette communication, nous présentons l'élaboration d'une méthodologie semi-automatisée visant à identifier et analyser les descripteurs permettant d'identifier et de comprendre les enjeux du territoire du Parc national des Calanques à partir de Twitter.

ABSTRACT. This paper presents the objectives, methodology and initial results of an interdisciplinary research project (geography, information and communication sciences) based on the site of the Calanques National Park. This project is founded by the LABEX DRIIHM-CNRS and the OHM Littoral Mediterraneen. Two objectives are sought: (1) to establish a database of digital contents concerning the relations between the national park and the populations, and (2) to analyze the contribution of digital networks to the rapprochement between the ecological stakes of preserving biodiversity, the natural heritage and the social stakes (accessibility and development) of uses in nature spaces. In this paper, we present a semi-automatic methodology to identify and analyze descriptors related to the territory of the Calanques National Park from the Twitter social network.

MOTS-CLÉS : Twitter, représentation territoriale, fouille de textes, participation, gouvernance, Parc national des Calanques.

KEYWORDS: Twitter, territorial sciences, text mining, participation, governance, Calanques National Park

1. Introduction

Dans cette communication, nous présentons l'élaboration d'une méthodologie semi-automatisée visant à identifier et analyser les descripteurs permettant d'identifier et de comprendre les enjeux du territoire du Parc national des Calanques à partir du réseau social Twitter. La notion de territoire fait référence à différents concepts tels que les informations spatiales et temporelles, les acteurs, les opinions, l'histoire, la politique, etc. Dans le cadre de ces travaux, nous nous focalisons sur la détection d'entités nommées (EN) de type acteurs, lieu (que l'on nomme entité spatiale), temporel et thématique. Plus précisément, nous proposons une approche interdisciplinaire mobilisant l'expertise thématique issue de la géographie à une approche de fouille de textes issue des sciences de l'information et de la communication pour extraire les descripteurs territoriaux. Nous proposons des premiers éléments de réponse aux questions suivantes : Quels sont les acteurs qui s'expriment sur les lieux/sujets en lien avec le Parc national des Calanques ? Quelles sont les relations entre ces acteurs ? Quelles sont les évolutions observées selon différentes temporalités ?

2. Travaux connexes

2.1. Le projet de Parc national des Calanques

De 2008 à 2011, des recherches interdisciplinaires en sciences sociales (sociologie, géographie, aménagement-urbanisme) ont analysé le processus de construction territoriale du Parc national des Calanques de Marseille-Cassis-La Ciotat et la concertation mise en œuvre pour réaliser la charte du parc national (Deldrève et Deboudt, 2012). Le processus de création du Parc national des Calanques a débuté en 2007. Nos résultats de recherche démontrent la difficulté d'articuler les enjeux globaux et locaux dans la construction d'un tel projet de territoire qui a principalement bénéficié aux usagers traditionnels et s'est accompagnée d'une exclusion des enjeux urbains et maritimes.

Nous avons souhaité réinterroger ces résultats (principalement obtenus à partir de méthodologies combinant des enquêtes par entretiens semi-directifs et de l'observation participante) en mobilisant des corpus de données numériques issues des réseaux sociaux (twitter). Est-ce que, depuis la création du Parc national en 2012, des acteurs se sont mobilisés ou exprimés sur les réseaux sociaux à propos du Parc national des Calanques ? Les événements récents associés aux rejets de résidus en mer (boues rouges), dans le périmètre du parc national, par un site industriel de production d'alumine calcinée (ALTEO) à Gardanne, montrent la difficulté d'articuler dans un projet de territoire des enjeux globaux avec des enjeux locaux, et sont susceptibles de provoquer des prises de paroles sur les réseaux sociaux. Est-ce que les discours, projets portés dans ces réseaux alimentent, rejoignent ou s'opposent aux projets inscrits dans les agendas politiques ou développés dans l'espace public

physique ?

Pour répondre à ces questions, il est important de pouvoir identifier et valider les descripteurs permettant de décrire le territoire d'études, à savoir les entités nommées de type acteurs, lieux, temporalités, et thématiques.

2.2. Extraction des entités nommées

Les Entités Nommées (EN) ont été définies comme des noms de personnes, des lieux et des organisations lors des campagnes d'évaluations américaines appelées MUC (Message Understanding Conferences), qui furent organisées dans les années 90. Un premier défi consiste à reconnaître dans les tweets les entités nommées (EN) de type lieu, organisation et date. De nombreuses méthodes permettent de reconnaître les EN à partir de textes, et parmi ces méthodes, les approches statistiques étudient généralement les termes co-occurrents par analyse de leur distribution dans un corpus ou par des mesures calculant la probabilité d'occurrence d'un ensemble de termes. Ces méthodes ne permettent pas toujours de qualifier des termes comme étant des EN, notamment les EN de type ES ou acteurs. Des méthodes de fouille de données fondées sur l'extraction de motifs permettent de déterminer des règles (appelées règles de transduction) afin de repérer les EN. Ces règles utilisent des informations syntaxiques propres aux phrases. Des approches récentes s'appuient sur le Web pour établir des liens entre des entités et leur type (ou catégorie). Par exemple, l'approche de repose sur le principe que les distributions de probabilités d'apparition des mots dans les pages associées à une entité donnée sont proches des distributions relatives aux types. Globalement, les relations peuvent être identifiées par des calculs de similarité entre leurs contextes syntaxiques, par prédiction à l'aide de réseaux bayésiens, par des techniques de fouille de textes ou encore par inférence de connaissances à l'aide d'algorithmes d'apprentissage. Ces méthodes sont efficaces, mais elles ne sont pas adaptées aux particularités des corpus de tweets, et notamment au langage utilisé contenant des abréviations, des fautes, et souvent peu de structure syntaxique. Enfin, pour la reconnaissance des classes d'EN, de nombreuses approches s'appuient sur des méthodes d'apprentissage supervisé. Ces méthodes d'apprentissage comme les SVM ou encore les champs aléatoires conditionnels notés CRF sont souvent utilisées dans le challenge Conference on Natural Language Learning (CoNLL). Les algorithmes exploitent divers descripteurs ainsi que des données expertisées/ étiquetées. Les types de descripteurs utilisés sont par exemple les positions des termes, les étiquettes grammaticales, les informations lexicales (par exemple, majuscules/minuscules), les affixes, l'ensemble des mots dans une fenêtre autour du candidat, etc. Bien qu'intéressantes, ce type d'approche nécessite un travail manuel d'étiquetage important que nous ne pouvons appliquer dans le cadre de ces travaux.

Dans cette recherche, nous proposons une méthode combinant une approche statistique à une approche de fouille de textes (Pak et al., 2014 ; Zenasni et al., 2016).

2.3. Twitter un corpus pour la construction et l'extraction de connaissances

Dans le domaine de traitement automatique de langues, plusieurs travaux de recherche utilisent Twitter comme corpus pour construire des ressources linguistiques et extraire des connaissances pertinentes. (Read, 2005 ; Pak & Paroubek, 2010) ont utilisé les émoticônes comme marqueur de polarité pour distinguer les textes positifs et négatifs depuis les tweets. Dans un premier temps, ils ont identifié une liste d'émoticônes positives (:), :-), :-D, etc.) et une liste d'émoticônes négatives (:(, :-(, etc). Ensuite, les deux listes ont été utilisées comme critère de recherche pour récupérer des messages positifs et négatifs depuis Twitter.

Dans des travaux plus récents (Mohammad, 2012 ; Qadira & Riloffe, 2013 ; Fraisse & Paroubek, 2014) ont utilisé une liste de mot-dièses (hashtag en anglais) (#sad, #happy, #angry, #fear, #anxious, #disappointed, #unhappy, etc.) pour collecter des corpus émotionnels et construire de façon automatique des lexiques affectifs. Les lexiques ont été ensuite utilisés dans des tâches de détection automatique des émotions.

3. Une méthode pour la construction d'une représentation territoriale

Pour la collecte du corpus, deux périodes temporelles ont été choisies :

– la période 2007-2011 correspondant à celle du processus de création du Parc national des Calanques et de l'organisation du processus de concertation pour élaborer la charte du Parc national. C'est durant cette période que se sont manifestés des groupes d'acteurs favorables ou opposés à la création d'un Parc national dans le territoire des Calanques ;

– la période 2012-2017 intégrant l'année de création du Parc national (en 2012) jusqu'aux années récentes marquées par plusieurs conflits d'usages et notamment celui fortement médiatisé, provoqué par des rejets de boues rouges issues de la production d'alumine par l'usine ALTEO de Gardanne, dans les espaces du cœur maritime du Parc national.

Avant de lancer la collecte automatique du corpus de tweets, nous avons identifié manuellement, avec l'expertise de géographes spécialistes du territoire d'études, une liste d'acteurs qui ont participé au processus de création du Parc national des Calanques à Marseille. Il s'agit essentiellement d'associations, de conseils de quartiers, de personnalités politiques, et de citoyens. Nous avons aussi identifié une liste de mots clés que nous avons utilisés sous forme de hashtags pour collecter les tweets comme par exemple #PNCalanques, #ParcNationalDesCalanques, #BouesRouges, etc. (cf. figure 1). Nous avons utilisé l'API Search[1] de Twitter pour collecter et filtrer les messages. L'Api permet de spécifier la langue de messages et une requête de recherche par mot clé. Ainsi, pour chaque hashtag h du tableau 1, nous collectons un certain nombre de tweets qui contiennent le hashtag h. Au total,

[1] https ://dev.twitter.com/docs/api/1/get/search

sur la période 2007-2011 (avant la création du parc national), nous avons collecté 5 000 tweets et pour la période 2012-aujourd'hui 2 900 tweets furent collectés.

TABLE 1. Extrait de la liste de Hashtags utilisés pour la construction de corpus de Tweets

Hashtags	**Description**
#ParcNationalCalanques	Parc National des Calanques
#PNCal	Parc National des Calanques
#BouesRouges	Boues Rouges
#Marseille	Marseille
#CreationPNCalanques	Création du Parc National des Calanques

Avec l'aide des experts du domaine nous avons analysé dans un premier temps le corpus collecté afin de vérifier la pertinence des hashtags utilisés pour la collecte des tweets. Ainsi, suite à cette analyse, nous avons modifié et rajouté certains hashtags et relancer le processus de collecte.

TABLE 2. Exemples de tweets extraits du corpus collecté

Exemples de tweets collectés
@cestrosi @MetropoleNCA **#stephaneBouillon** préfet #paca qui autorise la pollution dans la méditerranée de l'entreprise #altéo **#bouesrouge**
#ComitéSantéLittoral Sud (Marseille): Bulletin n°1 http://comitesantelittoralsud.blogspot.com/2014/01/bulletin-n1.html?spref=tw …
#marseille: Parc national des Calanques : les élus face à l'enquête publique http://bit.ly/sxqeOx
la création du **#PNCalanques** aura permis de préserver l'espèce des avocats et juristes en tout genre ! http://www.marsactu.fr/environnement/parc-des-calanques-un-recours-depose-contre-lelection-de-danielle-milon-30167.html …

En se basant sur une approche statistique, notre méthode consiste à extraire, à partir du corpus collecté, et pour chaque hashtags h, l'ensemble de mots qui lui est associé. Ces mots peuvent être des noms d'acteurs, des mots simples ou des noms de lieux. En effet, nous considérerons si un mot m est fortement corrélé à un hashtag h de notre liste alors ce mot est pertinent pour notre analyse de relations territoriales.

Afin de mesurer l'association entre un mot m du corpus et un hashtag h, nous nous sommes basés sur l'information mutuelle introduite par (Fano, 1961) qui, pour chaque couple de variables aléatoires (X, Y), mesure leur degré de dépendance au sens probabiliste. L'information mutuelle est donnée par l'équation 1.

$$\boldsymbol{IM(X,Y) = \log_2(\frac{P(X,Y)}{P(X).P(Y)})} \quad (1)$$

Ainsi, dans notre cas, il s'agit de mesurer le degré de dépendance entre un hashtag *h* et un mot *m* (*cf.* équation 2).

$$\boldsymbol{IM(h,m) = \log_2(\frac{freq(h,m)}{freq(h).freq(m)})} \quad (2)$$

Avec $freq(h,m)$ est le rapport entre le nombre de tweets contenant le mot *m* et le hashtag *h* ($|T_{h,m}|$) et le nombre total de tweets (|T|) (*cf.* équation 3).

$$\boldsymbol{freq(h,m) = \frac{|T_{h,m}|}{|T|}} \quad (3)$$

$freq(h)$ est le rapport entre le nombre total de tweets contenant le hashtag *h* ($|T_h|$) et le nombre total de tweets (*cf.* équation 4).

$$\boldsymbol{freq(h) = \frac{|T_h|}{|T|}} \quad (4)$$

$freq(m)$ est le rapport entre le nombre total de tweets contenant le hashtag *h* ($|T_m|$) et le nombre total de tweets (*cf.* équation 5).

$$\boldsymbol{freq(m) = \frac{|T_m|}{|T|}} \quad (5)$$

4. Expérimentations et résultats

Les messages Twitter peuvent contenir : des URLs, des mentions utilisateurs (les acteurs dans notre cas, par exemple, @MairiedeMarseille), des retweets, etc. Ainsi, avant de procéder à l'extraction des hashtags, des noms d'acteurs et des lieux cités dans les tweets, nous avons tout d'abord effectué certaines opérations de prétraitements automatiques : (1) suppression des liens URLs et les retweets, (2) segmentation et (3) suppressions des mots outils. Ensuite, nous avons calculé pour chaque mot *m* du corpus sa corrélation avec le hashtag qui a été utilisé dans la requête. En fonction de leurs degrés d'association, les mots sont ensuite ordonnés par ordre croissant : du plus pertinent au moins pertinent.

Afin de mieux visualiser les résultats obtenus, nous avons fait appel à un expert du domaine pour fixer un seuil en dessous duquel nous considérons qu'un mot m n'est pas pertinent dans notre analyse.

Nous avons procédé ensuite à la visualisation de ces résultats via des graphes : les nœuds représentent les mots, les hashtags et les noms d'acteurs dans le corpus et les arcs sont soit :

- Les relations entre les mots (mot clé ou hashtag) (*cf.* Figure 1) : cette relation permet de répondre à la question *Quelles sont les thématiques et les sujets abordés dans les tweets ?* par exemple Parc national des calanques et pollution (*cf.* figure 2).
- Les relations entre un hashtag et un acteur : cette relation permet de répondre à la question *Qui parle de Quoi ?*
- Les relations entre un acteur et un acteur : cette relation permet de répondre à la question *Qui parle à Qui ?*

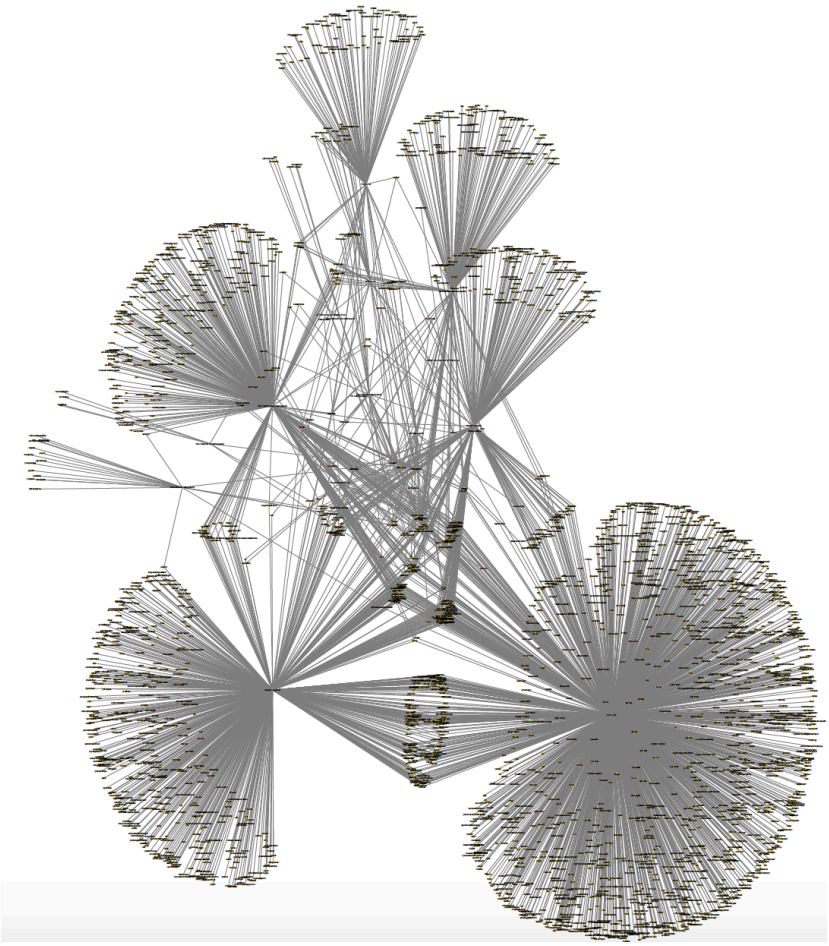

FIGURE 1. Visualisation des relations entre les mots et hashtags

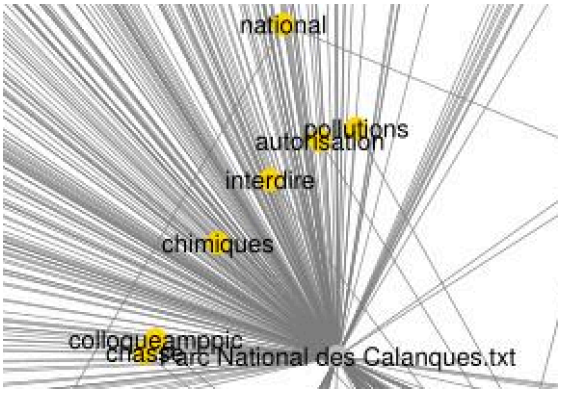

FIGURE 2. Visualisation des relations entre les mots Parc national des calanques et pollution

Les résultats sont actuellement en cours d'analyse fine en collaboration avec les experts du domaine.

4. Conclusion et perspectives

Cette communication a permis de livrer le processus de construction de la méthodologie pour créer le corpus de données numériques, identifier les connaissances pertinentes et nos premiers résultats concernant les thèmes les acteurs, leurs interactions à propos du Parc national des calanques dans le réseau social twitter. La mobilisation des données numériques de twitter permet de constituer un nouveau corpus de données numériques contemporain de la période de création du Parc national des Calanques selon une démarche rétrospective, et aussi de fournir des données numériques associées aux premières années de fonctionnement du parc national. L'interprétation fine des données récoltées, en cours de réalisation par les membres du projet, devrait permettre de compléter les analyses déjà réalisées et de produire de nouveaux résultats concernant : les interactions entre les acteurs (Berthelot *et al.,* 2016), les thématiques, les lieux, à propos du fonctionnement du parc national.

Bibliographie

Berthelot M.-A., Severo M., Kergosien E. (2016), Cartographier les acteurs d'un territoire : une approche appliquée au patrimoine industriel textile du Nord-Pas-de-Calais, *in 3ème colloque international du CIST (CIST 2016)*, pp. 6, Grenoble.

Deldrève V., Deboudt P. (coord.) (2012), Le parc national des Calanques. *Construction territoriale, concertation, usages*, Paris, QUAE, 2012, 248 p.

Fano R. (1961), *Transmission of information : A statistical theory of communications*. MIT Press, Cambridge.

Fraisse A., Paroubek P. (2014), Twitter as a Comparable Corpus to build Multilingual Affective Lexicons. *In Proceedings of the 7th International Workshop on*

Building and Using Comparable Corpora at LREC 2014 (BUCC 2014), Reykjavik, Iceland, pp. 17-21.

Mohammads M. (2012), Emotional tweets, *in Proceedings of First Joint Conference on Lexical and Computational Semantics*, pp.246–255.

Pak A., Paroubek P. (2010), Construction d'un lexique affectif pour le français à partir de twitter, *in Proceedings of TALN (Traitement Automatique des Langues Naturelles)*, Montréal, Canada.

Pak A., Paroubek P., Fraisse A., Francopoulo G. (2014), *Normalization of Term Weighting Scheme for Sentiment Analysis*. Book Chapter, Human Language technology Challenges for Computer Science and Linguistics. Series: Lecture Notes in Vetulani, Zygmunt, Mariani, Joseph (eds.), Artificial Intelligence, Springer, Vol. 8387.

Qadira. & Riloffe. (2013). Bootstrapped learning of emotion hashtags hashtags4you. *In The 4th Workshop on Computational Approaches to Subjectivity Sentiment and Social Media Analysis*, Atlanta.

Read J. (2005), Using emoticons to reduce dependency in machine learning techniques for sentiment classification, *in Proceedings of The 43rd Annual Meeting of the Association for Computational Linguistics*, pp.43–48.

Zenasni S., Kergosien E., Roche M., Teisseire M. (2016), Extracting new Spatial Entities and Relations from Short Messages, *in The 8th International ACM Conference on Management of Digital EcoSystems (MEDES'2016)*, Hendaye (France), pp. 8.

Calcul de similarité entre événements sociaux

A. Fotsoh , C. Sallaberry , A. Le Parc - Lacayrelle

LIUPPA / Université de Pau et des Pays de l'Adour
BP 1155
64013 Pau cedex
armel.fotsohtawofaing@univ-pau.fr
christian.sallaberry@univ-pau.fr
annig.lacayrelle@univ-pau.fr

RÉSUMÉ. *Nous considérons les événements comme des entités nommées complexes, récursivement composées d'entités nommées simples (personnes, lieux, organisations, dates, ...) et/ou d'entités nommées complexes. Nous avons mis au point une chaîne de traitement dédiée à l'extraction, l'indexation et la recherche d'événements sociaux dans des corpus de pages Web. C'est dans ce contexte que ce papier propose une fonction de calcul de similarité générique qui cible tout type d'entités nommées complexes.*

ABSTRACT. *We consider events such as complex named entities (people, places, organizations, dates, ...) and/or complex named entities. We have developed a processing chain dedicated to extracting, indexing and searching for social events in Web page corpuses. In this context, this paper is proposing a generic similarity computation fonction that targets any type of complex named entity.*

MOTS-CLÉS : *Calcul de similatité, entité complexe, évenement social*

KEYWORDS: *Named Entity Similarity Computation, Event Similarity Computation, Event Extraction, Information retrieval*

1. Introduction

Les événements sociaux sont des événements bien planifiés impliquant un grand nombre de participants : e.g. les événements culturels (festivals), de divertissement (concerts), éducatifs (forums), socio-économiques (expositions, salons ...), sportifs (match de football ...) ou même politique (meetings). Ils possèdent une dimension spatiale (où), une dimention temporelle (quand) et une dimension thématique (quoi, qui), et peuvent apparaître dans un catalogue d'annonces ou un site de ventes en ligne. Notre travail vise la mise à disposition

de services géolocalisés qui analysent des ressources du Web afin d'y repérer des événements sociaux, de les indexer et, enfin, de les exploiter au travers d'un système de recherche d'information géographique. Les flux d'informations à traiter comportent des redondances. En effet, un même événement social peut être décrit sur plusieurs pages Web avec ou sans différence. Par exemple, la figure 1 montre deux pages Web décrivant le même événement même si les titres ne sont pas identiques.

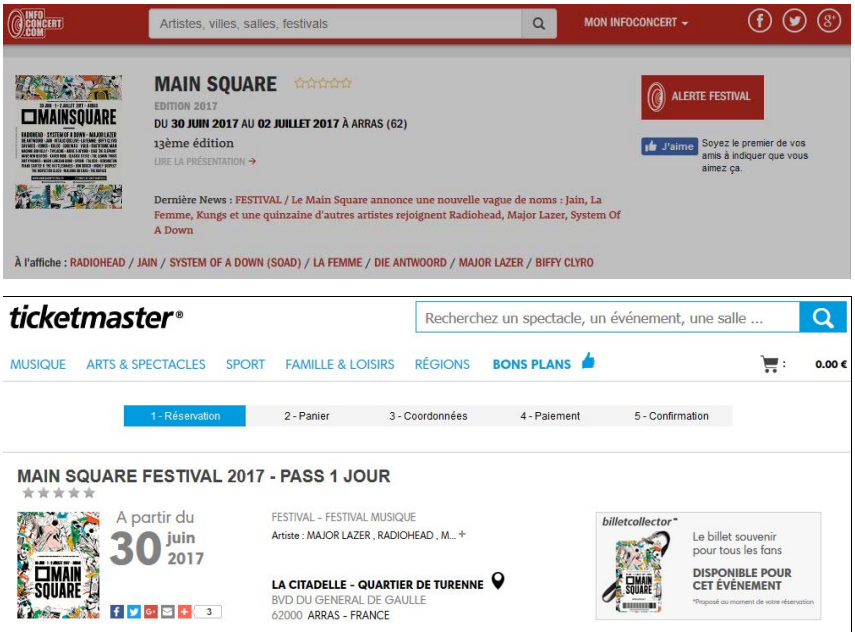

FIGURE *1. Exemple d'événements*

La chaîne de traitement que nous proposons vise la résolution de trois principaux verrous : (i) le filtrage de ressources du Web, (ii) l'identification des propriétés des événements dans des pages Web peu ou non structurées, (iii) le calcul de similarité entre deux événements. Le filtrage du Web nécessite la mise au point d'un processus qui cible des sites spécifiques et, au sein de ces sites, repère les pages Web de la catégorie "page événement". De même, l'identification des propriétés des événements est difficile du fait de l'utilisation de vocabulaires non contrôlés dans des textes peu structurés et/ou de structures différentes. De récents travaux (Foley *et al.*, 2015) proposent des approches d'extraction par apprentissage alignées sur le modèle d'événements de *schema.org*. De notre côté, nous reprenons des éléments de *schema.org* dans un nouveau modèle d'événements et proposons un processus d'extraction d'EN hybride en deux phases : (1) l'apprentissage supervisé pour le repérage des propriétés d'événements ; (2)

les patrons et ressources pour la normalisation et la correction des précédents marquages. Enfin, le calcul de similarité entre événements présente également de nombreuses difficultés : pas de standard pour la représentation des catégories d'événements ; des valeurs similaires de propriétés peuvent être exprimées de manières différentes (e.g. lieu de l'événement : nom d'une salle vs adresse). Les travaux de (Becker *et al.*, 2010 ; Wongsuphasawat *et al.*, 2012) traitent plus particulièrement d'événements de type "faits" (actualité, faits historiques) relatés dans les réseaux sociaux. Nous étendons ces travaux et proposons une approche de calcul de similarité générique qui cible tout type d'entités nommées complexes et que nous appliquons aux événements de type "sociaux" à titre d'expérimentation.

Dans le cadre de ce papier, nous allons nous focaliser sur le calcul de similarité entre deux événements à des fins (i) d'intégration sans doublon d'un nouvel événement dans un index et (ii) de recherche d'événements en vue de construire une réponse à un besoin d'information ("Local search"). La section 2 définit notre modèle d'événement. La section 3 détaille nos propositions pour le calcul de similarité entre entités nommées complexes, tandis que la section 4 détaille leur expérimentation sur un corpus d'événements sociaux. La section 5 conclut ces travaux et propose des perspectives.

2. Modèle d'événement

Contrairement aux modèles existants ((NIST, 2005),(Raimond *et al.*, 2007), (Troncy *et al.*, 2010) et (Hage *et al.*, 2011)) dans lesquels les quatre dimensions (quoi, quand, qui, où) sont au même niveau, notre modèle est construit sur deux niveaux : l'événement défini par sa dimension thématique (quoi) puis par les représentations correspondantes (les dimensions spatiale (où), agentive (qui) et temporelle (quand)). Ce choix est motivé par le fait que nous nous intéressons uniquement aux événements sociaux, et que ceux-ci peuvent se produire plusieurs fois. Par exemple, il est possible d'assister au même concert de l'artiste X à plusieurs dates dans des lieux différents. De plus, l'ensemble des artistes est rattaché à la représentation et non pas à l'événement de manière à tenir compte du fait que lors de la tournée d'une pièce de théâtre, par exemple, un acteur peut changer. Il en est de même pour l'ensemble des organisateurs. Il est important de noter que la description des personnes et des organisations, notamment, s'appuie sur des propositions de *schema.org*.

Ainsi, un événement social e possède un titre t, une description $desc$, un ensemble de catégories Cat_e et un ensemble de représentations R_e. Nous avons :

$$e =< t, desc, Cat_e, R_e >$$

Une représentation r d'un événement e ($r \in R_e$) est définie par un lieu l, une date d, une heure h, un ensemble d'agents A_r (chanteur, acteur, artiste, équipe

de sport, sportif, ...) intervenant lors de la représentation et enfin l'ensemble des organisateurs Org_r. Nous avons :

$$r =< l, d, h, A_r, Org_r >$$

3. Similarité entre entités complexes

Lors de la phase d'annotation, nous extrayons d'une page web un événement e_a (qui peut comporter une ou plusieurs représentations r_{a_i}). Lors de la phase d'indexation, un même événement ne doit pas être indexé plusieurs fois. Ainsi, 3 cas de figures peuvent se produire :

– e_a est déjà dans l'index ;

– e_a est partiellement dans l'index, c'est à dire que e_a possède, par exemple, une ou plusieurs représentations r_{a_i} qui ne sont pas encore dans l'index ;

– e_a n'est pas dans l'index.

Il est donc nécessaire de calculer la similarité à 2 niveaux : au niveau de l'événement et au niveau de la représentation. Un événement et sa représentation correspondant tous les deux à une entité complexe (un objet possédant une ou plusieurs propriétés), calculer la similarité entre deux événements ou entre deux représentations peut être formulé de la même façon.

Soient deux entités e_1 et e_2 décrites par le même ensemble de n propriétés (p_1, ..., p_n). Chaque propriété peut être monovaluée ou multivaluée et de type texte, lieu, date/heure ou concept ontologique (par exemple, la propriété *titre* est un texte et la propriété *agents* est un ensemble de textes). On a : $e_1 =< v_{11}, ..., v_{1n} >$ et $e_2 =< v_{21}, ..., v_{2n} >$, v_{1i} et v_{2i} étant respectivement la valeur pour la propriété p_i de e_1 et de e_2. Calculer la similarité entre deux entités complexes nécessite :

1. de comparer deux à deux leurs différentes propriétés afin de déterminer si leurs valeurs sont similaires ou pas. Nous avons, pour cela, défini une fonction de similarité individuelle pour chacun de ces types (métriques textuelle, sémantique et numérique), ainsi que la similarité entre ensembles (métrique pour les ensembles). Soit s_{p_i} la fonction de similarité associée à la propriété p_i;

2. de combiner les résultats précédents à l'aide d'une fonction f_{sim} pour obtenir la similarité entre e_1 et e_2.

Plusieurs fonctions f_{sim} peuvent être définies selon que l'on utilise une approche basée sur la somme pondérée ou une approche basée sur les techniques d'aide à la décision.

Pour la première approche, une première proposition pour la fonction f_{sim} est la somme pondérée des similarités des différentes propriétés : $f_{sim}(e_1, e_2) = \sum_{i=1}^{n} w_i.s_{p_i}(v_{1i}, v_{2i})$ avec w_i poids associé à la propriété p_i. Les poids w_i sont déterminés de façon empirique. Cette proposition est la plus utilisée (par ex.

dans (Becker *et al.*, 2010), (Serrano *et al.*, 2013) et (Wongsuphasawat *et al.*, 2012)) et permet à l'expert de donner plus d'importance à certaines propriétés.

Dans le cadre d'une seconde proposition, nous avons envisagé de déterminer les poids w_i en utilisant la régression linéaire. La fonction f_{sim} devant prendre des valeurs dans l'intervalle $[0;1]$, utiliser la régression linéaire aurait pu entraîner des effets de bord (valeurs négatives proches de 0 ou valeurs légèrement supérieures à 1) (Agresti, Kateri, 2011). C'est la raison pour laquelle nous avons conçu une seconde proposition basée sur la régression logistique. Dans ce cas, $f_{sim}(e_1, e_2)$ correspond à la probabilité que e_1 et e_2 soient similaires sachant X avec $X = (s_{p_1}(v_{11}, v_{21}), ..., s_{p_n}(v_{1n}, v_{2n}))$, le vecteur contenant le score de similarité pour chaque propriété de e_1 et e_2. On a donc : $f_{sim}(e_1, e_2) = P(1|X)$. Utiliser le *logit* (Aldrich, Nelson, 1984) de f_{sim} permet de travailler sur l'intervalle $]-\infty;+\infty[$ et ainsi d'éviter les effets de bord :

$$\begin{aligned} logit(f_{sim}(e_1, e_2)) &= ln(\frac{f_{sim}(e_1, e_2)}{1 - f_{sim}(e_1, e_2)}) \\ &= w_0 + w_1.s_{p_1}(v_{11}, v_{21}) + \cdots + w_n.s_{p_n}(v_{1n}, v_{2n}) \end{aligned}$$

Nous obtenons donc :

$$f_{sim}(e_1, e_2) = \frac{1}{1 + e^{-(w_0 + w_1.s_{p_1}(v_{11}, v_{21}) + \cdots + w_n.s_{p_n}(v_{1n}, v_{2n}))}}$$

Pour déterminer w_0, ..., w_n, nous construisons, à partir d'un échantillon représentatif de m couples d'entités complexes (e_1^1, e_2^1), ..., (e_1^m, e_2^m), la matrice A et le vecteur B. $A \in M_{m,n}(\mathbb{R})$ et ses coefficients $a_{i,j}$ correspondent à la similarité pour la propriété p_j du couple (e_1^i, e_2^i) : $a_{i,j} = s_{p_j}(v_{1j}^i, v_{2j}^i)$, avec $1 \leq i \leq m$ et $1 \leq j \leq n$. $B \in M_{m,1}(\mathbb{R})$ et ses coefficients b_i correspondent au score de similarité donné par un expert pour le couple (e_1^i, e_2^i). A partir de A et de B, nous utilisons la technique d'apprentissage basée sur la descente de gradiant (Friedman, Popescu, 2003) pour estimer le vecteur des poids W ($W \in M_{1,n}(\mathbb{R})$). Cette méthode incrémentale permet d'obtenir une solution sans contrainte sur l'échantillon.

Pour la deuxième approche, nous considérons que la forme de la fonction f_{sim} n'est pas connue a priori. Par conséquent, nous proposons d'utiliser une méthode d'apprentissage supervisé employée dans les systèmes de prédiction pour l'aide à la décision : les arbres de décision (Safavian, Landgrebe, 1991). Cette méthode consiste à analyser un jeu de données pour le segmenter en sous-ensembles homogènes. Ces sous-ensembles correspondent à des couples d'entités vérifiant un même groupe de conditions. Nous allons utiliser l'algorithme "tree-based regression" (Wang *et al.*, 2013) qui nous permettra de construire l'arbre de décision et d'inférer un modèle d'apprentissage permettant de déterminer la similarité entre deux entités complexes.

4. Expérimentation de f_{sim} sur les événements sociaux

Nous avons expérimenté nos différentes propositions, pour calculer la similarité sur les événements, lors de la phase d'indexation. Nous les avons ensuite comparées en calculant, pour chacune d'elles, le rappel, la précision et la f1-mesure.

4.1. Protocole

Pour déterminer si deux événements sont similaires ou pas, nous avons choisi de nous baser sur 3 propriétés : le titre t, l'ensemble des catégories Cat_e et les acteurs principaux Ap_e. Ainsi, si un événement e est défini par :

$$e =< t, desc, Cat_e, R_e > \ avec\ R_e = \{r\}\ et\ r =< l, d, h, A_r, Org_r >$$

Nous commençons par calculer e' tel que :

$$e' =< t, Cat_e, Ap_e > \ avec\ Ap_e = \{a/\forall r \in R_e, a \in A_r\}$$

Lors de la phase d'annotation, la valeur d'une ou plusieurs propriétés de l'événement peut ne pas avoir été identifiée. Ainsi, les deux événements à comparer peuvent avoir des propriétés pour lesquelles la valeur n'est pas connue. Nous avons considéré, comme condition nécessaire à la fonction f_{sim} (lors de l'indexation), que e'_1 et e'_2 possèdent une valeur pour la propriété $titre$ (t). Si ce n'est pas le cas, alors leur similarité est égale à 0 ($f_{sim}(e'_1, e'_2) = 0$).

Les trois propositions de fonctions f_{sim} à evaluer (voir section 3) sont :

1. *approche empirique* : le choix a été fait de donner plus de poids à la propriété $titre$. La combinaison de poids qui nous donne la meilleure f1-mesure est : $w_t = 0,5$, $w_{Cat} = 0,3$ et $w_{Ap} = 0,2$. Lorsque, pour une propriété, un des deux événements n'a pas de valeur, son poids est réparti équitablement entre les autres propriétés ;

2. *régression logistique* : cette approche nécessite de construire un jeu d'apprentissage. Ce dernier a été composé de 85 couples d'évenements dont 50 sont considérés comme similaires par un expert et 35 non similaires. De plus, nous utilisons un processus de validation croisée pour déterminer le taux d'apprentissage et le nombre maximum d'itération ;

3. *régression "tree-based"* : nous avons utilisé le même jeu d'apprentissage que pour la régression logistique. Le nombre maximum de noeud a été fixé, après expérimentation, à 10. En effet, à partir de 10 noeuds, la f1-mesure était stable.

L'évaluation de ces trois propositions a été réalisée sur un jeu de test (différents du jeu d'apprentissage) composé de 100 couples d'événements dont 58

sont considérés similaires par un expert et 42 non similaires. Nous comparerons nos résultats à ceux obtenus en utilisant une approche CombMNZ(Fox, Shaw, 1993) normalisée pour combiner les scores de similarité des différentes propriétés. L'approche CombMNZ, qui est une approche classique utilisée en recherche d'information multi-dimensionnelle pour la combinaison des critères de recherche, sera notre baseline.

4.2. Résultats

Nous avons tracé, figure 2, les scores de similarités pour chacun des 100 couples d'événements traités, obtenus respectivement avec l'approche CombMNZ, empirique, regression logistique et regression par arbres. Ces résultats sont comparés à ceux évalués par les experts. Ainsi, en observant ces courbes, nous constatons que la régression logistique et la régression "tree-based" semblent donner des résultats plus tranchés que CombMNZ et l'approche empirique. De plus, ces courbes nous ont permis d'estimer les seuils des scores de similarité optimaux pour chaque approche : nous avons proposé des intervalles au-delà desquels nous considérons qu'il y a similarité. Une fois ces intervalles déterminés, nous avons, pour chaque approche, calculé le rappel, la précision et la f1-mesure pour chaque valeur de seuil en la faisant varier par pas de 0,01.

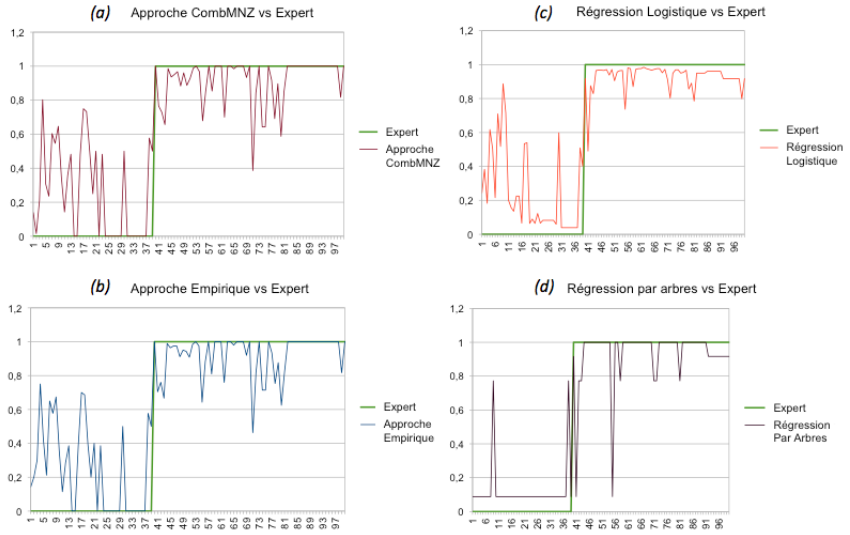

FIGURE *2. Approches CombMNZ (a), empirique (b), régression logistique (c) et régression par arbres (d)*

Le tableau 1 donne le seuil optimal calculé avec la précision, le rappel et la f1-mesure associés. Ces résultats corroborent les observations réalisées sur

les courbes. La meilleure précision est obtenue avec la régression logistique et le meilleur rappel avec l'approche empirique. Enfin, la f1-mesure étant une combinaison du rappel et de la précision, nous concluons que la régression logistique donne les meilleurs résultats.

$f_{sim}(e'_1, e'_2)$	Précision	Rappel	F1-mesure	Seuil
CombMNZ approach	86,76	96,72	91,47	0,64
Empirical approach	86,95	**98,36**	92,30	0,62
logistical regression	**90,90**	95,74	**93,25**	0,75
tree-based regression	90,62	95,08	92,79	0,77

TABLE *1. Résultats obtenus par les différentes méthodes sur tout le jeu de test*

5. Conclusion

Nos travaux concernent une chaîne de traitement dédiée à l'extraction d'événements sociaux dans des corpus de pages Web. Les événements détectés sont consolidés, enrichis et indexés à des fins de recherche d'information. L'étape d'indexation (ajout d'un événement dans l'index en évitant les doublons) et l'étape de recherche d'information (appariement événement-requête avec événements de l'index) s'appuient sur une fonction de calcul de similarité entre deux événements.

C'est dans ce contexte que nous proposons une fonction de calcul de similarité générique qui cible tout type d'entités nommées complexes : nous instancions cette fonction $f_{sim}(e'_1, e'_2)$ selon trois approches distinctes que nous avons nommées approches empirique, régression logistique et régression par arbres, respectivement. Une expérimentation menée, pour la phase d'indexation, sur 100 couples d'événements sociaux nous a permis de montrer que l'approche régression logistique se démarque clairement des autres approches. Cette approche est particulièrement efficace lorsqu'une ou plusieurs propriétés d'un événement ne sont pas renseignées. Ce qui est généralement le cas dans un contexte de recherche d'information, les caractéristiques de l'événement-requête pouvant, par exemple, se limiter à un lieu ou une zone et une date ou une période seulement. Nous allons donc maintenant mener une expérimentation similaire pour la phase de recherche d'information de façon à confirmer cette tendance.

Bibliographie

Agresti A., Kateri M. (2011). Categorical data analysis. In M. Lovric (Ed.), *International encyclopedia of statistical science*, p. 206–208. Berlin, Heidelberg, Springer Berlin Heidelberg.

Aldrich J. H., Nelson F. D. (1984). *Linear probability, logit, and probit models* (vol. 45). Sage.

Becker H., Naaman M., Gravano L. (2010). Learning similarity metrics for event identification in social media. In *Proceedings of the third acm international conference on web search and data mining*, p. 291–300.

Foley J., Bendersky M., Josifovski V. (2015). Learning to extract local events from the web. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, santiago, chile, august 9-13, 2015*, p. 423–432.

Fox E. A., Shaw J. A. (1993, février). Combination of Multiple Searches. In D. K. Harman (Ed.), *Trec-1: Proceedings of the first text retrieval conference*, p. 243–252. Gaithersburg, MD, USA.

Friedman J., Popescu B. E. (2003). *Gradient directed regularization for linear regression and classification*. Rapport technique. Citeseer.

Hage W. R. van, Malaisé V., Segers R. H., Hollink L., Schreiber G. (2011). Design and use of the simple event model (sem). *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 9, n° 2.

NIST. (2005). The ace 2005 (ace05) evaluation plan.

Raimond Y., Abdallah S. A., Sandler M., Giasson F. (2007, September 23-27). The music ontology. In *Proceedings of the 8th international conference on music information retrieval*, p. 417–422. Vienna, Austria.

Safavian S. R., Landgrebe D. (1991, May). A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 21, n° 3, p. 660-674.

Serrano L., Bouzid M., Charnois T., Brunessaux S., Grilhères B. (2013). Events extraction and aggregation for open source intelligence: From text to knowledge. In *2013 IEEE 25th international conference on tools with artificial intelligence, herndon, va, usa, november 4-6, 2013*, p. 518–523.

Troncy R., Shaw R., Hardman L. (2010). Lode: une ontologie pour représenter des événements dans le web de données. In *21ièmes journées francophones d'ingénierie des connaissances (ic 2010)*, p. 69–80.

Wang J., Chen K., Kayis E., Gallego G., Guerrero J., Wang R. *et al.* (2013). *Tree-based regression.* Google Patents. (US Patent App. 13/528,972)

Wongsuphasawat K., Plaisant C., Taieb-Maimon M., Shneiderman B. (2012). Querying event sequences by exact match or similarity search: Design and empirical evaluation. *Interacting with computers*, vol. 24, n° 2, p. 55–68.