14e édition des Journées francophones

EGC 2014

Extraction et Gestion des Connaissances

28-31 Janvier

RENNES

IRISA & Centre Inria Rennes - Bretagne Atlantique

Campus de Beaulieu, Rennes

Journées Ateliers/Tutoriels

FST : Fouille de données spatiales et temporelles

CERGEO : Construction, enrichissement et exploitation de ressources GEOgraphiques

Atelier Fouille de données Spatiales et Temporelles (FST) - Construction, enrichissement et exploitation de ressources GEOgraphiques pour l'analyse de données (CERGEO)

Organisateurs
FST :
Thomas Guyet (Agrocampus-Ouest/IRISA, Rennes),
René Quiniou (INRIA Rennes - Bretagne Atlantique/IRISA),
CerGeo :
Eric Kergosien (LIRMM, CNRS, Montpellier),
Mathieu Roche (Cirad, UMR TETIS, Montpellier)

PRÉFACE

Les ateliers «Fouille de données Spatiales et Temporelles (FST)» et «Construction, enrichissement et exploitation de ressources GEOgraphiques pour l'analyse de données (CERGEO)» ont été fusionnés pour proposer un atelier autour des thèmes de l'analyse et de la fouille des données spatiales et temporelles lors de l'édition 2014 d'EGC.

La thématique FST vise les problématiques liées à la prise en compte de l'information temporelle ou spatiale – quantitative ou qualitative – et les défis émergents dans les processus de gestion et d'extraction de ce type données. Les données visées sont principalement (mais pas exclusivement) :

Fouille de données Spatiales et Temporelles

Dans les processus d'extraction d'information, il est souvent essentiel de prendre en compte le contexte spatio-temporel dans lequel sont acquises les données pour espérer en extraire une information pertinente et utilisable. Depuis une dizaine d'année, on constate, d'une part, un développement croissant des capteurs, des réseaux de capteurs, du web des objets ou encore des réseaux sociaux qui produisent des traces de manière continue. D'autre part, les outils de géolocalisation se démocratisant, à ces données sont de plus en plus associées des informations sur leur localisation géographique et sur l'environnement spatial (par des images de télédétection ou des données géographiques de plus en plus facilement accessibles). Ces traces spatiales ou temporelles traduisent le fonctionnement de systèmes dont on cherche à comprendre et caractériser le comportement.

Les défis portent sur les aspects suivants :

- l'efficacité algorithmique pour extraire efficacement des informations temporelles ou spatiales d'une grande masse de données,
- l'intégration de connaissances expertes pouvant guider les processus d'extraction de connaissances (générales sur le domaine),
- la fiabilité et la qualité des méthodes/résultats auxquelles ces méthodes doivent particulièrement s'intéresser parce qu'elles traitent des données qui arrivent de manière continue et sont évolutives par nature et doivent prendre en compte le traitement du temps et de l'espace,
- l'expressivité des méthodes pour la représentation de l'information temporelle et spatiale dans les données et les motifs extraits. Il faudra, par exemple, prendre en compte les aspects métriques ou symboliques des relations temporelles et spatiales, l'hétérogénéité, la multi-modalité, la multi-dimensionnalité des données,
- la mise à disposition d'outils intégrés pour aider les utilisateurs à extraire des informations répondant à ses requêtes.

Construction, enrichissement et exploitation de ressources GEOgraphiques pour l'analyse de données

Dans le contexte du web et de l'open data, de nouvelles ressources géographiques comme GeoSud Theia, Geonames, GeoBase, OpenGeodata, M2ISA, Wikimapia, Geoportail, Temporal database, sont disponibles. Cette masse importante de données mise à disposition implique de faire des choix de ressources et de les adapter aux applications visées. Ainsi, la construction et/ou l'enrichissement de telles ressources constituent un enjeu crucial et s'inscrit notamment dans les problématiques des données ouvertes.

Intégrées au sein d'un processus d'extraction des connaissances, de nombreux verrous scientifiques (extraction/désambiguïsation d'entités spatiales, alignement/fusion de ressources géographiques, etc.) ont été mis en évidence, pour lesquels des méthodes spécifiques doivent être proposées. Par exemple, la construction et l'enrichissement de telles ressources à partir de textes engendre des difficultés liées à la qualité des données textuelles disponibles, la qualité et l'exhaustivité des ressources et l'adéquation des outils de TAL (Traitement Automatique du Langage) pour identifier les entités géographiques. Dans un tel contexte, il est alors essentiel de discuter et d'évaluer la qualité des ressources (par exemples, gazetteer) et outils dédiés dans un processus d'extraction et de gestion des connaissances (EGC).

L'objet de cet atelier est de réunir les étudiants, enseignants et chercheurs intéressés par cette problématique et soucieux d'en comprendre à la fois la complexité et la portée. Il s'agira également de dresser un panorama des travaux de la communauté francophone sur ce domaine afin d'identifier les enjeux et perspectives associés.

Thomas Guyet
Agrocampus-Ouest/IRISA
Rennes

Eric Kergosien
LIRMM, CNRS
Montpellier

René Quiniou
INRIA/IRISA
Rennes- Bretagne Atlantique

Mathieu Roche
Cirad, UMR TETIS
Montpellier

Membres du comité de lecture

Le Comité de Lecture FST est constitué de:

Jérôme Azé (Université Montpellier 2)
Hugo Alatrista-Salas (TETIS)
Sid-Ahmed Benabderrahmane (INRIA/IRISA)
Tassadit Bouadi (IRISA)
Peggy Cellier (INSA/IRISA)
Ahlame Douzal-Chouakria (UJF/LIG)
Frédéric Flouvat (Univ. Nouvelle-Calédonie)
élisa Fromont (Université Jean Monnet, Lab. Hubert Curien)
Catherine Garbay (CNRS/LIG)
Thomas Guyet (Agrocampus-Ouest/IRISA)
Medhi Kaytoue (LIRIS)
Florence Le Ber (ENGEES/I-Cube)
Simon Malinowski (Femto-ST)
Alice Marascu (IBM)
Jean-François Mari (INRIA)
Pierre-François Marteau (UBS/IRISA)
Yoann Pitarch (IRIT)
René Quiniou (INRIA/IRISA)
Christophe Rigotti (LIRIS)
Nazha Selmaoui-Folcher (Univ. Nouvelle-Calédonie)
Alexandre Termier (UJF/LIG)
Maguelonne Teisseire (IRSTEA/TETIS)
Cédric Wemmert (I-Cube)

Le Comité de Lecture CERGEO est constitué de:

Nicolas Béchet (IRISA, Vannes)
Sandra Bringay (LIRMM, Montpellier)
Sophie Bruguiere Fortuno (UMR TETIS, Montpellier)
Catherine Dominguès (COGIT-IGN, Paris)
Mauro Gaio (LIUPPA, Pau)
Jérôme Gensel (LIG, Grenoble)
Mouna Kamel (IRIT, Toulouse)
Cédric Lopez (Viseo, Grenoble)
Damien Palacio (Univ. de Zurich, Switzerland)
Pascal Poncelet (LIRMM, Montpellier)
Anne Ruas (IFSTTAR, Paris)
Fatiha Saïs (LRI, Paris)
Christian Sallaberry (LIUPPA, Pau)
Maguelonne Teisseire (UMR TETIS, Montpellier)

TABLE DES MATIÈRES

Conférence invitée

Accès par le contenu géographique à des corpus textuels : Complémentarité des dimensions spatiale, temporelle et thématique
Christian Sallaberry 1

CERGEO

Fouille de textes : une approche séquentielle pour découvrir des relations spatiales
Hugo Alatrista-Salas, Nicolas Béchet 5

Senterritoire : Détection et mise en valeur d?opinions liées à l?aménagement d?un territoire
Eric Kergosien, Amayas Abboute, Gilles Entringer 13

FST

Prédiction au plus tôt d?événements par règles d?épisodes
Lina Fahed, Armelle Brun, Anne Boyer 27

Un modèle sémantique spatio-temporel pour capturer la dynamique des environnements
Benjamin Harbelot, Helbert Arenas, Christophe Cruz 39

Trajectoires médianes
Laurent Etienne, Thomas Devogele 55

Caractérisation d?un parcellaire agricole : comparaison des sacs de noeuds obtenus par un chemin de Hilbert adaptatif et un graphe de voisinage
Thomas Guyet, Sébastien da Silva, Claire Lavigne, Florence Le Ber 67

Index des auteurs **81**

Accès par le contenu géographique à des corpus textuels : Complémentarité des dimensions spatiale, temporelle et thématique

Christian Sallaberry*

* Laboratoire Informatique LIUPPA (EA n°3000) - Equipe T2I, 64000 - Pau
{prénom.nom}@liuppa.fr

1 Résumé

Les travaux de la littérature relatifs à la recherche d'information géographique (RIG) dans des corpus textuels pointent notamment les verrous suivants : la reconnaissance et la résolution d'entités nommées spatiales et temporelles, l'indexation spatiale et temporelle à des fins de RI, l'appariement de couples document/requête et le calcul de scores de pertinence dédiés à la RI spatiale d'une part, et à la RI temporelle d'autre part, la RI multicritère combinant les dimensions spatiales, temporelles et thématiques, l'évaluation de tels systèmes de RIG.

La reconnaissance et la résolution d'entités nommées spatiales et temporelles (BATTISTELLI, 2011; LEIDNER et LIEBERMAN, 2011) dans des documents textuels est supportée par deux principales catégories d'approches. La première s'appuie sur un ensemble de règles, établies par des experts, permettant à un interpréteur de déterminer si un terme est une entité nommée ou non. La seconde, quant à elle, est basée sur un corpus d'apprentissage annoté manuellement permettant, à l'issu d'un traitement statistique, la constitution automatique de règles de découverte d'entités nommées applicables à des corpus plus larges. Conformément à la première catégorie d'approches, les travaux menés sous la direction de Mauro Gaio (GAIO et al., 2008) mettent en oeuvre un ensemble de règles, bâties manuellement, dédiées à l'expression de l'espace et du temps dans un corpus composé de récits de voyages : ces règles permettent le marquage et une première interprétation symbolique des entités détectées (classification puis analyse d'éventuelles relations spatiales ou temporelles associées). Dans cette interprétation, les entités absolues telles que « la ville de Pau » ou « l'an 2000 » se distinguent des entités dites relatives telles que « les environs de la ville de Pau » ou « au début de l'an 2000 ». Rappelons que ce traitements visent uniquement les contenus textuels de documents, indépendamment de leur structure ou des métadescriptions associées.

L'indexation associe une interprétation numérique (géométrie, période calendaire) aux entités spatiales et temporelles détectées dans les textes. L'organisation des index peut, par exemple, dissocier complètement les références à l'espace et au thème dans des index indépendants ou bien combiner ces deux dimensions dans des structures spécifiques stockées dans un seul et même index (VAID et al., 2005) Comme Paul Clough et al. (CLOUGH et al., 2006), nous avons fait le choix de travailler sur des index spatiaux, temporels et thématiques indépendants. Les algorithmes d'interprétation des représentations symboliques des entités tiennent compte

de l'aspect absolu ou relatif de leurs descriptions. La représentation numérique qui en résulte correspond au résultat d'une recherche dans des ressources de type gazetier dans le cas des entités spatiales absolues. Linda L. Hill décrit ces ressources et les services associés dans (HILL, 2006). Nous avons utilisé la BD-NYME et la BD-TOPO de l'IGN ainsi que l'ontologie géographique développée dans le cadre du projet ANR GEONTO. L'appariement et le calcul de scores de pertinence ont également fait l'objet de nombreuses propositions en RI spatiale (ANDOGAH, 2010) et en RI temporelle (ALONSOO, 2008; BATTISTELLI, 2011). Comme pour la majorité de ces propositions, nous avons mis au point une RI spatiale et une RI temporelle supportées par des formules ad hoc adaptées à nos corpus.

La combinaison des dimensions spatiale, temporelle et/ou thématique en RIG est généralement mise en oeuvre via des approches par filtrage (LIEBERMAN et al., 2007; VAID et al., 2005). Pour une plus grande expressivité dans le processus d'interrogation, nous avons proposé des opérateurs d'exigence et de préférence qu'il est possible d'associer à chaque critère de recherche. Nous avons mis au point une méthode d'agrégation de résultats issus de différents systèmes de RI, tenant compte des niveaux d'exigence et de préférence exprimés dans la requête. Cette méthode est inspirée des approches d'agrégation mises en oeuvre dans les systèmes d'aide à la décision (MARTEL, 1999) ainsi que dans les systèmes de recherche d'information multicritère (FARAH et VANDERPOOTEN, 2008).

La mise au point de premiers prototypes de RIG met en exergue la nécessité d'évaluer de tels systèmes (CARDOSO, 2011; MANDL, 2011). Toutefois, à l'exception de campagnes telles que TEMPEVAL (VERHAGEN et al., 2009) dédiée au temps ou GEOCLEF (GEY et al., 2005) dédiée au spatial et au thématique, il n'existe pas, à notre connaissance, de cadres d'évaluation de système de RIG combinant les dimensions spatiale, temporelle et thématique de l'information. Nous avons donc proposé un cadre expérimental dédié à ce type d'évaluation. Nous avons constitué une collection de test et mis au point un protocole d'expérimentation que nous mettons en oeuvre pour l'évaluation de nos prototypes.

Ces propositions sont décrites en détail dans (SALLABERRY, 2013).

Références

ALONSOO, R. (2008). Temporal information retrieval. In *PhD thesis*. Davis, Etats-Unis : University of California.

ANDOGAH, G. (2010). Geographically constrained information retrieval. In *PhD thesis*.

BATTISTELLI, D. (2011). Linguistique et recherche d'information, la problématique du temps. In *vol. Collection Traitement de l'information*. Lavoisier.

CARDOSO, N. (2011). Evaluating geographic information retrieval. In *SIGSPATIAL Special*, Volume 3, pp. 46–53. ACM.

CLOUGH, P., H. JOHO, et R. PURVES (2006). Judging the spatial relevance of documents for gir. In *ECIR'06 : Proceedings of the 28th European Conference on IR Research*, Volume 3936, pp. 548–552. LNCS, Springer.

FARAH, M. et D. VANDERPOOTEN (2008). An outranking approach for information retrieval. In *Information Retrieval*, Volume 11, pp. 315–334. Kluwer Academic Publishers.

GAIO, M., C. SALLABERRY, P. ETCHEVERRY, C. MARQUESUZAA, et J. LESBEGUERIES (2008). A global process to access documents'contents from a geographical point of view. In *Journal of Visual Languages and Computing*, Volume 19, pp. 3–23.

GEY, F., R. LARSON, M. SANDERSON, H. JOHO, P. CLOUGH, et V. PETRAS (2005). Geoclef'05 : The clef 2005 cross-language geographic information retrieval track overview. In *CLEF'05 : Proceedings of the 5th workshop on Cross-Language Evalution Forum*, Volume 4022, pp. 908–919. LNCS, Springer.

HILL, L. (2006). Georeferencing : the geographic associations of information, digital libraries and electronic publishing. In *Digital libraries and electronic publishing*. MIT Press.

LEIDNER, J. et M. LIEBERMAN (2011). Detecting geographical references in the form of place names and associated spatial natural language. In *SIGSPATIAL Special*, Volume 3, pp. 5–11.

LIEBERMAN, M., H. SAMET, J. SANKARANARAYANAN, et J. SPERLING (2007). Steward : Architecture of a spatio-textual search engine. In *GIS'07 : Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems*, pp. 1–8. ACM.

MANDL, T. (2011). Evaluating gir : geography-oriented or user-oriented ? In *SIGSPATIAL Special*, Volume 3, pp. 42–45. ACM.

MARTEL, J. (1999). L'aide multicritère la décision : méthodes et applications. In *Actes de la conférence CORS - SCRO*, pp. 6–16.

SALLABERRY, C. (2013). Geographical information retrieval in textual corpora. In *Geographical Information Systems Series*. Wiley-ISTE.

VAID, S., C. JONES, H. JOHO, et M. SANDERSON (2005). Spatio-textual indexing for geographical search on the web. In *SSTD'05 : Proceedings of the 9th international Symposium on Spatial and Temporal Databases*, Volume 3633, pp. 218–235. LNCS, Springer.

VERHAGEN, M., R. GAIZAUSKAS, F. SCHILDER, M. HEPPLE, J. MOSZKOWICZ, et J. PUSTEJOVSKY (2009). The tempeval challenge : identifying temporal relations in text. In *Language Resources and Evaluation*, Volume 43, pp. 161–179. Springer.

Fouille de textes : une approche séquentielle pour découvrir des relations spatiales

Hugo Alatrista-Salas*, Nicolas Béchet**

*UMR TETIS, 500 rue Jean-François Breton, F-34093 Montpellier
hugo.alatrista-salas@teledetection.fr
**IRISA, Université de Bretagne Sud, Rue André Lwoff, BP 573, 56017 VANNES Cedex
nicolas.bechet@univ-ubs.fr

Résumé. Dans cet article, nous présentons les premières étapes d'un projet de fouille de données textuelles. Plus précisément, nous appliquons un algorithme d'extraction de motifs séquentiels sous contraintes multiples afin d'identifier des relations entre entités spatiales. Les premiers résultats obtenus montrent l'intérêt de l'utilisation de cette approche et ses limites. Dans cet article, nous détaillons les premières bases de travaux plus ambitieux dont l'objectif est d'apporter des informations cruciales permettant de compléter l'analyse des images satellitaires.

1 Introduction

Les données satellites recèlent une multitude d'informations et leur analyse demande un investissement humain conséquent. La mise à disposition d'images avec une forte répétitivité temporelle pose le problème d'une gestion plus efficace, même si de nombreuses méthodes de classification automatique permettent d'aider l'expert dans cette tâche. De plus, les techniques actuelles de télédétection ne permettent pas une analyse efficace et rapide des images satellites dès qu'elles ont une forte répétitivité temporelle. Par exemple, grâce à la télédétection, nous pouvons déterminer des zones de cultures, mais nous ne pouvons différencier une culture de maïs ou de vignes pour un lieu donné. Il devient donc urgent de proposer des méthodes complémentaires à celles issues de la télédétection afin d'enrichir les connaissances associées aux images satellites. Dans ce contexte, l'utilisation de méthodes complémentaires à celles de télédétection telle que la fouille de textes semble pertinente.

Dans cet article, nous nous intéressons à la mise en place de méthodes d'extraction d'entités spatiales et de relations sémantiques entre ces dernières à partir de données textuelles. Actuellement, la plupart de ces méthodes s'appuient sur la mise en place de règles linguistiques (patrons lexico-syntaxiques). Seulement, étant donné le gros volume de données, de telles méthodes possèdent des limites et ne sont pas exhaustives. Dans le but de traiter une quantité importante de données, il semble essentiel de mettre en place des méthodes originales qui combinent des méthodes à base de motifs séquentiels et des approches de fouille de textes. Ce type de combinaison est proposé dans cet article.

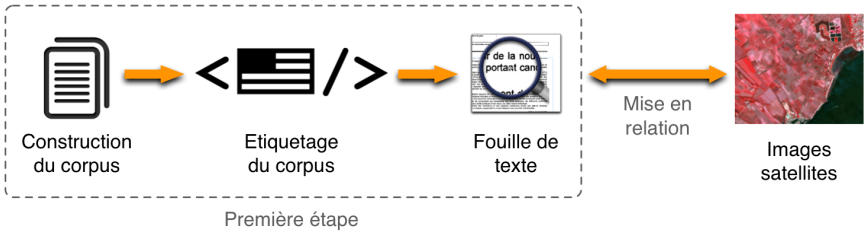

FIG. 1 – *Schema du projet de fouille de données textuelles*

2 Le corpus

Avec ses 7500 hectares, l'étang de Thau est, après la Méditerranée, la plus profonde et la plus grande étendue d'eau du Languedoc-Roussillon. Ce bassin s'étire le long de la côte lagune sur 19 km et 5 de large et plus de 30% du territoire de Thau est couvert par l'eau. Cet étang est un lieu dont le développement des activités économiques, tels que la production d'huitre et de moules, a augmenté énormément ces derniers années. Chaque année 13 000 tonnes d'huîtres, soit 10% de la production nationale sont extraits, ainsi que 3 000 tonnes de moules [1].

Dans un premier temps, nous nous intéresserons à des données textuelles issues du territoire côtier de Thau, déjà largement exploré par la communauté scientifique locale. Les données disponibles permettent de balayer les dispositifs numériques de cette région (journaux locaux, blogs citoyens ou d'associations, sites web des institutions locales, rapports, etc.) pour de premières extractions.

Le corpus à notre disposition est composé de 14 documents extraits du journal *Midi Libre* [2]. Il comporte des opinions, rapports et articles sur des différentes rubriques (économiques, sociales, politiques, etc.) de la région de *Sète*. Ces documents ont été nettoyés et postérieurement étiquetés en utilisant l'outil LinguaStream [3] qui permet de traiter les textes et identifier les données géo-spatiales des *Entités Spatiales* (ES). La Figure 2 illustre un exemple d'étiquetage du corpus. Les balises *XML* illustrées dans la Figure 2 sont décrites ci-dessous :

- sentence : l'identifiant de la séquence (une phrase),
- token : peut être une proposition (PRO nous), un verbe (V avons), etc.,
- es : Les Entités Spatiales (ES) exprimées dans un texte sont constituées d'au moins une Entité Nommée (EN) et d'un nombre variable d'indicateurs spatiaux, précisant leur localisation. Généralement, une ES se réfère à une EN de lieu dont la position est connue (par exemple, Selles-sur-Cher, au nord de Jacou, dans les zones montagneuses),
- nomCommun : un nom commun, composé des différents Tokens, par exemple, (NOM l'ensemble)(PREDET du)(N territoire),
- indirection : représente une position, par exemple, (PREP entre) ou (NOM est),
- vpe : comporte une expression incluant un verbe du type "perception", "position", etc. et un *NomCommun*, par exemple, (V vu)(CONJC que)(DET l')(NOM espace)(A urbain).

1. Syndicat Mixte du Bassin de Thau - 328, Quai des Moulins - 34200 SETE.
2. http ://www.midilibre.fr/
3. http ://www.linguastream.org/

FIG. 2 – *Étiquetage du corpus : exemple*

Concernant les entités spatiales ES, cinq types d'indications spatiales existent : l'orientation (au sud de), la distance (à 1 heure de marche de, à 20 km de), l'adjacence (près de, loin de, la périphérie de), l'inclusion (le quartier de, la frontière entre, le sommet de) et la figure géométrique qui définit l'union ou l'intersection liant au moins deux ES (entre A et B, entre le triangle A, B, C, à l'intersection de A et B, la frontière A et B, etc.).

Dans la section suivante, nous décrirons la méthode d'extraction de motifs séquentiels utilisée dans nos expérimentations.

3 Extraction de motifs séquentiels : panorama et définitions

Le problème de la recherche de motifs séquentiels a été introduit par R. Agrawal dans [2] et appliqué avec succès dans de nombreux domaines comme la biologie [3, 4], la fouille d'usage du Web [5, 6], la détection d'anomalie [7], la fouille de flux de données [8], l'extraction de motifs spatio-temporels permettant l'étude des épidémies [9] ou la description des comportements au sein d'un groupe [10]. Des approches plus récentes [11] utilisent les motifs séquentiels pour décrire les évolutions temporelles des pixels au sein des séries d'images satellites.

Dans cette section, nous introduisons les définitions relatives à l'algorithme de fouille de motifs séquentiels sous contraintes multiples selon [1].

Dans l'extraction de motifs séquentiels, un *itemset* est un ensemble non vide de littéraux appelés *items* noté par $I = (i_1, i_2, \ldots, i_n)$. Par exemple, *(NOM ville)* est un itemset composé des deux items : *NOM* et *ville*.

Une *séquence* S est une liste ordonnée, non vide, d'itemsets notée $< I_1 I_2 \ldots I_m >$ où I_j est un itemset. Par exemple, *S=<(DET l')(NOM urbanisation)(ADJ ces)(ADJ dernières)(NOM années)>* est une séquence d'itemsets. Une *n-séquence* est une séquence composée de n itemsets. Par exemple, *S* est une 4-séquence.

Une séquence $S =< I_1 I_2 \ldots I_p >$ est une sous-séquence d'une autre séquence $S' =< I'_1 I'_2 \ldots I'_m >$, représentée par $S \preceq S'$, s'il existe des entiers $k_1 < k_2 < \ldots < k_j < \ldots < k_p$ tels que $I_1 \subseteq I'_{k_1}, I_2 \subseteq I'_{k_2}, \ldots, I_p \subseteq I'_{k_p}$. Par exemple, la séquence S' = *<(ces)(NOM années)>* est une sous-séquence de *S* car *(ces)* $\subseteq$ *(ADJ ces)* et *(NOM années)* $\subseteq$ *(NOM années)*. Toutefois, *<(ADJ)(DET l')(NOM urbanisation)>* n'est pas une sous-séquence de S.

Une base de séquences *BDS* est un ensemble de tuples $(seqID, S)$ où *seqID* est l'identifiant de la séquence et S est une séquence. Par exemple, le Tableau 1 représente une base de séquences contenant trois séquences. Un tuple *(seqID, S)* contient une séquence S_1 si et seulement si $S_1 \preceq S$. Le support d'une séquence S_1 dans la base de séquences BDS, noté par $supp(S_1, BDS)$ est le nombre de tuples dans la base de séquences qui contiennent la séquence S_1. Par exemple, dans le Tableau 1, le support des sous-séquences *(PRO)(V)*, *(V correspond)(PREP à)* et *(PREP de)(NOM)* est 2 car ils apparaissent dans deux tuples différentes de la base de séquences.

Soit $minSupp$ le support minimal fixé par l'utilisateur. Une séquence qui vérifie le support minimal (i.e. dont le support est supérieur à $minSupp$) est une *séquence fréquente*.

seqID	Séquence
S1	(PRO Nous)(V avons)(V vu)(CONJC que)(DET l')(NOM espace)(ADJ urbain) (V correspond)(PREP à)(DET la)(NOM ville)
S2	(PRO Il)(V constituait)(NOM 11%)(PREP de)(DET l')(NOM ensemble)(ART du) (NOM territoire)
S3	(DET L')(NOM urbanisation)(V correspond)(PREP à)(DET une)(NOM expansion) (PREP de)(NOM 528)(NOM m2)

TAB. 1 – *Exemple d'une base de séquences*

Le problème de la recherche de motifs séquentiels dans une base de données consiste à trouver les séquences dont le support est supérieur au support minimal spécifié. Chacune de ces séquences fréquentes est souvent appelée *motif séquentiel*. Le nombre de séquences fréquentes extraites peut s'avérer très important. La représentation de séquences condensées telle que les séquences *séquences fréquentes fermées* (closed patterns) [12] peut donc être utilisée pour éviter la redondance sans perte d'information. Une séquence fréquente S est dit fermée s'il n'existe pas une autre séquence fréquente S' tel que $S \preceq S'$ et $supp(S, BDS) = supp(S', BDS)$. Par exemple, pour un support minimal $minSupp = 2$ la séquence *<(V correspond)>* du Tableau 1 n'est pas fermée, tandis que la séquence *<(V correspond)(PREP à)(DET)>* est fermée.

Le paradigme d'extraction de motifs séquentiels sous contraintes [13] apporte des techniques utiles pour inclure des contraintes choisies pour l'utilisateur (autres que le support minimal) afin de se concentrer sur les motifs les plus intéressants pour lui. Par exemple, il est possible de définir des contraintes pratiques telles que la contrainte de *l'écart*. L'écart (gap) dans une séquence est la nombre d'itemsets qui peuvent être ignorés compris entre deux itemsets d'une séquence fréquente S. $g(M, N)$ représente l'écart dont la taille est comprise dans l'intervalle $[M, N]$ où M et N sont des nombres entiers. L'intervalle $[M, N]$ est appelé *contrainte d'écart*. Une séquence fréquente satisfaisant la contrainte d'écart $[M, N]$ est dénotée par $P_{[M,N]}$. A savoir, il existe un écart $g(M, N)$ entre tous les itemsets voisins de $P_{[M,N]}$. Par exemple, dans le Tableau 1, $P_{[0,2]} = (V)(PREP)$ et $P_{[1,2]} = (V)(PREP)$ sont deux séquences satisfaisant la contrainte d'écart. Cependant $P_{[0,2]}$ est trouvée dans les séquences 1 et 3 tandis que la séquence $P_{[1,2]}$ est trouvée dans la séquence 2.

4 Extraction de motifs séquentiels textuels

A partir de documents textuels, ce travail a pour objectif d'identifier des relations entre les informations spatiales et tout autre mot contenu dans les documents. Dans ce cadre, une information spatiale est définie comme un nom toponymique précédé ou non d'un indicateur spatial et/ou d'une ou plusieurs relation(s) spatiale(s). Nous indiquons qu'un syntagme est lié à une information spatiale si la fréquence de la relation entre ce syntagme et l'ensemble des éléments de type information spatiale est supérieure à un seuil minimal fixé manuellement.

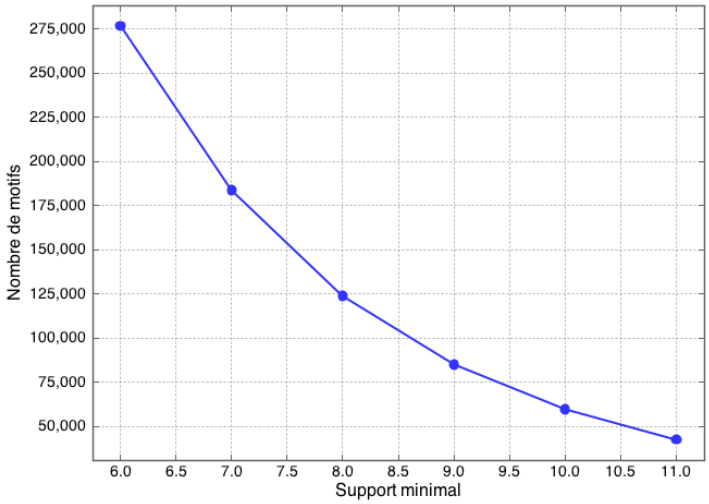

FIG. 3 – *Impact du support minimal sur le nombre de motifs extraits*

Les premières expérimentations ont été menées sur le corpus décrit dans la Section 2. Pour les expérimentations, nous avons appliqué un algorithme d'extraction de motifs séquentiels sous contraintes multiples (décrit dans la Section 3) où chaque phrase étiquetée du corpus est une séquence d'itemsets, par exemple, *(DET L')(NOM espace)(ADJ urbain)...(PRE de)(TOPO)*. La base de séquences ainsi construite est composée de 241 séquences.

Pour les tests de cette première étape, nous avons expérimenté différents paramètres portant sur les contraintes :

- le support minimal : plusieurs valeurs de seuil ont été utilisées,
- nous avons sélectionné les séquences (phrases) contenant au moins un toponyme caractérisé pour la balise TOPO,
- la contrainte d'écart (espace entre itemsets) a été fixée à 0, c'est-à-dire des itemsets consécutifs,
- la valeur de la longueur minimale (le nombre minimal d'itemsets comportant la séquence) a été fixée à 3,
- les motifs séquentiels doivent obligatoirement contenir l'item *TOPO* (contrainte d'appartenance).

La Figure 3 représente l'impact du support minimal sur le nombre de motifs extraits. Les motifs fréquents illustrés dans le Tableau 2 correspondent aux experimentations avec un support minimal de 7 (un motif est inclus dans au moins 7 séquences de la base de séquences).

Les premiers résultats semblent cohérents, cependant, ils sont trop génériques même pour un support très faible. Il y a très peu d'itemsets dans les résultats et quand il y en a, il s'agit de termes très généraux comme "de" ou "et". Le résultat semble difficilement exploitable, d'autant que nous avons beaucoup de motifs (environ 183 000 motifs pour un seuil minimal de 7). Un outil de filtrage ou ranking de motifs (par exemple, les *top-k*) semble intéressant à mettre en place.

Les motifs qui finissent par des mots vides (conjonction, préposition, déterminant) ne sont pas intéressants. Pour résoudre ce problème, nous pouvons : a) faire un ranking en mettant en avant ceux avec une composition intéressante, par exemple, *(NOM)(ADV)(TOPO)* ; ou b) garder uniquement les préfixes des motifs, par exemple, *(NOM)(ADV)(TOPO)(DET)* devient *(NOM)(ADV)(TOPO)*.

Motif séquentiel	**Support**
<(TOPO)(DET)(NOM)(PREP)(DET)(TOPO)(NOM)>	7
<(TOPO)(DET)(PREP)(DET)(TOPO)(TOPO)>	7
<(TOPO)(DET)(PREP de)(N)(PREP)(TOPO)>	7
<(V)(NOM)(PREP)(NOM)(CONJC)(PREP)(DET)(TOPO)>	7
<(TOPO)(DET)(NOM)(DET la)(TOPO)>	8
<(DET)(NOM)(PREP)(NOM)(PREP)(TOPO)(CONJC)(DET)(TOPO)>	8
...	...

TAB. 2 – *Motifs séquentiels obtenus pour un support minimal de 7*

Finalement, pour limiter le nombre d'itemsets à analyser et évaluer, une solution est d'extraire uniquement les motifs qui se terminent par un toponyme *(TOPO)* car on trouve très souvent cette structure dans la langue française.

5 Conclusion

Dans cet article, nous avons présenté les premières étapes d'un projet de fouille de données textuelles. Nous avons plus particulièrement appliqué un algorithme d'extraction de motifs séquentiels pour extraire des relations spatiales. Ces relations spatiales seront postérieurement utilisées pour enrichir à faible "coût" humain les connaissances associées aux images satellites, plus précisément, nous allons déterminer les descripteurs linguistiques pertinents à partir des textes, puis les mettre en correspondance avec les données images via l'identification des segments de textes du corpus spatialement pertinents pour une image donnée. Ce travail préliminaire conduira par la suite à une généralisation de la méthode sur un large spectre couvrant les régions de Strasbourg, Pau, Caen et Montpellier offrant ainsi un corpus de données massives et hétérogènes.

Les perspectives de ce travail préliminaire sont nombreuses. Tout d'abord, nous devons valider la démarche proposée sur un jeu de données plus conséquent. Nous nous attacherons

également à proposer une mesure de validation objective permettant une selection des motifs obtenus.

Il est important de noter que ces travaux sont menés dans le cadre du projet ANIMITEX (CNRS-MASTODONS) [4].

Références

[1] Nicolas Béchet, Peggy Cellier, Thierry Charnois, and Bruno Crémilleux. Discovering linguistic patterns using sequence mining. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 7181 of *Lecture Notes in Computer Science*, pages 154–165. Springer Berlin Heidelberg, 2012.

[2] Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In Philip S. Yu and Arbee L. P. Chen, editors, *Proceedings of the Eleventh International Conference on Data Engineering, March 6-10, 1995, Taipei, Taiwan*, pages 3–14. IEEE Computer Society, 1995.

[3] Ke Wang, Yabo Xu, and Jeffrey Xu Yu. Scalable sequential pattern mining for biological sequences. In *CIKM '04 : Proceedings of the thirteenth ACM International Conference on Information and Knowledge Management*, pages 178–187, New York, NY, USA, 2004. ACM.

[4] Paola Salle, Sandra Bringay, and Maguelonne Teisseire. Mining discriminant sequential patterns for aging brain. In Carlo Combi, Yuval Shahar, and Ameen Abu-Hanna, editors, *Artificial Intelligence in Medicine, 12th Conference on Artificial Intelligence in Medicine, AIME 2009, Verona, Italy, July 18-22, 2009. Proceedings*, Lecture Notes in Computer Science, pages 365–369, 2009.

[5] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, and Hua Zhu. Mining access patterns efficiently from web logs. In Takao Terano, Huan Liu, and Arbee L. P. Chen, editors, *Knowledge Discovery and Data Mining, Current Issues and New Applications, 4th Pacific-Asia Conference, PADKK 2000, Kyoto, Japan, April 18-20, 2000, Proceedings*, Lecture Notes in Computer Science, pages 396–407. Springer, 2000.

[6] Florent Masseglia, Pascal Poncelet, Maguelonne Teisseire, and Alice Marascu. Web usage mining : extracting unexpected periods from web logs. *Data Mining and Knowledge Discovery (DMKD)*, 16(1) :39–65, 2008.

[7] Julien Rabatel, Sandra Bringay, and Pascal Poncelet. Aide à la décision pour la maintenance ferroviaire préventive. In Sadok Ben Yahia and Jean-Marc Petit, editors, *Extraction et gestion des connaissances (EGC'2010), Actes, 26 au 29 janvier 2010, Hammamet, Tunisie*, Revue des Nouvelles Technologies de l'Information, pages 363–368. Cépaduès-Éditions, 2010.

[8] Alice Marascu and Florent Masseglia. Mining sequential patterns from data streams : a centroid approach. *Journal of Intelligent Information Systems*, 27(3) :291–307, 2006.

4. http ://www.lirmm.fr/ mroche/ANIMITEX/

[9] Hugo Alatrista-Salas, Sandra Bringay, Frédéric Flouvat, Nazha Selmaoui-Folcher, and Maguelonne Teisseire. The pattern next door : Towards spatio-sequential pattern discovery. In *PaKDD'12, LNAI-7302*, volume 2, pages 154–168. Springer-Verlag, 2012.

[10] Dilhan Perera, Judy Kay, Irena Koprinska, Kalina Yacef, and Osmar R. Zaïane. Clustering and sequential pattern mining of online collaborative learning data. *IEEE Transactions on Knowledge and Data Engineering*, 21(6) :759–772, 2009.

[11] A. Julea., N. Meger, and Ph. Bolon. On mining pixel based evolution classes in satellite image time series. In *Proc. of the 5th Conf. on Image Information Mining : pursuing automation of geospatial intelligence for environment and security (ESA-EUSC 2008)*, page 6, 2008.

[12] Xifeng Yan, Jiawei Han, and Ramin Afshar. Clospan : Mining closed sequential patterns in large databases. In Daniel Barbará and Chandrika Kamath, editors, *Proceedings of the Third SIAM International Conference on Data Mining, San Francisco, CA, USA, May 1-3, 2003*. SIAM, 2003.

[13] Guozhu Dong and Jian Pei. *Sequence Data Mining*, volume 33 of *Advances in Database Systems*. 2007.

Senterritoire : Détection et mise en valeur d'opinions liées à l'aménagement d'un territoire

Eric Kergosien[*,**], Amayas Abboute[**], Gilles Entringer[**]

[*] LIRMM, UMR 5506, 161 rue Ada, 34392 Montpellier - France
{prénom.nom}@lirmm.fr
[**] UMR TETIS, 500 rue Jean-François Breton, 34093 Montpellier - France
{prénom.nom}@teledetection.fr

Résumé. De nombreux travaux ont été réalisés en extraction d'informations et plus particulièrement en fouille de données d'opinions dans des contextes spécifiques tels que les critiques de films, les évaluations de produits commerciaux, les discours électoraux... Dans le cadre du projet SENTERRITOIRE, nous nous posons la question de l'adéquation de ces méthodes pour des documents associés à l'aménagement des territoires. Ces documents renferment différents types d'informations se rapportant à des acteurs, des opinions, des informations géographiques, et tout autre aspect lié plus généralement à la notion de territoire. Cependant, il est extrêmement difficile d'identifier puis de lier les opinions à ces informations. Dans cet article, nous décrivons (1) la méthode semi-automatique nommée OPILAND (OPinion mIning from LAND-use planning documents) qui combine une chaîne de Traitements Automatiques du Langage Naturel et des techniques de Fouilles de Textes pour détecter les opinions relatives aux informations géospatiales formant un territoire, et (2) le démonstrateur associé.

1 Introduction

La notion de territoire, et plus spécifiquement d'aménagement du territoire, fait référence à différents concepts tels que les entités spatiales et temporelles, les acteurs, les opinions, l'histoire, la politique, etc. Dans cet article, nous nous intéressons à la détection d'opinions liées aux entités nommées de type Lieu, que l'on nomme Entité Spatiale, pour des corpus décrivant des territoires politiques et administratifs à l'échelle des collectivités locales, voire régionales.

L'extraction d'entités nommées ainsi que la fouille d'opinions ont été intensivement appliquées dans divers domaines tels que l'analyse automatique des critiques de films, articles politiques, tweets, etc. Dans le contexte de l'aménagement du territoire, même si les informations publiées sur le web (blogs, forums, etc.) et dans les médias (journaux, comptes, etc.) expriment des sentiments, les approches traditionnelles de fouille de textes ne parviennent pas à extraire des opinions de façon exhaustive et précise de ces domaines spécialisés. Aussi, il n'existe pas à notre connaissance d'outil permettant de visualiser les opinions qu'on les acteurs sur leur territoire.

Nous avons donc proposé une approche originale, appelée OPILAND (Tahrat et al., 2013; Kergosien et al., 2013, 2014) (OPinion mIning from LAND-use planning documents, ce qui signifie en français « Fouille d'opinions pour les documents liés à l'aménagement du territoire »), définie pour identifier semi-automatiquement des opinions relatives aux entités nommées dans des contextes spécialisés. Ce travail est réalisé dans le cadre du projet SENTERRITOIRE [1] qui vise à développer un environnement de prise de décision pour les documents d'aménagement du territoire. Le démonstrateur mis en place nommé SENTERRITOIREVIEWER doit appuyer les experts du territoire Thau Agglomération (Syndicat Mixte du Bassin de Thau SMBT, chercheurs géographes, etc.) dans leur travail de diagnostic du territoire. L'objet de cet article est de présenter le démonstrateur SENTERRITOIREVIEWER permettant de représenter les opinions extraites des textes selon les composantes spatiales et temporelles.

Le document est structuré comme suit. Dans la section 2, un aperçu des méthodes d'extraction d'opinions est présenté. Dans la section 3, les différentes étapes de la méthode OPILAND sont résumées et un pointeur est donné vers les articles détaillant chacune des étapes. La section 4 présente les modules principaux du démonstrateur SENTERRITOIREVIEWER. Nous concluons et présentons les perspectives de ces travaux en section 5.

2 Etat de l'art

Les **Entités Nommées** *EN* ont été définies comme des nom des personnes, des lieux et des organisations lors des campagnes d'évaluations américaines appelées MUC (Message Understanding Conferences), qui furent organisées dans les années 90. De nombreuses méthodes permettent de reconnaître les ENs en général et les ES en particulier (Nadeau et Sekine, 2007) selon des méthodes supervisés ou non supervisées.

Concernant la **fouille de données d'opinion** et le **TALN**, l'analyse de la subjectivité et des opinions exprimées par des personnes dans les textes (journaux, documents techniques, blogs, commentaires des internautes, lettres aux éditeurs, etc.) est connue comme l'analyse des sentiments ou des opinions (Pang et Lee, 2010; Liu, 2012). La reconnaissance de polarité tente de classer les textes selon la positivité ou la négativité des opinions qui y sont exprimées. Deux approches principales peuvent être identifiées : l'une fondée sur le recensement des termes positifs et négatifs dans chaque texte (Turney, 2002), et l'autre sur l'apprentissage automatique à partir de textes annotés (Esuli et Sebastiani, 2006). Les approches hybrides semblent proposer les meilleurs résultats (Kennedy et Inkpen, 2006). Outre la classification de textes d'opinions, des travaux se sont concentrés sur la construction automatique de vocabulaires spécialisés liés à l'opinion (Husaini et al., 2012; Duthil et al., 2011). Les approches incrémentales proposées reposent sur des méthodes de fouille du web afin d'apprendre un vocabulaire d'opinion lié à un thème ou sous-thème donné.

Concernant la problématique applicative associée à l'**analyse des ressentis des acteurs** à propos de l'**aménagement des territoires**, le panorama des travaux existants soulignent l'implication de différentes communautés. En effet, depuis une trentaine d'années, le concept de territoire a été largement utilisé et discuté, sous des acceptations diverses, tant, dès l'origine, par les éthologues et écologues, que par les géographes, les sociologues (Barel, 1981), les économistes (Pecqueur, 2007) les politistes (Alliès, 1980), les agronomes (Deffontaines et al.,

1. http ://msh-m.fr/programmes-2012/senterritoire/

2001), que les philosophes (Deleuze et Guattari, 1980)... La géographie en particulier a été particulièrement prolixe, partageant sa production entre une analyse du territoire « sujet politique » et une analyse du territoire « sujet social » (Vanier, 2009) : d'un côté, la géographie sociale a analysé la dimension identitaire du territoire, les rapports d'appartenance et d'ancrage (Buléon et Méo, 2005) ; de l'autre, une géographie plus politique s'est efforcée d'éclairer la dimension de représentation du territoire, à travers une analyse des dispositifs d'action publique (Debarbieux et Vanier, 2002).

Concernant la visualisation, différents travaux ont pour objet l'analyse de l'opinion et la classification de sujets. Certains de ces travaux utilisent la visualisation comme un outil de navigation (Fry, 2008). Différents travaux s'appliquent à analyser les opinions concernant la politique. Presidential Watch 2008 [2] est un outil de représentation graphique du débat présidentiel aux états unis sur Internet qui permet d'identifier et d'analyser les tendances importantes de la campagne, l'orientation des opinions des citoyens et leurs évolutions en temps réel. Plus proche de nous, la blogosphère politique [3], regroupant l'ensemble des sites et blogs de citoyens qui alimentent le débat politique sur l'internet Français, donne à voir plus de 2000 sites web classés par leur appartenance politique. Des liens politiques identifiés dans les sites et blogs sont explicités dans l'application. Un autre outil nommé WebAtlas (Ghitalla, 2004) propose une carte de mots qui montre un espace politique « commun » à partir de huit controverses analysées par des étudiants de Sciences Po. Dans le secteur de l'actualité, des travaux s'appuient sur l'algorithme de treemap pour proposer une vue globale de l'agrégateur google News (Ong et al., 2005). L'outil permet ainsi de comparer l'importance de différents sujets dans les journaux. Dans le domaine des sciences sociales, l'outil White Spectrum (BBCTwo, 2008) propose d'analyser le débat suscité par les programmes de la BBC2 en 2008. Chaque particule représente des opinions tirées de commentaires postés par les utilisateurs. Dans le même ordre d'idées, WeFeelFine (Kamvar et Harris, 2011) est un site Web interactif qui identifie en continue les expressions de l'émotion à travers les blogs, puis les affiche dynamiquement sous forme d'un nuage de points. OpinionSeek (Wu et al., 2010) propose également une visualisation des opinions sous forme d'un nuage de points enrichis automatiquement à partir de l'analyse en continue des avis postés par les clients d'une chaine d'hôtels. Plus proche de nos travaux, LoveWillConquer est un projet développé pour célébrer la Saint Valentin. Il récupère en continue les tweets qui mentionnent les mots « amour » et « haine », puis les représente visuellement via une barre temporelle. Un filtre spatial est également disponible mais la représentation visuelle dans l'espace n'est pas proposée. (Offenhuber, 2005) propose un navigateur en 3 dimensions qui permet de se déplacer dans une représentation spatiale des opinions relatives à la Communauté Européenne.

Aussi loin que portent nos recherches, il n'existe pas de travaux concernant la fouille de données d'opinions en lien avec l'aménagement du territoire. Dans cet article, nous nous focalisons sur un démonstrateur associé à l'approche OPILAND. Ce démonstrateur est un moteur permettant de naviguer dans un corpus selon trois composantes (Opinion - Temps - Espace) afin de visualiser les opinions dans le temps et l'espace et cela de façon dynamique.

2. http ://politicosphere.net/map/
3. http ://blogopole.observatoire-presidentielle.fr/

3 OPILAND : extraction d'opinions relatives aux entités nommées

Dans l'objectif de définir des vocabulaires spécifiques relatifs à des entités nommées dans le cadre de l'aménagement du territoire, nous avons proposés l'approche OPILAND qui se décompose en trois étapes (*cf.* Figure 1).

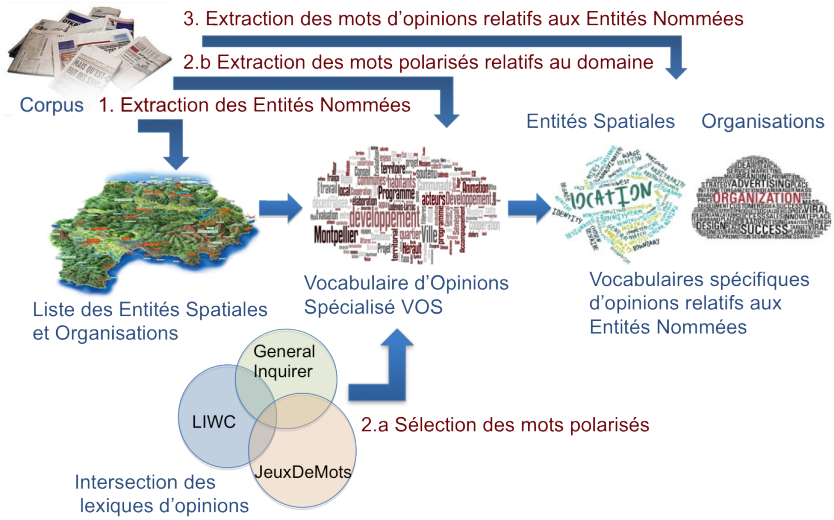

FIG. 1 – *Approche* OPILAND *pour l'extraction des opinions relatives aux Entités Nommées*

La première étape consiste à extraire semi-automatiquement des entités spatiales telles que : « Le bassin de Thau », « étang de Thau » , « ville de Sète », « en France » et « dans les alentours de Sète et Montpellier », ainsi que des organisations telles que « le SMBT » et « Conseil Général de la Pêche en Méditerranée ». À cette fin, nous proposons de fusionner une chaîne de TALN et une méthode d'apprentissage supervisé pour améliorer l'extraction d'entités nommées (Tahrat et al., 2013).

Ces entités doivent être enrichies par des informations territoriales liées aux expressions temporelles et aux opinions que les acteurs ont sur leur propre territoire mais cette seconde tâche reste difficile. Pour ce faire, nous définissons dans une deuxième étape un vocabulaire spécialisé d'opinions sur lequel nous nous appuyons ensuite pour calculer un score de polarité pour chaque document du corpus. Cette évaluation doit nous permettre de valider le vocabulaire défini (Kergosien et al., 2013). Par exemple, les termes suivants sont liées à notre domaine d'application : « magnifique », « croissante », « belle », « fièvre », « solution », craint », « critiqué » et « polluées ». Les vocabulaires d'opinions spécialisés liés au domaine sont construits sur la base de lexiques d'opinions généralistes produits par la communauté scientifique. Concernant les entités temporelles, nous intégrons la chaine de traitements linguistiques définie par (Lacayrelle et al., 2007). Nous restreignons la notion de composante temporelle aux données calendaires absolues (date, intervalle de dates ou période) et nous ne prenons donc pas en

compte les entités nommées tels que les évènements dans l'histoire (Par exemple « La révolution française »).

Dans une troisième étape, notre objectif est d'identifier les opinions relatives aux entités spatiales et aux organisations (Kergosien et al., 2014). Cette étape consiste à identifier les couples candidats Opinion-Entités Nommées, puis identifier la taille de fenêtre contenant le plus de couples pertinents, et enfin filtrer les couples correspondantes présents dans la fenêtre sélectionnée. Comme par exemple, « magnifique-bassin de Thau » , « belle-ville de Sète » , « SMBT-craint », « critiqué-Conseil général de la Pêches en Méditerranée » et « SMBT-solution ». Parmi les opinions identifiées dans ces extraits, les termes « magnifique » et « belle » sont liés aux ES tandis que « craint », « critiqué » et « solution » sont liés à des entités Organisations.

3.1 Extraction des Entités Spatiales et des Organisations

Nous adoptons un module standard de Traitement Automatique du Langage Naturel dans le domaine de la Recherche d'Information Géographique (RIG) (Abolhassani et al., 2003) : (I) la lemmatisation, (ii) l'analyse morpho-lexicale, (iii) l'analyse syntaxique, (iv) l'analyse sémantique. Dans ce processus, l'étape de lemmatisation détermine le lemme pour un mot donné. L'analyse morpho-lexicale consiste à identifier pour chaque mot la catégorie grammaticale (nom, adjectif, etc.) ainsi que les paramètres de flexion (nombre, temps, etc.). L'analyse syntaxique a pour objectif d'identifier le rôle des termes ou des syntagmes dans la phrase. Elle s'appuie pour cela sur des grammaires pour trouver les relations entre les mots. Enfin, l'analyse sémantique permet de réaliser une interprétation plus spécifique des syntagmes retenus. Nous obtenons ainsi une liste contenant des entités spatiales et une seconde liste composée des organisations. La deuxième étape consiste maintenant à construire un vocabulaire d'opinions spécialisées à partir de lexiques généralistes d'opinions disponibles dans la communauté.

3.2 Extraction d'un vocabulaire d'opinions spécialisé relatif à l'aménagement du territoire

Notre objectif est d'identifier automatiquement les opinions dans corpus relatifs à un domaine d'application spécifique. Le projet SENTERRITOIRE traite un corpus décrivant le territoire de l'étang de Thau. Les méthodes d'extraction d'opinions classiques échouent lorsqu'elles sont appliquées sur ce type de corpus peu volumineux et contenant un vocabulaire complexe. Aussi, les lexiques d'opinions généralistes ne sont pas adaptés à ce domaine spécifique.

Pour identifier une opinion dans un texte, nous avons défini trois types de vocabulaire :

1. Vocabulaire d'Opinions Pivots généraliste (VOPG) qui correspond à une liste de mots pivots polarisés ;
2. Vocabulaire d'Opinions Contextualisé (VOC) qui correspond à une liste de mots qui seront polarisés au regard du contexte dans lesquels ils se trouvent ;
3. Vocabulaire d'Opinions Spécialisé (VOS) qui correspond à une liste de mots polarisés liés au domaine. Ce vocabulaire a été validé par les géographes qui travaillent sur le projet SENTERRITOIRE.

Une fois les différents vocabulaires définis, l'étape suivante consiste à attribuer un score de polarité à chaque document. Deux types de pré-traitement sont réalisés : (1) pré-traitement statistique : suppression des mots non discriminants en utilisant leur score IDF (Inverse Document Frequency), (2) pré-traitement linguistique : Attribution de poids aux mots selon leur catégorie grammaticale (*cf. Cat*) identifiée via Tree-Tagger.

3.3 Identification de vocabulaires d'opinions spécifiques aux Entités Spatiales et Organisations

Nous avons proposé ensuite dans une troisième étape de construire des vocabulaires d'opinions spécifiques aux Entités Spatiales ES et aux OE. La difficulté ici est d'identifier des relations entre les opinions et les entités nommées. L'approche, fondée sur des traitements statistiques, se décompose en quatre étapes : (1) indexation des documents, (2) identification des couples correspondantes Opinion-ENs, (3) sélection de la taille d'une fenêtre pertinente pour l'identification des couples, et (4) filtrage des opinions pertinentes.

Concernant la composante temporelle, nous relions une entité calendaire identifiée via la chaine de traitement linguistique (Lacayrelle et al., 2007) à un couple pertinent Opinion - ES lorsque ceux-ci se trouvent dans la même unité de texte, la phrase dans notre cas.

3.4 Les données du projet SENTERRITOIRE

Nous avons sélectionné 300 articles de journaux relatifs à l'aménagement du territoire décrivant l'étang de Thau situé au Sud de la France. À partir de ce corpus, 100 documents (SENT_100) ont été divisés en deux classes d'opinions : positifs et négatifs. L'opinion exprimée est liée à la formation d'une nouvelle agglomération regroupant plusieurs communes (Montpellier Agglomération). Le corpus a été validé par des experts géographes travaillant sur le projet SENTERRITOIRE. Le corpus ne contient pas de textes polarisés neutres.

La section suivante décrit le démonstrateur SenterritoireViewer qui permet de représenter l'ensemble des informations extraites via l'approche OPILAND.

4 L'application SenterritoireViewer pour la représentation des opinions dans le temps et l'espace

Le projet SENTERRITOIRE a pour objectif final de mettre à disposition un outil d'aide à la décision dans le contexte de l'aménagement du territoire basé sur une analyse automatique de textes. Pour cela, nous avons développé l'application Web SENTERRITOIREVIEWER selon l'architecture 3-Tiers (Client, Serveur, Données) car elle donne la possibilité à plusieurs utilisateurs d'interroger l'application en simultané depuis leur poste client via un simple navigateur Web. Côté serveur, le développement est réalisé en Java et l'intégration est réalisée par l'intermédiaire du serveur d'application Tomcat. Le tiers données est structuré en bases de données PostGres/PostGis et il s'appuie sur le moteur géographique Geoserver.

Les opinions des acteurs sont exprimées dans les documents du fonds documentaire selon deux axes sémantiques : le temps et l'espace. L'application doit apporter des informations

sur l'environnement territorial afin d'appuyer les experts du territoire Thau Agglomération (Syndicat Mixte du Bassin de Thau SMBT, chercheurs géographes, etc.) dans leur travail de diagnostic du territoire. Pour cela, l'outil intègre trois composantes principales : une première qui nous donne une vision générale des opinions, une deuxième permettant une lecture spatiale des opinions, et une dernière en offrant une lecture temporelle.

4.1 Composante opinion

La composante *Opinion* permet de représenter l'opinion générale du fonds documentaire exprimée par rapport aux contraintes temporelles et spatiales. Pour cela, la moyenne des opinions résultantes des différentes contraintes (espace spatial et espace temporel sélectionnés) est calculée de façon dynamique. Pour la représentation, nous utilisons un graphique circulaire principal qui est coloré selon la part des pourcentages calculés des opinions : rouge pour négative, jaune pour neutre et vert pour positive. Cinq petits graphiques circulaires sont ajoutées sous le graphique principal pour représenter l'évolution de l'opinion générale dans le temps, de l'opinion la plus ancienne à l'opinion la plus récente exprimées dans un contexte spatial et temporel donné.

La figure 2 décrit un échantillon de documents avec 21% d'opinions négative et 71% d'opinions positives. En passant au-dessus des différentes parties du graphique (colorées en fonction de l'opinion générale), le nombre d'opinions respectives (positives, négatives et neutres) est affiché via une info-bulle.

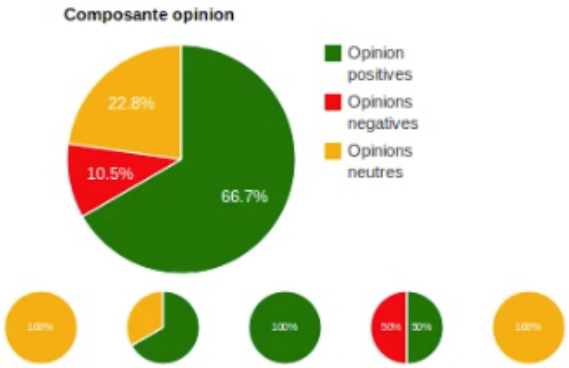

FIG. 2 – *Représentation de la composante Opinion sous forme de graphique circulaire*

Au niveau technique, le module est développé à l'aide du framework ajax ExtJS [4].

4.2 Représentation de l'opinion dans la composante spatiale

La composante spatiale représente les opinions du vocabulaire spécialisé organisées selon la zone géographique décrite. L'application se base sur le moteur cartographique Geoserver qui

4. http ://www.sencha.com/products/extjs

fournit les services cartographiques (**WMS** et **WFS**) qui peuvent être interrogés par notre application. Ces services permettent d'afficher des informations vectorielles ou au format raster, indispensables pour ce type d'application. Un tel service cartographique permet par exemple d'afficher, de colorer ou de modifier de façon dynamique les contours des départements français en fonction des requêtes effectuées par les utilisateurs.

Les fonctionnalités intégrées sont :

- Zoom(s) : 3 outils permettent de zoomer et de dé-zoomer : une loupe pour sélectionner une zone, un "slider" pour augmenter ou diminuer le zoom ou encore le double-clique de la souris ;
- Déplacement : L'utilisateur peut se "déplacer" sur la carte en cliquant et en glissant le curseur dans la direction souhaitée ;
- Interrogation d'objets : L'utilisateur peut cliquer sur une icône représentant un thème (par exemple un hôpital) et une info-bulle s'ouvre dans laquelle sont affichées les différentes opinions relatives à ce thème ;
- Sélection de l'échelle souhaitée : Une liste déroulante permet à l'utilisateur de choisir parmi plusieurs échelles prédéfinies ;
- Sélection / dé-sélection d'informations : L'utilisateur peut choisir par exemple de ne pas afficher les découpages administratifs sur la carte ou de ne pas afficher les icônes représentants les différents thèmes. L'utilisateur peut choisir librement les informations affichées sur la carte de la composante spatiale ;
- Différents fonds de carte : Intégration de différentes couches de fonds de cartes (vue satellite, vue plan etc.). Nous mettons à la disposition des utilisateurs différents fonds de cartes issus de l'API GoogleMaps (Street, Satellite et Hybrid) ainsi que les fonds de cartes Open Source OpenSreetMap.

La figure 3 décrit un exemple de représentation spatiale de l'opinion sur l'ensemble de la zone Thau Agglomération. Concernant les aspects développement et ergonomie (boutons pour zoomer, choix de l'échelle etc.), nous utilisons les technologies GeoExt [5], ExtJS et OpenLayers [6] qui permettent la manipulation et visualisation d'un volume important de données. GeoExt et ExtJS fournissent un ensemble de composants JavaScript graphiques (Widgets) prêts à l'emploi. OpenLayers est une API disponible sous licence libre qui propose différents composants pour créer des cartes dynamiques. Le code est exécuté entièrement côté client par le navigateur.

Pour chaque couple Opinion - Entité Spatiale identifié via l'approche OPILAND, un graphique circulaire est ajouté sur la carte sur l'entité spatiale pointée. Ce graphique circulaire respecte les mêmes caractéristiques que celui représentant l'opinion globale présenté Section 4.1. Par exemple ici, le circulaire vert positionné sur la commune de Pézenas indique qu'un couple "opinion positive - Pézenas" a été identifié dans un document de notre corpus. Si plusieurs couples sont identifiés pour une même ES, un seul circulaire est affiché et regroupe l'ensemble des opinions. La taille de chaque circulaire est alors proportionnelle au nombre d'opinions liées à l'ES pointée. Par exemple dans la figure 3, Une majorité d'opinions posi-

5. http ://geoext.org/
6. http ://www.openlayers.org/

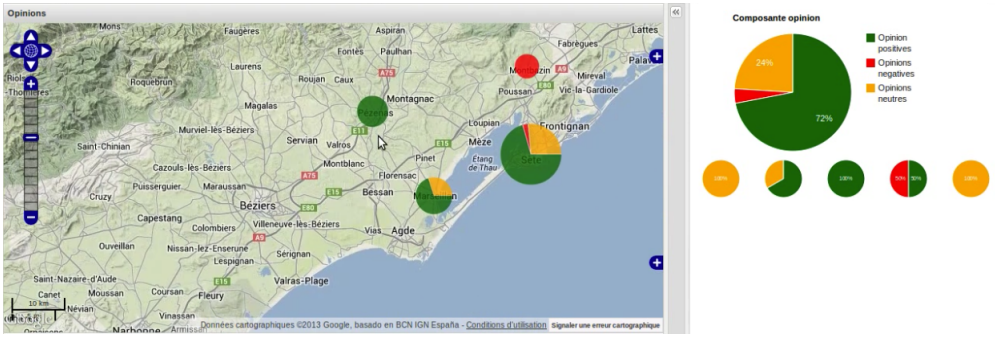

FIG. 3 – *Représentation de l'opinion selon la composante Spatiale*

tives sont liées à la zone bassin de Thau.

Enfin, en utilisant la fonction de zoom, il est possible de sélectionner une zone plus ou moins précise. Les graphiques circulaires décrivant les opinions sur la carte et celui présentant l'opinion globale sont alors re-générés en fonction de la zone sélectionnée. La figure 4 présente la région Thau Agglomération à différentes échelles (de la vue globale à la vue de la commune de Sète).

4.3 Représentation de l'opinion dans la composante temporelle

La composante temporelle est représentée sous forme d'une frise chronologique développée via l'API javascript TimelineJS. Le module permet d'avoir une représentation linéaire d'informations (ici des opinions) dans le temps. De plus, il permet d'associer une opinion sur une échelle graduée. Nous avons intégré la possibilité de spécifier l'échelle du temps, c'est-à-dire de modifier l'unité de temps (jour, mois, année). Par exemple, l'utilisateur pourra zoomer sur la plage de temps qui l'intéresse pour que l'échelle soit graduée au niveau "jour", ou au contraire dé-zoomer pour avoir une échelle graduée au niveau "mois".
La navigation dans le temps est implémentée avec la possibilité de faire défiler la frise de manière horizontale. Il est ainsi possible de faire glisser la frise vers la gauche pour avancer dans le temps, ou vers la droite pour reculer dans le temps. La figure 5 décrit des opinions extraites de notre corpus représentées dans le temps.

Dans la version actuelle du module, chaque document est représenté par une ligne dans le temps avec pour début la première date identifiée dans le contenu texte et pour fin la dernière date. Cela pourra poser des problèmes de lisibilité si l'on souhaite traiter un corpus volumineux. Pour obtenir des informations sur le document dont l'opinion est extraite, un système d'info-bulle est implémenté. A chaque fois que l'utilisateur clique sur un item, une info-bulle affiche la description du document (la phrase extraite du document, dans lequel l'opinion est exprimée et URI du document). Aussi, la composante temporelle intègre les mêmes principes que la composante spatiale pour représenter les opinions (couleurs, graphiques circulaires, regroupement de plusieurs opinions identifiées sur une même date/période, etc.).

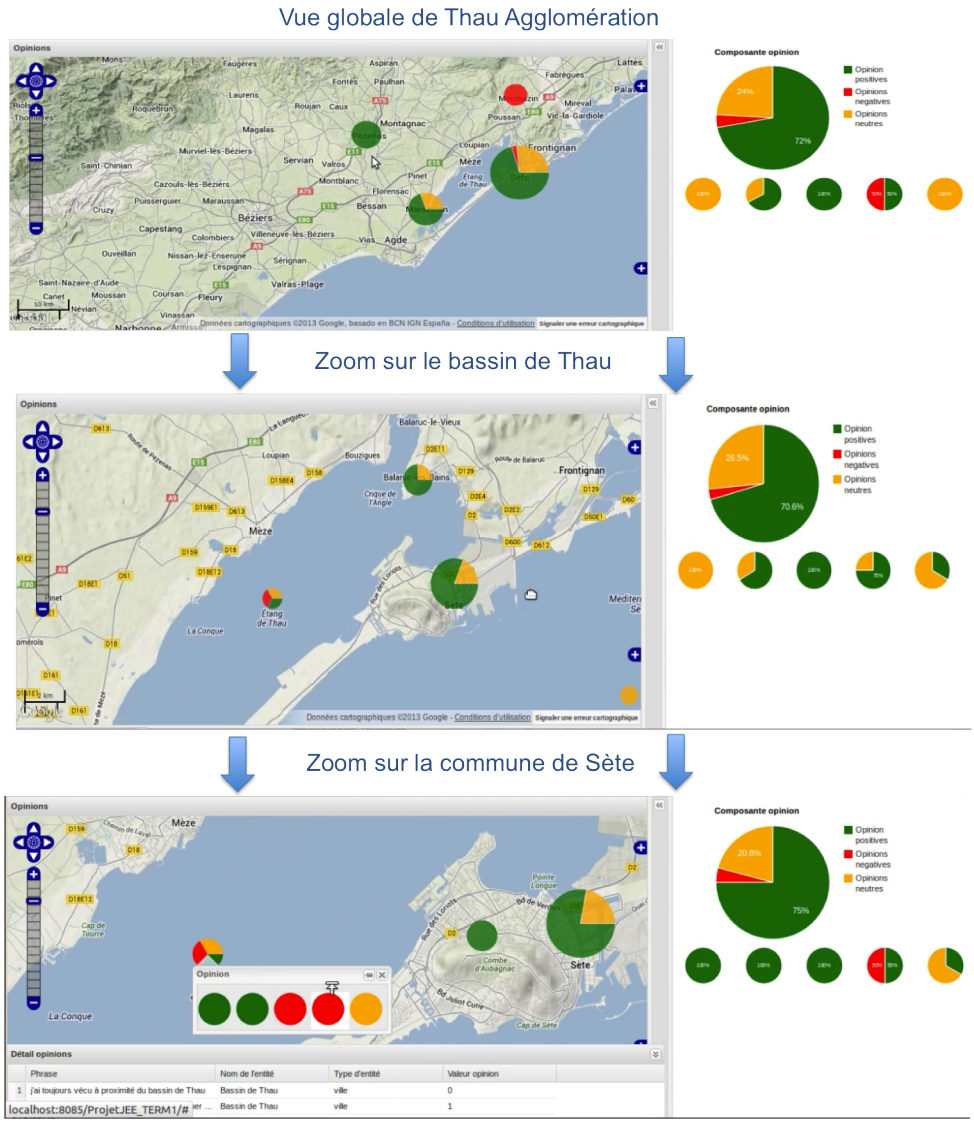

FIG. 4 – *Fonction de zoom sur la zone géographique du bassin de Thau*

4.4 Interaction entre les différentes composantes

Les différentes composantes sont liées dynamiquement entre elles. Plus précisément, les contraintes temporelles spécifiées dans la frise chronologique sont prises en compte par les autres composantes (opinion et spatiale) dans leur affichage respectif. Ainsi, lorsque l'on se déplace dans la frise chronologique, les composantes opinion et spatiale sont automatiquement mises à jour en affichant uniquement les opinions identifiées dans la période concernée. La fi-

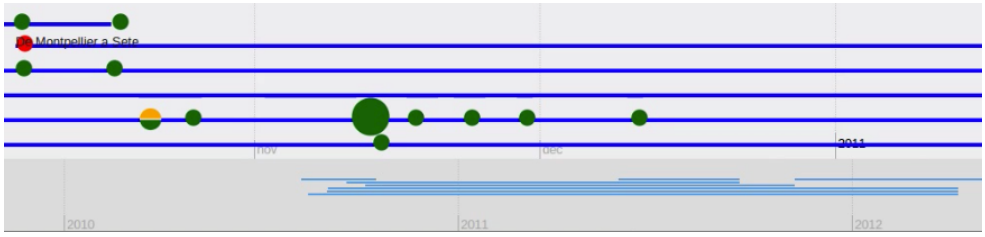

FIG. 5 – *Représentation de l'opinion selon la composante temporelle*

gure 6 décrit l'interaction des trois modules (temps - opinions - espace) intégrés à l'application SenterrritoireViewer.

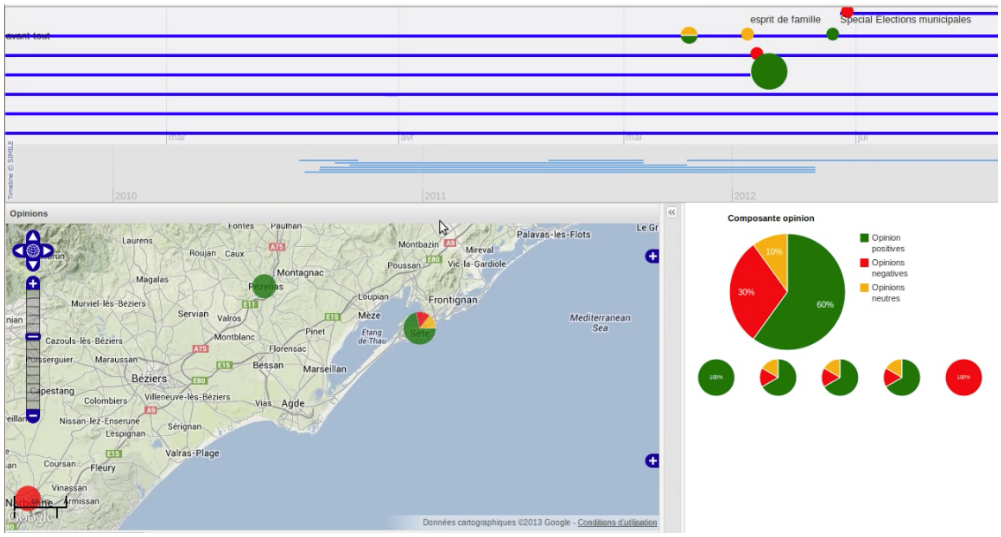

FIG. 6 – *Représentation des composantes Temps-Espace-Opinions de l'application SenterritoireViewer*

L'extrait présenté figure 6 met ainsi en avant les opinions extraites des documents de notre corpus entre mars et juin 2011 sur l'ensemble de la zone géographique de Thau Agglomération.

5 Conclusion et perspectives

L'application de l'approche OPILAND sur un corpus de l'aménagement du territoire de l'étang de Thau a fourni des résultats prometteurs. Nous avons proposé un démonstrateur permettant de représenter des opinions dans le temps et l'espace. L'application permet de naviguer dynamiquement dans le corpus via les composantes spatiales et temporelles afin d'arriver à terme aux documents eux mêmes. Les travaux à venir sont tout d'abord consacrés à évaluer le démonstrateur en le faisant tester par les « experts terrain » du projet Senterritoire, à savoir les

membres du SMBT et un échantillon d'acteurs du territoire de Thau Agglomération. Nous travaillons actuellement à la définition des scénarios d'évaluation avec les géomaticiens du projet SENTERRITOIRE. Nous souhaitons ensuite pouvoir traiter un volume important de données et la couche Geoserver nous garantie la possibilité de le faire pour les composantes spatiales et opinion. Des actions vont être menées au niveau de la composante temporelle pour prendre en compte la montée en charge.

Remerciements

Les auteurs remercient Pierre Maurel (UMR TETIS) pour son expertise sur le corpus de référence, Maguelonne Teisseire (IRSTEA, UMR TETIS), Mathieu Roche (CIRAD, UMR TETIS) pour leur implication dans le projet. Ce travail est soutenu par le labex *NUMEV* et la *Maison des Sciences de l'Homme de Montpellier*.

Références

Abolhassani, M., N. Fuhr, et N. Gövert (2003). Information extraction and automatic markup for xml documents. In *Intelligent Search on XML Data*, pp. 159–178.

Alliès, P. (1980). *L'Invention du territoire*. 184 pages.

Barel, Y. (1981). Modernité, code, territoire, annales de la recherche urbaine. Volume 10-11, pp. 3–21.

BBCTwo (2008). Bbc news white spectrum.

Buléon, P. et G. D. Méo (2005). *L'espace social*. Annales de la recherche urbaine : Armand Colin.

Debarbieux, B. et M. Vanier (2002). *Ces territorialités qui se dessinent*. Datar : Editions de l'Aube, 267 pages.

Deffontaines, J., E. Marcelpoil, et P. Moquay (2001). *Le développement territorial : une diversité d'interprétations*. Maurel, P. Lardon, S., Piveteau, V. (Eds) 184 pages, Paris, Hermès Sciences Publications.

Deleuze, G. et F. Guattari (1980). *Mille plateaux*. Paris : Editions de Minuit. coll. Capitalisme et schizophrénie, 645 pages.

Duthil, B., F. Trousset, M. Roche, G. Dray, M. Plantié, J. Montmain, et P. Poncelet (2011). Towards an automatic characterization of criteria. In *International Conference on Database and Expert Systems Applications (DEXA'11)*, Volume 1, Toulouse, France, pp. 457–465. Springer-Verlag, LNCS.

Esuli, A. et F. Sebastiani (2006). Sentiwordnet : A publicly available lexical resource for opinion mining. In *5th Conference on Language Resources and Evaluation*, pp. 417–422.

Fry, B. (2008). *Visualizing Data* (1st ed.).

Ghitalla, F. (2004). La géographie des agrégats de documents sur le web.

Husaini, M., A. Kocyigit, D. Tapucu, B. Yanikoglu, et Y. Saygin (2012). An aspect-lexicon creation and evaluation tool for sentiment analysis researchers. In *Proc. of the 2012 Eu-*

ropean conference on Machine Learning and Knowledge Discovery in Databases, Volume Part II, pp. 804–807.

Kamvar, S. et J. Harris (2011). We feel fine and searching the emotional web. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, New York, NY, USA, pp. 117–126. ACM.

Kennedy, A. et D. Inkpen (2006). Sentiment classification of movie reviews using contextual valence shifters. In *Computational Intelligence*, Volume 22(2), pp. 110–125.

Kergosien, E., B. Laval, M. Roche, et M. Teisseire (2014). Are opinions expressed in land-use planning documents ? In *International Journal of Geographical Information Science*, Volume à paraître.

Kergosien, E., P. Maurel, M. Roche, et M. Teisseire (2013). Opiter : Fouille de données d'opinion pour les territoires. In *Spatial Analysis and GEOmatics (Sagéo'13)*, Brest, France.

Lacayrelle, A. L. P., M. Gaio, et C. Sallaberry (2007). La composante temps dans l'information géographique textuelle. In *Revue Document Numérique*, Volume 10(2), pp. 129–148.

Liu, B. (2012). Sentiment analysis and opinion mining. pp. 167. Morgan and Claypool Publishers.

Nadeau, D. et S. Sekine (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes 30*(1), 3–26.

Offenhuber, D. (2005). Thought landscape.

Ong, T., H. Chen, W. Sung, et B. Zhu (2005). Newsmap : a knowledge map for online news. *Decision Support Systems 39*(4), 583–597.

Pang, B. et L. Lee (2010). Opinion mining and sentiment analysis. In *Found. and Trends in IR*.

Pecqueur, B. (2007). L'économie territoriale : une autre analyse de la globalisation. In *Alternatives économiques, lÕéconomie politique*, Volume 1 (33), pp. 41–52.

Tahrat, S., E. Kergosien, S. Bringay, M. Roche, et M. Teisseire (2013). Text2geo : from textual data to geospatial information. In *The 3rd International Conference on Web Intelligence, Mining and Semantics (WIMS'13)*.

Turney, P. (2002). Thumbs up or thumbs down ? semantic orientation applied to unsupervised classification of reviews. In *Proc. of ACL*, pp. 417–424.

Vanier, M. (2009). Territoires, territorialité, territorialisation - controverses et perspectives. In *PUR*, pp. 417–424.

Wu, Y., F. Wei, S. Liu, N. Au, W. Cui, H. Zhou, et H. Qu (2010). Opinionseer : Interactive visualization of hotel customer feedback. *IEEE Transactions on Visualization and Computer Graphics 16*(6), 1109–1118.

Summary

A great deal of research on information extraction from textual datasets has been performed in specific data contexts, such as movie reviews, commercial product evaluations, campaign speeches, etc. In the SENTERRITOIRE project, we raise the question on how appropriate these

methods are for documents related to land-use planning. The kind of information sought concerns the stakeholders, opinions, geographic information, and everything else related more generally to the territory. However, it is extremely challenging to link opinions to these kinds of informations. In this article, we present (1) the semi-automatic method called Opiland (OPinion mIning from LAND-use planning documents) combining a NLP process and Text Mining tools in order to extract opinions related to land-use planning, and (2) the associated Web application.

Prédiction au plus tôt d'événements par règles d'épisodes

Lina Fahed, Armelle Brun, Anne Boyer

Université de Lorraine - LORIA - Équipe KIWI
Campus scientifique BP 239 54506 Vandoeuvre-lès-Nancy Cedex
{Lina.Fahed, Armelle.Brun, Anne.Boyer}@loria.fr

Résumé. Le nombre de messages déposés quotidiennement sur les réseaux sociaux et les blogs est colossal. Ces messages constituent une source d'informations précieuse. Leur exploration se place dans le domaine de la fouille de données temporelles. Cette fouille peut être utilisée dans un but de prédiction afin d'anticiper l'apparition de certaines informations. Notre objectif dans cet article est de proposer un algorithme permettant la prédiction d'informations au plus tôt et de façon fiable.

Dans ce travail préliminaire, nous nous focalisons sur l'extraction de règles d'épisodes minimales, pouvant être identifiées rapidement dans un flux de données. Ces règles auront aussi une conséquence éloignée temporellement de l'antécédent, pour anticiper au plus tôt la conséquence, tout en ayant une confiance élevée.

1 Introduction

Avec l'émergence du web 2.0, les internautes ne sont plus de simples consommateurs, ils sont également acteurs par le biais des messages qu'ils peuvent déposer, des tags ou commentaires qu'ils peuvent laisser et de toute action qu'ils peuvent effectuer. Dans ce cadre, les messages laissés dans les réseaux sociaux représentent une source importante et précieuse d'informations, que de nombreuses recherches cherchent à analyser dans le but d'en comprendre le contenu, d'en extraire les relations cachées, mais aussi de prédire de l'information.

Le flux de données (de messages) peut être considéré comme une séquence ordonnée par la date de création des messages. On appellera "item" un élément représentant un message (mot du message, opinion ou sujet extrait, etc.). À un temps t donné, plusieurs items apparaissent donc dans cette séquence : l'ensemble des items du message créé au temps t. Ce type de séquence est appelée "séquence complexe".

La fouille de données temporelles concerne la fouille de données séquentielles ordonnées selon certains critères comme par exemple le temps ou la position (Laxman et Sastry, 2006). Par conséquent, l'analyse du flux de messages issus des réseaux sociaux peut être vue comme une tâche de fouille de données temporelles. Les traces de navigation (Priya et Vadivel, 2012), les séquences biologiques (Bathoorn et al., 2010), ou les rapports météo (Basak et al., 2004), peuvent représenter d'autres exemples de données temporelles.

L'extraction de motifs (et surtout de motifs "séquentiels" dans lesquels l'ordre est considéré) est une tâche essentielle dans la fouille de données temporelles pour extraire des items

corrélés et des relations (ordonnées) entre items. Dans certains cas, les motifs sont utilisés pour former des règles d'association qui peuvent être exploitées pour prédire des informations : la conséquence des règles (Agrawal et al., 1994).

Une sous-tâche de l'extraction de motifs, adaptée aux données séquentielles représentée sous la forme d'une unique séquence, est l'extraction d'épisodes. Un épisode est un motif temporel composé d'items "relativement proches", qui apparaît souvent tout au long de la séquence ou sur une partie de cette séquence (Mannila et al., 1997). L'extraction d'épisodes diffère de celle de motifs séquentiels par la mesure de fréquence des épisodes. À l'opposé des premiers algorithmes d'extraction d'épisodes, qui extrayaient un ensemble complet d'épisodes, certains algorithmes récents visent à extraire, pour des raisons de complexité, un sous-ensemble d'épisodes respectant certaines caractéristiques. Par exemple, nous pouvons citer l'ensemble des épisodes fermés (Tatti et Cule, 2012) ou des épisodes maximaux (Iwanuma et al., 2005). Tout comme il est possible d'extraire des règles d'association à partir de motifs, des règles d'épisodes peuvent être extraites à partir d'épisodes. Ces règles d'épisodes peuvent aussi être utilisées dans un objectif de prédiction (Daurel, 2003).

La majorité des règles d'épisodes construites à partir d'épisodes fermés ou maximaux ont la caractéristique d'avoir un antécédent long (composé de nombreux items). De notre point de vue, elles ne sont pas adaptées à la prédiction au plus tôt d'informations. En effet, pour prédire une conséquence, il faut attendre l'apparition de la totalité des items de l'antécédent. Cependant, notre objectif est d'anticiper, au plus tôt, des événements. Dans ce cas, pour détecter rapidement des événements, une règle d'épisodes avec un antécédent plus petit est plus pertinente qu'une règle avec un antécédent plus long. Par conséquent, notre premier objectif est d'extraire des règles d'épisodes composées d'un antécédent de taille minimale, que nous appellerons "règles d'épisodes minimales" .

Dans la tâche d'identification au plut tôt d'informations, nous faisons l'hypothèse qu'il est inutile de chercher à former des règles d'épisodes très complexes ou très précises, mais des règles d'épisodes permettant d'identifier les premiers signaux. Nous faisons donc l'hypothèse qu'il est inutile d'extraire les règles d'épisodes contenant plusieurs fois le même item. Il est vrai qu'une apparition multiple d'un item porte plus d'informations, mais cette caractéristique permettra de diminuer la complexité de l'algorithme. Nous appelons "épisode sans répétition" un épisode composé d'items uniques. Notre second objectif est donc d'extraire des épisodes sans répétition.

Nous souhaitons pouvoir anticiper/prédire des informations "lointaines". Notre troisième objectif est donc d'extraire des règles d'épisodes ayant une conséquence éloignée temporellement de l'antécédent.

La suite de cet article est organisée comme suit : la section 2 présente plusieurs travaux de l'état de l'art sur la fouille de données temporelles et en quoi ils ne peuvent répondre à notre besoin. Dans la section 3, nous définissons quelques notions qui seront utilisées dans notre algorithme et la section 4 détaille l'algorithme EMMA, dont s'inspire notre travail. Notre algorithme est présenté dans la section 5. Nous conclurons et fournirons quelques perspectives dans la section 6.

2 État de l'art

Les méthodes d'extraction de motifs varient en fonction des données disponibles. Dans le cas de données transactionnelles, premier type de données auxquelles s'est intéressée la fouille de données, des motifs non séquentiels (itemsets) sont extraits de l'ensemble des transactions. Plusieurs algorithmes sont très populaires dans l'état de l'art, comme Apriori (Agrawal et al., 1994) et FPGrowth (Han et al., 2000).

Dans le cas de données séquentielles, dans lesquelles l'ordre entre les transactions est considéré, des motifs séquentiels sont extraits. Les algorithmes sont en règle générale une adaptation des algorithmes d'extraction d'itemsets, comme prefixspan (Han et al., 2001). Plusieurs méthodes proposent l'extraction de règles d'association à partir de motifs (règles d'association séquentielles dans le cas de données séquentielles). Une règle d'association est constituée de deux parties : un antécédent et une conséquence, et signifie que si l'antécédent est vrai, alors la conséquence est vraie. Pour mesurer l'efficacité de règles d'association, deux mesures principales sont utilisées : le support et la confiance (Agrawal et al., 1994), où le support est le nombre de transactions contenant à la fois l'antécédent et la conséquence, et la confiance est la probabilité d'apparition de la conséquence lorsque l'antécédent est apparu.

Dans le cas de données composées d'une unique longue séquence de données ordonnée par le temps, les motifs extraits sont des épisodes et des règles d'épisodes peuvent en être déduites (Mannila et al., 1997). À la différence des motifs séquentiels, dans lesquels seul l'ordre est utilisé, les méthodes d'extraction d'épisodes utilisent le temps d'apparition. De plus, la fréquence d'apparition des motifs séquentiels est évaluée comme étant le nombre de transactions dans lesquelles le motif est apparu, alors que la fréquence d'apparition des épisodes est évaluée en fonction du nombre d'occurrences de l'épisode dans l'unique séquence de données.

Mannila (Mannila et al., 1997) a proposé les premiers algorithmes d'extraction d'épisodes : $Winepi$ et $Minepi$ qui seront la base de nombreux autres algorithmes proposés par la suite. Les deux algorithmes extraient dans un premier temps les épisodes les plus petits, puis cherchent itérativement des épisodes plus grands en se basant sur leur fréquence. Les deux algorithmes diffèrent par la méthode utilisée pour compter le nombre d'occurrences des épisodes. $Winepi$ fait glisser une fenêtre de largeur w et compte le nombre de fenêtres qui contiennent l'épisode. $Minepi$ n'utilise pas la notion de fenêtre mais cherche les intervalles les plus petits de l'épisode qui ne contiennent aucune autre occurrence du même épisode ou d'un de ses sous-épisodes, ce qui est appelé "l'occurrence minimale d'un épisode". Ces méthodes ont la caractéristique d'extraire un ensemble complet d'épisodes.

La fouille d'épisodes dans des séquences de données complexes est une problématique récente. Elle nécessite un algorithme adapté qui doit prendre en compte l'existence de plusieurs items à chaque temps afin d'extraire les épisodes. Huang et Chang (Huang et Chang, 2008) proposent un algorithme appelé EMMA qui extrait un ensemble complet d'épisodes à partir d'une séquence complexe. Dans les deux premières phases, EMMA pré-traite la séquence complexe : il extrait un ensemble de motifs fréquents de chaque temps. À partir de ces motifs, les épisodes sont construits itérativement pendant une troisième phase. EMMA est un algorithme rapide et facile à adapter. La communauté d'extraction de motifs fréquents admet que dans la plupart des applications, il n'est pas nécessaire d'extraire tous les épisodes mais qu'il est suffisant d'extraire un ensemble d'épisodes condensé et significatif (Pasquier et al., 1999). En effet, non seulement les algorithmes qui extraient un ensemble complet d'épisodes sont

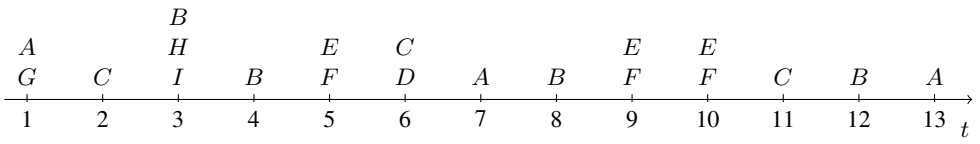

FIG. 1 – *Séquence complexe.*

complexes, mais de plus la taille de l'ensemble résultat affecte leur bonne compréhension et leur analyse. Afin d'avoir un ensemble d'épisodes plus concis et pertinent et un algorithme de plus faible complexité, les travaux récents se focalisent sur la détection d'épisodes comportant certaines caractéristiques et contraintes de façon à rendre cet ensemble plus significatif. Cette notion varie selon les auteurs. Cependant, dans la plupart des travaux, elle a pour but d'extraire un ensemble d'épisodes fermés (Tatti et Cule, 2012) ou maximaux (Iwanuma et al., 2005).

Certains travaux ont proposé d'extraire un ensemble condensé d'épisodes dans une séquence complexe, en impliquant certaines caractéristiques comme le fait que les épisodes soient maximaux (Gan et Dai, 2011) ou approximativement fermés et non dérivables (Gan et Dai, 2012). Cependant, dans ces épisodes, le même item peut apparaître plusieurs fois. Pour certaines applications, cette caractéristique peut rendre les règles non pertinentes, et on peut ne pas souhaiter l'apparition multiple d'un item dans un épisode. Dans ce cas, les algorithmes proposés dans ces deux travaux sont non applicables, et cette caractéristique ne peut pas être modifiée parce qu'elle représente le cœur des algorithmes.

Comme nous l'avons mentionné dans l'introduction, nous voulons extraire les règles d'épisodes ayant certaines caractéristiques, comme le fait d'être minimales. La notion "minimale" n'est pas nouvelle dans l'état de l'art. Elle a été proposée par Rahal (Rahal et al., 2004) dans le cadre d'extraction de règles d'association fiables, peu fréquentes ayant l'antécédent le plus petit et la conséquence fixée par l'utilisateur à partir de données transactionnelles dans le but de prédire au plus tôt la conséquence. Cette approche est différente de notre travail. En effet, nous voulons extraire les règles d'épisodes (et non pas des règles d'association). De plus, nous cherchons les règles d'épisodes sans répétition et avec une conséquence temporellement éloignée de l'antécédent.

Nous présentons maintenant quelques définitions, appuyées par un exemple de données, qui seront utiles pour la suite de cet article.

3 Définitions et exemples

Séquence complexe : Soit I un ensemble fini d'items de taille m. I_t est l'ensemble des items qui apparaissent au temps t. S est une suite ordonnée de paires (temps t, item i_t), appelée une séquence de données. S est qualifiée de séquence complexe quand apparaît au temps t un ensemble d'items I_t (voir figure 1). $S = < (t_1, I_{t_1}), (t_2, I_{t_2}), ..., (t_n, I_{t_n}) >$ avec $t_1 < t_2 < ... < t_n$. Le nombre de paires dans la séquence S est n .

Fenêtre : Dans l'extraction d'épisodes à partir d'une séquence S, la notion de fenêtre est souvent utilisée pour représenter un segment de la séquence. Notons $Win(S, st, w)$ le segment

dans la séquence S qui commence au temps st (start) et qui a la longeur w. Certains algorithmes d'extraction d'épisodes utilisent des fenêtres glissantes où st varie de 1 à n. Dans la figure 1, une fenêtre glissante de longeur 6 se décale tout au long de la séquence S, portant sur les intervalles suivants : ([1,6], ... ,[11,13],[12,13],[13,13]).

Épisode, sous-épisode, épisode sans répétition, concaténation :

Un épisode $P =< p_1, p_2, ..., p_k >$ sur I^k est une suite ordonnée d'ensembles d'items. Soit l'épisode $Q =< q_1, q_2, ..., q_s >$. P est un sous-épisode de Q si chaque $p_i \subseteq q_i$ (p_i peut être $\emptyset$). Par exemple, l'épisode $P = (< A >, < BD >)$ est un sous-épisode de $Q = (< AF >, < BD >)$. Au contraire, l'épisode $R = (< AB >)$ n'est pas un sous-épisode de Q. Dans certains cas, il est possible de trouver $p_i \subseteq p_j$ où $i \neq j$, ce qui signifie une apparition multiple d'un item dans le même épisode. Par exemple, dans l'épisode $(< A >, < AF >)$ l'item A apparaît plusieurs fois. Nous appelons "épisode sans répétition" un épisode qui ne contient pas d'items apparaissant plusieurs fois. Un épisode P peut apparaître plusieurs fois dans la séquence S et il peut être considéré comme un sous-épisode de la séquence S. On dit que $Win(S, st, w)$ contient $P =< p_1, ..., p_k >$ si p_1 est un sous-ensemble de I_{st} et le reste de P est contenu dans $< I_{st+1}, ..., I_{st+w} >$. Le nombre de fenêtres dans lesquelles P apparaît s'appelle le support de P. P est fréquent dans la séquence S si $support(P) \geq minsupp$ où $minsupp$ est le seuil minimum de support pré-déterminé.

Soient les deux épisodes P, Q. La concaténation de P et Q est $P.Q =< p_1, ..., p_k, q_1, ..., q_s >$ de taille $k + s$.

Règle d'épisodes : Soit un épisode fréquent $R = P.Q$. À partir de R, une règle d'épisodes $P \rightarrow Q$ signifie que Q apparaît après P avec une probabilité élevée. Nous appelons "règle d'épisodes sans répétition" une règle $P \rightarrow Q$ si $P.Q$ est un épisode sans répétition.

Mesures de règles d'épisodes : Les mesures d'évaluation de règles d'épisodes sont globalement les mêmes que pour les règles d'association. Le $support$ représente la fréquence de la règle et la $confiance$ représente la probabilité d'apparition de Q après P. Nous définissons le support et la confiance de la règle $P \rightarrow Q$ pour une séquence de taille n comme suit :

$$support(P \rightarrow Q) = \frac{support(P.Q)}{n} \qquad confiance(P \rightarrow Q) = \frac{support(P.Q)}{support(P)} \tag{1}$$

$(P \rightarrow Q)$ est considérée **fréquente** si $support(P \rightarrow Q) \geq minsupp$, et **fiable** si $confiance(P \rightarrow Q) \geq minconf$. Nous appelons **règle acceptable** une règle fréquente et fiable. Par exemple, l'épisode $< A >< B >$ a un nombre d'occurrences de 4 dans 2 fenêtres de largeur $w = 6$. Comme le support d'un épisode représente le nombre de fenêtres dans lesquelles il apparaît (pas l'occurrence de l'épisode), donc $support(< A >< B >) = 2/13 = 15.3\%$. À partir de cet épisode, une règle d'épisodes $RE =< A >\rightarrow< B >$ peut être construite. Le support de RE est $2/13 = 15.3\%$ et la confiance $2/3 = 66\%$. Pour un $minsupp = 15\%$ et $minconf = 60\%$, RE est fréquente et fiable et donc acceptable.

Liste de bornes : Une borne d'un épisode P dans la séquence S est un intervalle $[t_s, t_e]$, où $(t_e - t_s) < w$ et où p_1 (premier élément de P) est un sous-épisode de I_{t_s} (ensemble des items apparaissant au temps t_s), p_k (dernier élément de P) est un sous-épisode de I_{t_e}, et

$p_2, ...p_{k-1}$ est un sous-épisode de $(I_{t_s+1}, ..., I_{t_e-1})$. La liste de bornes $Borne(P)$ de P est la liste de tous les intervalles dans S dans lesquels P apparaît, et $Borne_i(P)$ est la $i^{ème}$ borne de P. Dans la figure 1, l'épisode $< A >$ a la liste de bornes $[1, 1], [7, 7], [13, 13]$, et l'épisode $< A >< B >$ a la liste de bornes $[1, 3], [1, 4], [7, 8], [7, 12]$ pour $w = 6$. Les bornes permettent de prendre en compte des mesures de temps et de distance entre items. Notons que l'ensemble final d'épisodes ne contient pas ces mesures, seul l'ordre est représenté.

Règle d'épisodes minimale : Soit la règle d'épisodes $RE = (P \rightarrow Q)$. Si RE est acceptable, RE est minimale s'il n'existe pas de règle $RE' = (R \rightarrow Q)$ acceptable où $R \subset P$.

Nous présentons maintenant l'algorithme de l'état de l'art EMMA, dont est inspiré notre algorithme.

4 Algorithme existant : EMMA

EMMA (Episode Mining using Memory Anchor) (Huang et Chang, 2008) est un algorithme d'extraction d'épisodes fréquents dans une séquence complexe. Il a la caractéristique d'extraire un ensemble complet d'épisodes. EMMA extrait d'abord les motifs fréquents, pour chaque temps t, dans l'ensemble des items apparaissant à t. Cette première caractéristique permet de diminuer la complexité de l'algorithme en diminuant la grande taille des motifs extraits. Les épisodes sont ensuite construits incrémentalement en prenant en compte la liste de bornes de chaque 1-uplet épisode. L'algorithme EMMA comporte trois phases résumées dans les sous-sections suivantes.

4.1 Extraction de motifs fréquents

EMMA utilise un algorithme appelé FIMA (Frequent Itemset mining using Memory Anchor) pour extraire un ensemble de motifs fréquents locaux : LFP (Local Frequent Patterns). Tous les items dans un motif fréquent ayant le même temps d'apparition t, le motif est dit "local". FIMA commence par une lecture de la séquence de données pour trouver les motifs locaux de taille 1. Ensuite, pour chaque motif local, une combinaison est faite avec un autre motif du même temps t. Cette procédure de combinaison est appliquée récursivement afin d'avoir des motifs plus longs. À chaque itération, les motifs sont choisis en fonction de leur support. Ensuite, une liste de bornes de chaque motif local fréquent est créée. Notons qu'un motif fréquent extrait représente un 1-uplet épisode.

4.2 Encodage de la séquence de données

Une fois l'ensemble des 1-uplet épisodes extrait, un identifiant id est associé à chacun (voir tableau 1). La séquence de données (figure 1) est encodée par les id de 1-uplet épisodes dans chaque intervalle selon leur liste de bornes (voir tableau 2). L'ensemble des id est appelé ID.

4.3 Phase principale : Extraction des épisodes fréquents

La dernière étape de l'algorithme EMMA extrait les épisodes fréquents en utilisant la séquence de données encodée. Chaque id représente un préfixe d'un épisode à construire (le

id	1-uplet épisode	Liste de bornes
#1	A	[1,1], [7,7], [13,13]
#2	B	[3,3], [4,4], [8,8], [12,12]
#3	C	[2,2],[6,6], [11,11]
#4	E	[5,5], [9,9], [10,10]
#5	F	[5,5], [9,9], [10,10]
#6	EF	[5,5], [9,9], [10,10]

TAB. 1 – *Encodage des* 1*-uplet épisodes.*

Temps	1	2	3	4	5	6	7	8	9	10	11	12	13
id	#1	#3	#2	#2	#4 #5 #6	#3	#1	#2	#4 #5 #6	#4 #5 #6	#3	#2	#1

TAB. 2 – *Encodage de la séquence de données.*

premier élément de l'épisode). Pour chaque id et selon sa liste de bornes, on cherche les autres id fréquents dans sa liste de bornes projetées définie comme suit :

Liste de bornes projetées : Pour une borne d'un épisode P, une borne projetée représente l'intervalle de temps (de taille au plus w) où les épisodes candidats (super-épisodes de P) sont recherchés. Pour l'épisode P, qui a sa $i^{ème}$ borne $Borne_i(P) = [t_{s_i}, t_{e_i}]$, la borne projetée de $Borne_i(P)$ est $ProjBorne_i(P) = [t_{s_i} + 1, t_{s_i} + w - 1]$. Par exemple, pour $w = 6$, le 1-uplet épisode $\#1$ a la liste de bornes projetées suivante : $ProjBorne(\#1) = \{[2, 6], [8, 12]\}$.

Un nouvel épisode candidat est alors construit en étendant chaque id avec chacun des id fréquents dans sa liste de bornes projetées. Dans ce cas, $\#1$, $\#2$, $\#3$, $\#4$, $\#5$, $\#6$ qui sont fréquents sont tous étendus, pour $minsupp = 15\%$ (tableau 2). Pour chaque épisode candidat, sa liste de bornes est ensuite calculée. Si cet épisode est fréquent, il est de nouveau étendu avec un nouvel id. Par exemple, pour construire l'épisode $< \#1, \#2 >$, sa liste de bornes est $Borne(< \#1, \#2 >) = [1, 3], [1, 4], [7, 8], [7, 12]$. $support(< \#1, \#2 >) = 2/13 = 15.3\%$, donc $< \#1, \#2 >$ est un épisode fréquent et il sera étendu avec d'autres id. EMMA utilise un parcours en profondeur pour étendre les épisodes, et l'itération s'arrête quand l'épisode candidat n'est pas fréquent ou s'il n'y a plus d'id avec lequel l'étendre. Cette façon de construire les épisodes peut mener à des épisodes contenant plusieurs occurrences du même item.

5 Notre approche : extraction de règles d'épisodes minimales

Rappelons que notre objectif est de pouvoir anticiper au plus tôt des événements. Pour cela, nous proposons un algorithme d'extraction de règles d'épisodes qui, en plus d'être fréquentes et fiables comme celles extraites par les algorithmes de l'état de l'art, sont sans répétition, minimales et avec une conséquence éloignée temporellement de l'antécédent.

Les règles d'épisodes extraites par notre algorithme sont minimales, c'est-à-dire avec un antécédent le plus petit possible, car nous partons de l'hypothèse que plus l'antécédent est petit, plus tôt nous l'identifierons dans une séquence, et plus tôt nous pourrons prévoir/anticiper la conséquence. À notre connaissance, cela est contraire aux algorithmes d'extraction de règles d'épisodes de l'état de l'art dont l'objectif est d'extraire des règles d'épisodes maximales ou les plus longues possibles tout en respectant certaines caractéristiques.

Les règles d'épisodes extraites par notre algorithme ont également la caractéristique d'avoir une conséquence la plus éloignée possible de l'antécédent, ce qui permettra de détecter les événements lointains.

Les règles d'épisodes extraites ont enfin la caractéristique d'être sans répétition car nous considérons que cette unicité permettra de réduire la complexité de l'algorithme.

Nous proposons une approche novatrice pour extraire des règles d'épisodes avec les caractéristiques souhaitées. Notre algorithme utilise les mêmes étapes d'initialisation et d'encodage qu'EMMA. Nous différons d'EMMA en deux points : notre algorithme extrait des règles d'épisodes, au contraire d'EMMA qui extrait des épisodes. De plus, grâce aux caractéristiques souhaitées, notre algorithme extrait un ensemble réduit pas complet de règles d'épisodes, au contraire d'EMMA qui extrait un ensemble complet d'épisodes.

Le principe de notre algorithme est le suivant : chaque règle d'épisodes est construite en fixant un 1-uplet épisode comme préfixe de la future règle, puis en fixant la conséquence, qui est la plus éloignée possible de ce préfixe, et enfin en complétant l'antécédent avec des 1-uplet épisodes les plus proches possibles du préfixe. L'algorithme 1 présente notre proposition. Nous détaillons ci-dessous cet algorithme.

5.1 Déroulement de l'algorithme

Identification du préfixe : Chaque id obtenu dans la deuxième étape d'EMMA représente un préfixe d'un épisode potentiel. Soit id_i un préfixe, nous construisons $P_{roj}ID_{fin}(id_i)$ la liste des id présents à la fin des bornes projetées de id_i (dernières positions des bornes projetées de id_i). Ces id sont stockés dans cette liste dans l'ordre décroissant du nombre de bornes dans lesquelles ils apparaissent. Nous construisons également $P_{roj}ID_{deb}(id_i)$, la liste des id présents au début des bornes projetées de id_i. Les id sont stockés dans cette liste dans l'ordre décroissant du nombre de bornes dans lesquelles ils apparaissent. La taille des intervalles "début" et "fin" est une proportion de w. Cette représentation nous permettra d'accélérer la construction des règles d'épisodes minimales car elle va éviter des itérations inutiles.

Par exemple, soit la taille de l'intervalle égale à 2, alors pour $\#1$, c'est $\#3$ qui est apparu le plus souvent à la fin des bornes projetées de $\#1$, il est apparu 2 fois. Ensuite, $\#2, \#4, \#5, \#6$ ne sont apparus qu'une seule fois à la fin des bornes projetées de $\#1$. Donc, $P_{roj}ID_{fin}(\#1)=$ $(\#3, \#2, \#4, \#5, \#6)$. C'est cette liste qui va nous servir pour identifier les conséquences de toutes les règles d'épisodes qui auront $\#1$ comme préfixe.

Identification de la conséquence : id_i est étendu avec les éléments de $P_{roj}ID_{fin}(id_i)$ qui représentent une conséquence potentielle d'un épisode dont le premier élément est id_i. id_i est étendu avec id_j (qui sera la conséquence de la règle), formant ainsi un épisode candidat $< id_i, id_j >$. La fréquence de cet épisode candidat est calculée en utilisant le nombre de bornes dans sa liste de bornes. Si $< id_i, id_j >$ n'est pas fréquent, on arrête cette itération et on

considère que id_j ne peut pas être une conséquence de id_i. On itère de nouveau pour étendre id_i avec d'autres $id \in P_{roj}ID_{fin}(id_i)$. Si $< id_i, id_j >$ est fréquent, alors la règle $id_i \rightarrow id_j$ est construite. Si elle est fiable, elle est considérée comme acceptable. Elle n'est donc plus étendue et elle est ajoutée à la liste finale des règles d'épisodes LRE. En effet, elle est non seulement fiable mais en plus elle est minimale, elle remplit donc notre objectif. Si la règle est fréquente mais pas fiable, alors le 2-uplet épisode $< id_i, id_j >$ est étendu pour compléter l'antécédent, comme suit :

Complétion de l'antécédent : Rappelons que nous cherchons à compéter l'antécédent de façon à obtenir non seulement une règle fiable mais aussi une règle minimale (avec un antécédent minimal). Les itérations de complétion de l'antécédent s'arrêtent donc dès que la règle d'épisodes est $fiable$, ou s'il n'y a plus d'épisode candidat.

À partir de $< id_i, id_j >$, id_i est étendu avec les id dans sa liste $P_{roj}ID_{deb}(id_i)$ itérativement. Soit $id_s \in P_{roj}ID_{deb}(id_i)$, si l'épisode candidat $< id_i, id_s, id_j >$ n'est pas fréquent, alors le préfixe de l'antécédent id_i est étendu avec un autre $id \in P_{roj}ID_{deb}(id_i)$, sinon la confiance de la règle $id_i, id_s \rightarrow id_j$ est calculée. Si la règle n'est pas fiable, alors l'épisode $< id_i, id_s >$ est étendu avec d'autres $id \in P_{roj}ID_{deb}(id_i, id_s)$, sinon elle est acceptable, et par définition minimale, son antécédent n'est donc plus étendu et elle est ajoutée à la liste LRE.

Épisode sans répétition : Comme nous l'avons mentionné, notre algorithme a pour but d'éviter l'apparition multiple d'un item dans les règles d'épisodes. À chaque fois qu'un épisode doit être étendu, une vérification est faite pour éviter de l'étendre avec des id ayant des items en commun, ce qui aura également l'avantage de réduire la complexité de l'algorithme.

En résumé : Notre algorithme permet non seulement d'obtenir des règles avec les trois caractéristiques souhaitées, mais il est également moins complexe qu'EMMA. Dans la plupart des algorithmes d'extraction d'épisodes, la complexité est dépendante de la taille de l'épisode. En extrayant les règles d'épisodes portant nos caractéristiques, l'algorithme consomme théoriquement moins de temps et de mémoire, ce qui signifie pour nous une complexité diminuée, par rapport aux autres algorithmes de l'état de l'art. En effet, en cherchant à extraire des règles d'épisodes minimales, sans répétition et avec une conséquence éloignée, les stratégies choisies pour former ces règles permettent d'éviter plusieurs itérations. Par exemple, les stratégies de recherche de la conséquence et de la construction de l'antécédent (utilisant les deux listes de début et de fin dans les bornes projetées) rendent l'algorithme moins complexe parce qu'elles réduisent les intervalles de recherche des id candidats.

5.2 Exemple

Reprenons l'exemple précédent. Nous avons $P_{roj}ID_{fin}(\#1)$=$(\#3, \#2, \#4, \#5, \#6)$, et $P_{roj}ID_{deb}(\#1)$ =$(\#2, \#3, \#4, \#5, \#6)$. Nous commençons par la phase d'identification de la conséquence du préfixe $\#1$, en l'étendant avec les id dans sa liste $P_{roj}ID_{fin}(\#1)$. Pour la conséquence candidate $\#3$, $< \#1, \#3 >$ est fréquent parce que $support(< \#1, \#3 >) = 15.3\% > minsupp\%$, donc nous fixons la conséquence $\#3$ pour le préfixe $\#1$. Ensuite, la règle $\#1 \rightarrow \#3$ est construite et sa confiance est calculée : $confiance(\#1 \rightarrow \#3) = 66\% <$

$minconf$, donc elle n'est pas fiable, dans ce cas il faut procéder à l'étape de complétion de l'antécédent afin de construire une règle acceptable et minimale.

Nous complétons donc $\#1$ avec $\#2$ (le premier id dans $P_{roj}ID_{deb}(\#1)$). $support(< \#1, \#2, \#3 >) = 15.3\% > minsupp$, donc l'épisode $< \#1, \#2, \#3 >$ est fréquent. $confiance(\#1, \#2 \rightarrow \#3) = 100\% > minconf$. Donc, la règle $\#1, \#2 \rightarrow \#3$ est acceptable et donc minimale, et elle est ajoutée à LRE. Le préfixe $\#1$ qui a une conséquence fixée $\#3$ est encore étendu avec d'autres id dans sa liste $P_{roj}ID_{deb}(\#1)$ pour compléter l'antécédent afin d'extraire d'autres règles d'épisodes minimales. Selon l'étape de vérification d'apparition multiple d'un item dans un épisode, $\#3 \in P_{roj}ID_{deb}(\#1)$ ne peut pas être un complément candidat de l'antécédent de la règle $\#1 \rightarrow \#3$. Lorsque l'on complète cette règle avec les id $\#4, \#5, \#6$ de $P_{roj}ID_{deb}(\#1)$, les règles construites ne sont pas acceptables et donc pas minimales. Nous procédons à une autre itération d'identification d'une autre conséquence de $\#1$. Cette procédure est répétée pour chaque $id \in P_{roj}ID_{fin}(\#1$.

Algorithme 1 : Extraction de règles d'épisodes minimales

Données : S : séquence complexe, LFP : motifs fréquents locaux, ID : liste des id
Résultat : LRE : liste de règles d'épisodes minimales
Procédure *extraction de règles épisodes minimales* **est**
  pour chaque $id_i \in LFP$ **faire**
    Construire $P_{roj}ID_{fin}(id_i), P_{roj}ID_{deb}(id_i)$
    `conséquence(`id_i`)`

Procédure *conséquence*(id_i) ; `/* `id_i`: préfixe */`
est
  pour chaque $id_j \in P_{roj}ID_{fin}(id_i)$ **faire**
    si $id_i \bigcap id_j = \emptyset$; `/* vérifier l'apparition multiple */`
    alors
      si $< id_i, id_j >$ *est fréquent* **alors**
        si $id_i \rightarrow id_j$ *est fiable* **alors**
          Ajouter $id_i \rightarrow id_j$ à LRE
        sinon
          `antécédent(`id_i, id_j`)`

Procédure *antécédent*(id_i, id_j) ; `/* `id_i`: préfixe, `id_j`: conséquence */`
est
  pour chaque $id_s \in P_{roj}ID_{deb}(id_i)$ **faire**
    si $id_i \bigcap id_s \bigcap id_j = \emptyset$; `/* vérifier l'apparition multiple */`
    alors
      si $< id_i, id_s, id_j >$ *est fréquent* **alors**
        si $id_i, id_s \rightarrow id_j$ *est fiable* **alors**
          Ajouter $id_i, id_s \rightarrow id_j$ à LRE
        sinon
          `antécédent(`$< id_i, id_s >, id_j$`)`

6 Conclusion et perspectives

Dans ce travail préliminaire, nous proposons un algorithme de fouille de flux de données dans le but de prédire au plus tôt et de façon fiable, des événements. Notre objectif à court terme est de fouiller des messages issus des réseaux sociaux et de prédire l'apparition d'informations. Nous avons mis en évidence que l'extraction de règles d'épisodes est adaptée à nos données, qui se présentent sous la forme d'une séquence complexe ordonnée selon le temps. Pour atteindre nos objectifs, nous avons déterminé plusieurs caractéristiques des règles d'épisodes à extraire : elles doivent être sans répétition, fréquentes, fiables, ayant l'antécédent le plus petit possible et la conséquence la plus éloignée temporellement de l'antécédent. Nous avons proposé un algorithme inspiré d'un autre de l'état de l'art "EMMA". Notre algorithme est bien plus rapide qu'EMMA, et donc que plusieurs autres algorithmes parce que les caractéristiques souhaitées permettent de réduire le nombre d'itérations nécessaires.

Nous envisageons de tester notre algorithme sur des données réelles de réseaux sociaux afin de prédire au plus tôt d'événements lointains. L'algorithme, tel que nous l'avons défini, ne permet d'extraire que des conséquences de taille 1 (1-uplet épisodes). Il est évident qu'une conséquence plus longue sera plus porteuse d'informations. Une de nos perspectives est donc d'adapter cet algorithme de façon à ce qu'il puisse extraire des conséquences plus longues. Cela va permettre d'anticiper des conséquences contenant plus d'informations sur les futurs événements. Nous envisageons également de modifier la première étape d'extraction des motifs fréquents locaux afin de ne pas extraire tous les motifs et d'éliminer les motifs longs, afin de favoriser l'extraction de règles minimales et de diminuer la complexité de l'algorithme. En effet, certains de ces motifs longs sont éliminés lors de la phase d'extraction de règles d'épisodes minimales de notre algorithme.

Références

Agrawal, R., R. Srikant, et al. (1994). Fast algorithms for mining association rules. In *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, Volume 1215, pp. 487–499.

Basak, J., A. Sudarshan, D. Trivedi, et M. Santhanam (2004). Weather data mining using independent component analysis. *The Journal of Machine Learning Research 5*, 239–253.

Bathoorn, R., M. Welten, M. Richardson, A. Siebes, et F. J. Verbeek (2010). Frequent episode mining to support pattern analysis in developmental biology. In *Pattern Recognition in Bioinformatics*, pp. 253–263. Springer.

Daurel, T. (2003). *Mining confident minimal rules with fixed-consequents*. Ph. D. thesis, Institut National des Sciences Appliquées de Lyon.

Gan, M. et H. Dai (2011). Fast mining of non-derivable episode rules in complex sequences. In *Modeling Decision for Artificial Intelligence*, pp. 67–78. Springer.

Gan, M. et H. Dai (2012). Mining condensed sets of frequent episodes with more accurate frequencies from complex sequences. *International Journal of Innovative Computing, Information and Control 8*(1), 453–470.

Han, J., J. Pei, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, et M. Hsu (2001). Prefix*S*pan : Mining sequential patterns efficiently by prefix-projected pattern growth. In *proceedings of*

the 17th international conference on data engineering, pp. 215–224.

Han, J., J. Pei, et Y. Yin (2000). Mining frequent patterns without candidate generation. In *ACM SIGMOD Record*, Volume 29, pp. 1–12. ACM.

Huang, K.-Y. et C.-H. Chang (2008). Efficient mining of frequent episodes from complex sequences. *Information Systems 33*(1), 96–114.

Iwanuma, K., R. Ishihara, Y. Takano, et H. Nabeshima (2005). Extracting frequent subsequences from a single long data sequence a novel anti-monotonic measure and a simple on-line algorithm. In *Data Mining, Fifth IEEE International Conference on*, pp. 8–pp. IEEE.

Laxman, S. et P. S. Sastry (2006). A survey of temporal data mining. *Sadhana 31*(2), 173–198.

Mannila, H., H. Toivonen, et A. I. Verkamo (1997). Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery 1*(3), 259–289.

Pasquier, N., Y. Bastide, R. Taouil, et L. Lakhal (1999). Discovering frequent closed itemsets for association rules. In *Database Theory-ICDT'99*, pp. 398–416. Springer.

Priya, V. et A. Vadivel (2012). User behaviour pattern mining from weblog. *International Journal of Data Warehousing and Mining (IJDWM) 8*(2), 1–22.

Rahal, I., D. Ren, W. Wu, et W. Perrizo (2004). Mining confident minimal rules with fixed-consequents. In *Tools with Artificial Intelligence, ICTAI*, pp. 6–13. IEEE.

Tatti, N. et B. Cule (2012). Mining closed strict episodes. *Data Mining and Knowledge Discovery 25*(1), 34–66.

Summary

The number of daily posted messages on social networks and blogs is huge, providing a valuable source of information. Temporal data mining is adapted to the exploration of this dataset. The result of this mining can be used for a prediction purpose, to anticipate the emergence of information. Our goal in this article is to propose an algorithm with the characteristic of early and confident information prediction.

In this preliminary work, we focus on extracting minimal episode rules to be quickly identified in a data stream. These rules also have the characteristic of having a consequence temporally distant of the antecedent, to be able to anticipate the consequence as soon as possible, and also have a high confidence.

Un modèle sémantique spatio-temporel pour capturer la dynamique des environnements

Benjamin Harbelot*, Helbert Arenas**, Christophe Cruz ***

*benjamin.harbelot@gmail.com
**helbert.arenas@checksem.fr
***christophe.cruz@u-bourgogne.fr

Résumé. Les Systèmes d'Information Géographique se sont peu à peu imposés comme des outils performants pour organiser, représenter, analyser et visualiser des données géographiques. Toutefois, l'intégration d'une dimension temporelle dans les SIG reste un défi de la recherche en sciences de l'information géographique. Dès lors, le développement de modèles spatio-temporels adaptés à l'étude de phénomènes géographiques réels devient un enjeu majeur dans la conception de systèmes d'informations dédiés à l'évolution d'entités spatiales. Dans ces travaux, nous proposons un nouveau modèle spatio-temporel basé sur une ontologie intégrant les connaissances des experts sur les données géographiques représentées. A terme, les capacités sémantiques proposées dans ce modèle permettent d'assister les experts dans la représentation et l'analyse d'un phénomène spatio-temporel en prenant en compte les informations contextuelles de l'environnement géographique.

1 Introduction

L'information géographique suscite un intérêt croissant, tant de la part des particuliers et des entreprises que des autorités publiques. Si les besoins en données géographiques ont toujours existé, la quantité d'informations générée a été tellement massive que les outils disponibles ont longtemps connu des difficultés à traiter efficacement ces données. Plus récemment, l'apparition d'une dimension temporelle dans certains Systèmes d'Information Géographique a étendu les champs d'étude et d'application liés à la géomatique. Désormais les objets à référence spatiale peuvent être étudiés dans leur dynamique qu'il s'agisse d'un mouvement ou d'un changement. Cependant, conjointement à l'ajout de la dimension temporelle, la quantité de données à stocker et à traiter s'est à nouveau considérablement accrue soulevant ainsi de nouveaux défis scientifiques et techniques. Le traitement d'informations aussi massives et complexes que les données spatiotemporelles nécessite d'être vigilant sur la modélisation afin de faciliter l'intégration et l'exploitation de ces données. Le suivi dans le temps d'objets spatiaux est l'un des enjeux majeurs de la modélisation spatiotemporelle. L'évolution dans le temps d'un objet spatial se compose d'une succession de transitions induisant le passage d'un état à un autre (changement de forme, de localisation, changement sur les attributs). Chaque transition de chaque objet se doit de représenter une évolution dans le respect de la réalité modélisée.

Il est fréquent de trouver des jeux de données partiellement complets et par conséquent, partiellement exploitables. Les jeux de données incomplets sont souvent dûs à la réutilisation de données spatiales historiques, on obtient alors l'état des objets à différentes époques, cependant les transitions d'un état à un autre pour chaque objet sont absentes et doivent être inférées afin d'assurer le suivi des objets dans le temps.

Dans ce papier, nous introduirons brièvement quelques notions de la modélisation spatio-temporelle. Puis, nous étudierons les différents modèles proposés dans la littérature pour répondre aux besoins de la modélisation spatio-temporelle, nous verrons ensuite en quoi l'utilisation d'une ontologie semble adaptée à ces besoins. Nous présenterons alors le modèle Continuum basé sur une ontologie afin de renforcer les capacités de représentation et d'analyse en modélisant de manière structurée l'environnement géographique dans son contexte et en fournissant des outils de raisonnement capables de générer des connaissances à partir des informations stockées dans l'ontologie. Enfin, nous présenterons un cas d'utilisation de notre modèle dans le cadre de l'étude de la décolonisation.

2 Modélisation spatiotemporelle

Afin de développer notre approche de modélisation, nous introduisons dans cette section quelques notions spatiales et temporelles.

2.1 Entités spatio-temporelles

Une entité spatio-temporelle est une représentation des entités du monde réel composée d'une identité, de propriétés descriptives et de propriétés spatiales. Tandis que l'identité décrit une sémantique fixe de l'entité, les propriétés descriptives et spatiales peuvent varier dans le temps et constituent la partie dynamique de l'entité. Lorsque l'identité d'une entité varie, on assiste à un type particulier d'évolution où l'entité spatio-temporelle est transformée en une nouvelle entité. Dans la littérature, on trouve deux principaux types d'entités spatio-temporelles : 1) les objets mouvants, comme par exemple un taxi sillonant les rues de la ville, et 2) les objets changeants, par exemple, une région dont la disposition des frontières administratives évolue dans le temps. Dans ce papier, nous nous intéresserons à ce deuxième cas.

2.2 Identité

Les applications spatio-temporelles ont vu le jour pour répondre au besoin de représenter des phénomènes impliquant l'espace et le temps. La modélisation d'un phénomène est intimement liée à celle de ses entités ainsi qu'à l'ensemble des changements intervenants sur ces entités. Dès lors, le concept d'identité a progressivement été introduit afin de caractériser les changements (Del Mondo et al. (2013)),(Del Mondo et al. (2010)),(Muller (2002)). L'identité peut se définir comme le caractère unique d'un objet, indépendamment de ses attributs ou de ses valeurs. C'est le trait qui distingue un objet de tous les autres. L'identité est essentielle dans la conceptualisation ainsi que dans la modélisation d'un phénomène. Cependant, cette notion est très subjective car elle dépend des critères choisis par l'utilisateur pour définir l'identité d'une entité. Généralement l'identité dépend du type d'entité et du phénomène.

2.3 Les relations de filiations

La relation de filiation définit le lien de succession qui existe entre les différentes représentations d'un même objet à différents moments du temps. Les relations de filiations trouvent leur intérêt dans le cadre de changements élémentaires tels que des divisions ou des fusions d'entités. D'autres changements spatiaux plus ou moins complexes existent et nécessitent d'identifier les entités "parents" et "enfants". A ce stade, La filiation se base uniquement sur des relations spatiales, c'est pourquoi elle peut être qualifiée de filiation spatiale dans le contexte des changements spatiaux. Ces changements spatiaux peuvent, en outre, révéler des changements sur la nature de l'entité, c'est pourquoi la relation de filiation est intimement liée à la notion d'identité. Cette relation est essentielle pour maintenir l'identité d'une entité qui évolue ainsi que de suivre l'évolution dans le temps. Il est également nécessaire d'identifier les nouvelles entités qui peuvent résulter d'une évolution. La mise en place d'une relation de filiation induit une dépendance à l'identité. Il y a alors deux types généraux de relations de filiation : continuation et dérivation (Harbelot et al. (2013a)),(Stell et al. (2011)),(Hornsby et Egenhofer (2000)). Dans le premier cas, l'identité est conservée. L'entité continue d'exister, mais a connu un changement. Alors que dans le deuxième cas, une nouvelle entité est créée à partir d'une autre après avoir été soumis à une évolution. Contrairement aux relations de continuation, les relations de dérivation peuvent concerner plusieurs entités en même temps.

3 Etude des différentes approches de modélisation spatio-temporelle

L'évolution d'une entité spatiale dans le temps peut être vue soit comme une succession d'états (ou représentations) de l'entité, soit comme une succession de transition intervenant sur cette entité au cours du temps. Les modèles de la première proposition se basent sur des approches continues ou discrètes. Des exemples de modèles sont le modèle snapshot (Armstrong (1988)), le modèle Space-Time Composites (STC) (Langran et Chrisman (1988)), le modèle Spatio-temporal Object (Worboys (1994)). L'inconvénient de ces modèles est qu'ils ne représentent que des changements soudains au travers desquels il est difficile d'identifier des processus tels que le changement ou le mouvement d'une entité de l'environnement géographique. D'autres modèles basés sur les approches continues ou discrètes ont été proposés pour gérer les changements liés à l'identité (Hornsby et Egenhofer (2000)) ou encore les changements sur les relations topologiques étudiées à l'aide de matrices d'intersection (Egenhofer et Al-Taha (1992)), toutefois l'analyse des causes du changement nécessaire à l'étude des phénomènes sont difficiles à déduire en utilisant ce type de modélisation. Par conséquent, cette première approche de modélisation ne permet pas une analyse complète de l'évolution. Pour pallier à ce problème, l'approche de modélisation basée sur les évènements et processus a peu à peu vu le jour. Cette approche considère que les entités spatiales évoluent sous l'impulsion d'un évènement ou d'un processus et dont l'objectif est d'analyser les causes et les conséquences. Parmi ces modèles, il est possible de citer le modèle Event-Based Spatiotemporal Data Model (ESTDM) (Peuquet et Duan (1995)), les processus composite (Claramunt et Theriault (1996)) ou encore le modèle de changement topologique basé sur les évènements (Jiang et Worboys (2009)). Le modèle ESTDM décrit un phénomène au travers d'une liste d'évènement, un nouvel évènement est créé en bout de liste à chaque fois qu'un changement est détecté. Toutefois

ce modèle ne prend en compte que des données de type raster et les liens de causalité entre les évènements sont difficilement mis en évidence dans ce modèle. Pour répondre à ce problème, le modèle des processus composites a pour objectif de représenter les liens entre les évènements ainsi que leurs conséquences, en outre, l'auteur avance que le modèle de donnée doit différencier ce qui est spatial, temporel et thématique. Le modèle des processus composites se base sur un langage permettant de décrire la sémantique liée à un phénomène du monde réel. Ce phénomène est assimilé à un processus composite, c'est-à-dire une suite de processus qui décrivent la dynamique du phénomène. Un processus composite est, par exemple, la trajectoire d'un bateau et peut se décomposer en 3 processus : stabilité, déplacement, rotation. Enfin, le modèle de changement topologique basé sur les évènements décrit chaque snapshot de l'environnement géographique à des instants de temps consécutifs sous la forme d'un arbre. Chaque arbre est lié au suivant et au précédent au travers de ses noeuds. Le lien entre deux arbres est un changement topologique qui révèle la création d'une entité sur l'environnement géographique, la suppression d'une entité, la division ou la fusion d'une entité ou aucun changement. La succession de ces changements topologiques permet de représenter des changements complexes.

4 L'ontologie comme support de la modélisation spatiotemporelle

En réalité, les deux approches que nous venons d'étudier sont complémentaires dans la modélisation d'un phénomène réel au sens où une transition correspond au passage d'une entité d'un état à un autre. En outre, les entités d'un espace géographique sont susceptibles d'entretenir différentes relations spatiales ou sémantiques entre elles, et ce, quelque soit l'état de leur évolution. Dès lors, il parait naturel de constater que la modélisation d'un phénomène dans le temps peut être assimilée à un réseau dans lequel les entités et leurs états sont vus comme des sommets tandis que les relations correspondent aux arêtes. Plusieurs travaux liés à la géomatique se sont appuyés sur la théorie des graphes pour des applications telles que la modélisation d'espaces urbains (Jiang et al. (2000)), l'étude de réseaux routiers (Jiang et Claramunt (2004a)) ou encore la modélisation de la structure d'un territoire au travers de graphes de connectivités (Jiang et Claramunt (2004b)). Cependant, tous ces exemples se focalise sur la structure spatiale et omettent la dimension temporelle qui impose de définir différemment le graphe modélisant le phénomène. Les transitions entre les différents états pour chaque entité sont au coeur de la modélisation spatio-temporelle car elles permettent de suivre une entité dans le temps. Dans la suite de ce papier, ces transitions sont confondues avec les relations de filiations présentées précédemment. La modélisation d'un phénomène génère un graphe complexe composés de relations telles que les relations temporelles, sémantiques, spatiales, ou encore des relations de filiation. L'ontologie se présente comme un modèle de données capable de représenter les concepts d'un domaine, ainsi que les relations entre ces concepts. Les concepts au sein d'un ontologie sont organisés dans un graphe permettant de modéliser l'ensemble des relations nécessaires à la modélisation spatio-temporelle. Traditionnellement utilisées pour représenter les éléments d'un domaine de connaissance, les ontologies se fondent sur les notions d'individus, classes, attributs, relations et évènements. Dès lors, les entités peuvent être assimilées à des individus et regroupées au sein de classes définies selon le type d'entité. Les attributs constituent les propriétés, caractéristiques ou paramètres que les entités peuvent posséder et partager

tandis que les relations définissent les liens qu'entretiennent entre elles les entités. Enfin, les évènements correspondent aux changements subis par des attributs ou des relations. Les informations disponibles sur des entités à modéliser peuvent être plus ou moins précises, par exemple, une entité de type "étendue d'eau" peut être spécialisée en un type "étang", "lac", "mer", "océan", etc... Une même spécialisation/généralisation peut être observée sur les relations entre les entités, par exemple, une relation appelée "relation spatiale" entre deux entités peut être spécialisée comme une "intersection", "adjacence", etc... Dans les deux cas, l'ontologie apporte une hiérachie basée sur la subsomption qui permet de représenter ou exploiter la connaissance à différents niveaux de détails. Si l'ontologie s'impose comme un support particulièrement complet pour représenter et classifier la connaissance liée à un domaine, elle est également largement utilisée pour raisonner à propos des entités modélisées et présente des capacités d'analyses supérieures aux systèmes d'informations traditionnels .

5 Le modèle Continuum

L'évolution d'entités spatiales dans le temps génére un graphe complexe de relations qui peut être modélisé à l'aide d'une ontologie (Harbelot et al. (2013c)),(Harbelot et al. (2013b)). La construction de l'ontologie est soumise à un langage de spécification. Dans ces travaux, nous utilisons le langage OWL au niveau DL (Ontology Web Language Description Logics) permettant une expressivité adaptée aux besoins de la modélisation et également des algorithmes d'inférence décidables pour raisonner sur l'ontologie. Si les relations sémantiques et les relations de subsomptions sont naturellement présentes et faciles à définir au sein d'une ontologie, le stockage de l'information géographique et la représentation d'une même entité à différentes époques ainsi que les relations de filiation entre les différents états de cette entité nécessitent des méthodes spécifiques. En outre, la modélisation spatio-temporelle s'impose comme le support de l'étude de l'environnement et des phénomènes géographiques. Dès lors, le système se doit de fournir des fonctionnalités d'analyses spatiales et temporelles au travers de l'outil de recherche mit en place sur le système d'information.

5.1 Modèles de données pour la gestion de l'information spatiale et l'évolution temporelle d'objets

La gestion de données géospatiales dans le Web Sémantique est un domaine émergent pour lequel une large majorité de la communauté s'est réunit autour de la définition d'un standard : GeoSPARQL (Consortium et al. (2012)). GeoSPARQL se compose d'une petite ontologie destinée à fournir une représentation spatiale aux entités définies dans une ontologie de domaine. Outre le stockage de l'information géographique, GeoSPARQL sert également de support au raisonnement spatial quantitatif et qualitatif au sein du langage de requête SPARQL. Cependant, GeoSPARQL ne fournit aucun support pour l'intégration de la dimension temporelle. Des ontologies de temps comme OWL-Time ou SWRL Temporal Ontology permettent d'introduire des données temporelles au sein d'une ontologie au travers d'instants ou intervalles de temps, de concepts temporels et des relations topologiques temporelles définies par Allen (Allen (1983)).

Mais la représentation du temps ne suffit pas à modéliser l'évolution d'une entité ainsi que ses propriétés et relations. Il existe deux principales théories philosophiques pour la représen-

tation d'entités évoluant dans le temps : l'*endurantisme* et le *perdurantisme*. L'endurantisme considére tout objet comme une entité en trois dimensions qui existe de manière invariante à n'importe quel instant de son existence. Le perdurantisme, quant à lui, est considéré comme une vue des objets en quatres dimensions dans laquelle les entités sont divisées en parties temporelles appelées *timeslices* (Welty et al. (2006)). Chacun des *timeslices* est une représentation partielle des entités, et valide sur une période ou à un instant de temps donné. Par conséquent, la représentation complète d'une entité au cours du temps est le résultat de l'aggrégation de tous les *timeslices*. La représentation du changement des propriétés d'un objet se réalise à travers l'utilisation de *fluents* qui sont des propriétés valides uniquement durant un certain intervalle. Le perdurantisme offrant plus d'avantages que l'*endurantisme* en permettant une représentation plus complète des phénomènes du monde réel (Al-Debei et al. (2012)), nous basons notre modèle sur cette approche. Pour cela, des ontologies de haut niveau basées sur les ontologies de temps ont vu le jour pour modéliser des entités selon une approche perdurantiste. Ces ontologies sont connues sous le nom d'ontologie de *fluents* parmi lesquelles nous retiendrons l'approche *4D-Fluent* (Welty et al. (2006)) ou encore *TOWL* (Milea et al. (2007)). Toutefois, ces ontologies de *fluents* n'ont pas été définies à destination de la modélisation spatio-temporelle et doivent être ajustées pour répondre aux besoins du domaine.

GeoSPARQL permet une représentation spatiale des entités et fournit des fonctions d'analyses spatiales, cependant l'usage de ce standard reste destiné à des applications statiques. Les ontologies de *fluents*, à l'inverse, ont été proposée pour représenter l'évolution dans le temps au travers des *timeslices* et des *fluents*, toutefois, l'application au domaine géospatial nécessite quelques adaptations pour répondre pleinement aux besoins de la modélisation. Enfin, les ontologies de *fluents* représente l'évolution d'une entité via une succession de *timeslices* sans exprimer de liens explicites entre les *timeslices* consécutifs. Ce lien entre *timeslices* appliqué au domaine du géospatial est assimilé à la relation de filiation et s'avère essentiel pour exprimer et comprendre des phénomènes géographiques tels que la fusion ou la division d'entités spatiales.

5.2 Présentation du modèle Continuum

Le modèle Continuum tire avantage des capacités spatiales de GeoSPARQL ainsi que des méthodes proposées dans les ontologies de *fluents* pour représenter l'évolution. En outre, ce modèle permet de représenter le lien de filiation entre les *timeslices* consécutifs liés à une entité. Le modèle Continuum a été proposé pour répondre au besoin de modélisation des changements qui s'opèrent sur les entités spatiales. Ces changements sont caractérisés par des modifications spatiales (liées à la position ou à l'empreinte spatiale de l'entité), liées à la sémantique ou liées à l'identité des entités. Dès lors, le modèle intégre et distingue les relations liées à l'espace (relations topologiques), au temps (relations d'Allen (Allen (1983))), à la sémantique et à l'identité (relation de filiation).

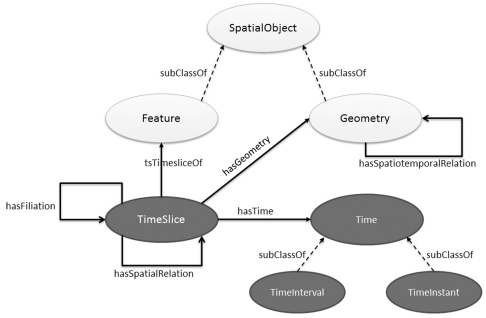

Figure 1 : Ontologie noyau du modèle Continuum

Pour présenter notre modèle, nous utiliserons la définition ontologique de Karlsruhe (Stumme et al. (2003)). Une ontologie OWL DL est une structure $O := (SO, \sigma O, FO)$ qui consiste en :

- Un ensemble sous-jacent SO contenant : Six ensembles disjoints sC, sT, sR, sA, sI, sV, sKR et sKA appelés respectivement concepts, type de donnée, relations, attributs, instances, valeurs de donnée, relations caractéristiques (telles que Symétrique, Fonctionnel, Inversement Fonctionnel, Transitif) et attributs caractéristiques (Fonctionnel). Quatres ordres partiels $\leq C, \leq T, \leq R$ et $\leq A$, respectivement sur sC appelée hiérarchie de concepts ou taxonomie, sur sT appelée hiérachie de type, sur sR appelée hiérarchie de relations et sur sA appelée hiérarchie d'attributs, tels que :
  $SO := (sC, \leq C), (sT, \leq T), (sR, \leq R), (sA, \leq A), sI, sV, sKR, sKA$
- La signature σO contenant : deux fonctions $\sigma R : sR \rightarrow sC^2$ appelée signature de la relation et $\sigma A : sA \rightarrow sC \times sT$ appelée signature d'attributs, telle que $\sigma O := \sigma R, \sigma A$,
- La fonction d'interprétation FO contenant les fonctions :
  - $\iota C : sC \rightarrow 2^{\text{si}}$ correspondant à l'instantiation de concepts,
  - $\iota T : sA \rightarrow 2^{\text{sv}}$ correspondant à l'instantiation de types de données,
  - $\iota R : sC \rightarrow 2^{\text{si} \times si}$ correspondant à l'instantiation de relations,
  - $\iota A : sC \rightarrow 2^{\text{si} \times sv}$ correspondant à l'instantiation d'attributs,
  - $\kappa R : sR \rightarrow 2^{\text{skr}}$ correspondant à la caractérisation de relations,
  - $\kappa A : sA \rightarrow 2^{\text{ska}}$ correspondant à la caractérisation d'attributs,
  - $\text{-}R : sR \rightarrow 2^{\text{sr}}$ correspondant à la spécification de la relation inverse,
- sC= TopConcept, SpatialObject, Feature, Geometry, TimeSlice, Time, TimeInterval, TimeInstant,
- sT=TopType, dateTime, String,
- $\leq C$ =(TopConcept, SpatialObject), (TopConcept, TimeSlice), (TopConcept, Geometry), (TopConcept, Time), (SpatialObject, Feature), (SpatialObject, Feature), (Time, TimeInterval), (Time, TimeInstant),
- $\leq T$=(TopType, Float) (TopType, String),
- sR= TopRelation, hasFiliation, hasSpatialRelation, hasTemporalRelation, isTimesliceOf, hasGeometry, hasTime , asWKT, hasDisjoint , hasOverlaps, hasTouches, hasContains, hasWithin, hasEqual, hasContinuation, hasDerivation, hasExpansion, hasContraction, hasSplits, hasFusion, hasDerivSeparation, hasContiSeparation, hasDerivAnnexation, hasContiAnnexation, before, meets, overlaps, during, starts, finishes, after, metBy, overlappedBy, contains, startedBy, finishedBy, Equal ,
- sA=TopAttribute, hasBeginInstant, hasEndInstant, hasTimeInstant ,

- σR=(hasFiliation (TimeSlice, TimeSlice)), (hasSpatialRelation, (TimeSlice, TimeSlice)), (isTimesliceOf,(TimeSlice, Feature)), (hasGeometry,(TimeSlice, Geometry)), (hasTime, (TimeSlice, Time)) , (hasDisjoint,(TimeSlice, TimeSlice)),(hasOverlaps,(TimeSlice, TimeSlice)), (hasTouches,(TimeSlice, TimeSlice)), (hasWithin,(TimeSlice, TimeSlice)),(isWithin,(TimeSlice, TimeSlice)), (hasDisjoint,(Geometry, Geometry)),(hasOverlaps,(Geometry, Geometry)), (hasTouches,(Geometry, Geometry)), (contains,(Geometry, Geometry)),(within,(Geometry, Geometry)) ,(equal,(Geometry, Geometry)) ,(hasContinuation,(TimeSlice, TimeSlice)), (hasDerivation,(TimeSlice, TimeSlice)), (hasExpansion,(TimeSlice, TimeSlice)), (hasContraction,(TimeSlice, TimeSlice)), (hasSplits,(TimeSlice, TimeSlice)), (hasFusion,(TimeSlice, TimeSlice)), (hasDerivSeparation,(TimeSlice, TimeSlice)), (hasContiSeparation,(TimeSlice, TimeSlice)), (hasDerivAnnexation,(TimeSlice, TimeSlice)), (hasContiAnnexation,(TimeSlice, TimeSlice)),(equal,(TimeInterval, TimeInterval)), (before,(TimeInterval, TimeInterval)), (after,(TimeInterval, TimeInterval)), (meets,(TimeInterval, TimeInterval)), (metBy,(TimeInterval, TimeInterval)), (overlaps,(TimeInterval, TimeInterval)), (overlappedBy,(TimeInterval, TimeInterval)), (during,(TimeInterval, TimeInterval)), (contains,(TimeInterval, TimeInterval)), (starts,(TimeInterval, TimeInterval)), (startedBy,(TimeInterval, TimeInterval)), (finishes,(TimeInterval, TimeInterval)), (finishedBy,(TimeInterval, TimeInterval)),
- σA=(hasBeginInstant (TimeInterval, dateTime)), (hasEndInstant (TimeInterval, dateTime)), (hasTimeInstant (TimeInstant, dateTime)), (asWKT (Geometry, String)),
- $\leq R$=(TopRelation, hasFiliation), (TopRelation, hasSpatialRelation), (TopRelation, isTimesliceOf), (TopRelation, hasGeometry), (TopRelation, asWKT), (TopRelation, hasTime), (hasSpatialRelation, hasDisjoint), (hasSpatialRelation, hasOverlaps), (hasSpatialRelation, hasTouches), (hasSpatialRelation, contains), (hasSpatialRelation, within), (TopRelation, hasTemporalRelation), (hasTemporalRelation, before) , (hasTemporalRelation, after) , (hasTemporalRelation, meets) , (hasTemporalRelation, metBy) , (hasTemporalRelation, overlaps) , (hasTemporalRelation, overlappedBy) , (hasTemporalRelation, during) , (hasTemporalRelation, contains) , (hasTemporalRelation, starts) , (hasTemporalRelation, startedBy) , (hasTemporalRelation, finishes) , (hasTemporalRelation, finishedBy) , (hasTemporalRelation, equal)
- $\leq A$=(TopAttribute, hasBeginInstant),(TopAttribute, hasEndInstant) ,(TopAttribute, hasTimeInstant),
- sKR=Transitive, Symmetric, Functional,
- sKA=Functional,
- κR=(hasDisjoint, Symmetric), (hasOverlaps, Symmetric), (hasTouches, Symmetric) , (hasEqual, Symmetric), (hasContains, Transitive), (hasWithin,Transitive), (hasContinuation,Functional), (hasExpansion,Functional), (hasContraction,Functional), (hasContiSeparation,Functional), (hasContiAnnexation,Functional), (equal, Symmetric), (during, Transitive), (contains,Transitive),
- κA=(hasBeginInstant, Functional), (hasEndInstant, Functional), (hasEndInstant, Functional), (asWKT, Functional) ,
- -R=(hasContains,hasWithin), (before,after), (meets,metBy), (overlaps,overlappedBy), (during,contains), (starts,startedBy), (finishes,finishedBy) ,

5.3 Détection de motifs spatio-temporels au sein du modèle Continuum

Les phénomènes du monde réel à modéliser sont soumis à des contraintes. Plus un phénomène présente de contraintes pour être représenté ou détecté et plus ce phénomène se distingue des autres pour révéler sa nature unique. A l'inverse, la réduction du nombre de contraintes permet de regrouper des phénomènes selon des caractéristiques communes. S'il est évident qu'extraire la nature unique d'un phénomène s'avère l'option la plus intéressante, cette analyse est généralement limitée par le nombre de contraintes exploitables à disposition dans le système d'information. D'autre part, une analyse lançée sur la base de nombreuses contraintes se heurte généralement à des temps de calculs relativement long. En réponse aux problématiques exposées précédemment, notre approche se base sur un nombre de contraintes suffisamment générique pour regrouper un nombre conséquent de résultat, et suffisamment spécifique pour servir de support à des analyses plus poussées. Le résultat de l'analyse est stocké sous la forme de relations organisées en couche. En somme, notre approche agit comme un pré-traitement des données permettant d'accélérer les temps de calcul d'analyses de plus haut niveau ainsi que de fournir de nouveaux supports d'analyse exploitables. Dans notre modèle, nous proposons de spécialiser la relation de filiation au sein d'une hiérachie de subsomption. La hiérarchie se compose de trois couches basées sur des contraintes :

- La couche *filiation* : seule la relation de filiation est établie afin de suivre le cycle de vie d'une entité au travers de ses évolutions. Aucune connaissance spécifique n'est portée par cette relation.
- La couche *identité* : la relation de filiation est spécialisée soit en une relation de dérivation si l'entité change sa nature, soit en relation de continuation si cette entité conserve sa nature au travers de l'évolution représentée.
- La couche *topologie et identité* : cette couche regroupe des contraintes spatiales ainsi que des contraintes d'identité pour fournir six motifs d'évolution d'une entité spatiale. Ces motifs ont été proposés dans Del Mondo et al. (2013).

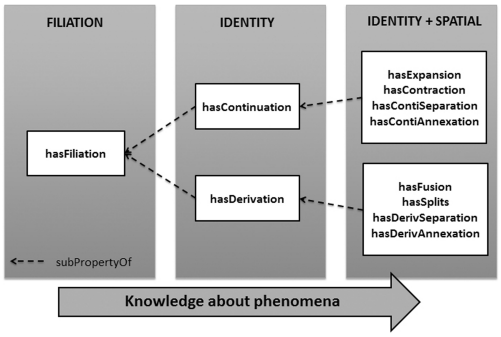

Figure 2 : Couches de qualification de la relation de filiation

La troisième couche de la hiérarchie est la plus expressive car elle décrit les motifs les plus complexes (voir Figure 3). Ces motifs font intervenir des contraintes spatiales et identitaires. Les divisions et les fusions spatiales sont détectées en vérifiant que l'union des géométries *enfants* est contenu dans la géométrie *parent* dans le premier cas et que l'union des géométries *parents* est contenu dans la géométrie *enfant* dans le second. Les fonctions d'analyses spatiales

nécessaires à cette détection sont présentes dans GeoSPARQL. Les contraintes d'identités, quant à elles sont définies par l'utilisateur selon les besoins de l'application :

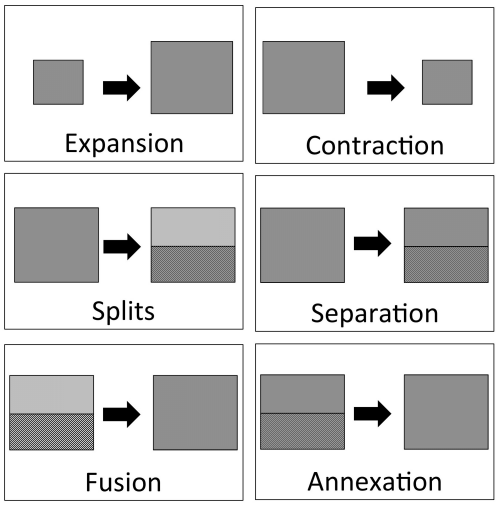

Figure 3 : Différents motifs de la couche *topologie et identité*

- **Expansion :** l'entité continue d'exister mais la géométrie s'étend. Cette relation est une relation de continuation.
- **Contraction :** l'entité continue d'exister mais la géométrie se réduit. Cette relation est une relation de continuation.
- **Division :** l'entité parent cesse d'exister, sa géométrie se sépare en deux nouvelles géométries correspondants chacune à une nouvelle entité qui n'existait pas auparavant. Cette relation donne lieu à deux dérivations.
- **Séparation :** l'entité parent continue d'exister mais sa géométrie se sépare et une nouvelle géométrie correspondants à une nouvelle entité apparait. Cette relation donne lieu à une continuation et au moins une dérivation.
- **Fusion :** les deux entités parents fusionnent et cessent dexister pour donner lieu à une nouvelle géométrie correspondant à une nouvelle entité. Cette relation donne lieu à deux dérivations.
- **Annexation :** les deux entités parents fusionnent mais une seule continue d'exister. Cette relation donne lieu à une continuation et au moins une dérivation.

Si la couche *filiation* ne porte aucune connaissance liée à la nature de l'évolution et permet seulement de suivre une entité à deux instants de temps consécutifs, les deux autres couches, quant à elles, augmentent progressivement cette connaissance. Chaque couche peut être vue comme un support d'analyse de l'évolution selon des contraintes distinctes. A terme, les motifs décrits dans les couches peuvent être mis en relation avec les informations contextuelles stockées dans l'ontologie afin de décrire un phénomène spécifique du monde réel.

6 Exemple de l'évolution de la décolonisation

La décolonisation est un des phénomènes historiques majeurs de la seconde moitié du XXème siècle. Dans notre exemple, nous souhaitons modéliser l'évolution des territoires occupés. Puis, dans un second temps, nous souhaitons fournir une analyse de l'espace modélisé afin de détecter les mécanismes de décolonisation et enrichir le système d'information avec les connaissances extraites de cette analyse. Pour les besoins de l'exemple, nous intégrons une ontologie de domaine au sein de notre modèle de données. L'ontologie de domaine se compose de la classe *Landparcel* spécialisée en *Territoire colonisé* ou *état indépendant*. Ces deux classes sont disjointes l'une de l'autre car un état ne peut être à la fois colonisé et indépendant. Enfin, la classe *état indépendant* est spécialisée en *état colonisateur*. Cette relation de subsomption implique que pour être un état colonisateur, il faut être un état indépendant.

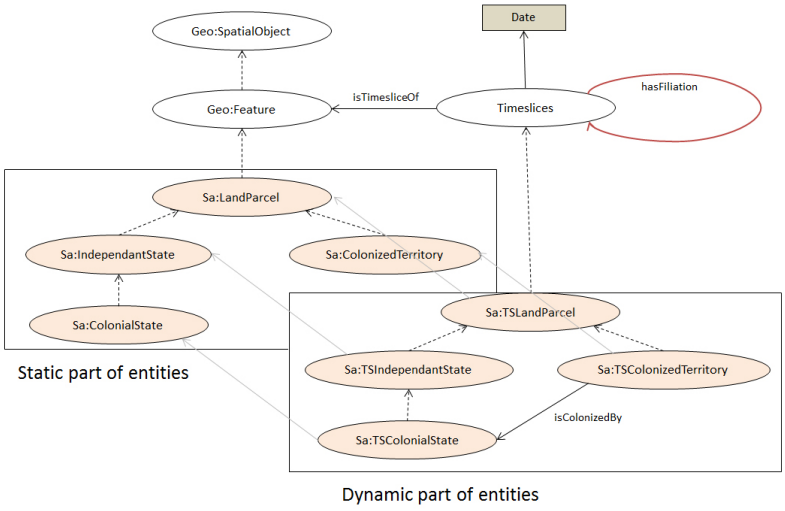

Figure 4 : Ontologie utilisée pour illustrer le phénomène de décolonisation

Par ailleurs, notre modèle fournit la classe *Timeslices* afin de permettre aux entités d'évoluer dans le temps. Toutefois, contrairement à l'approche *4D-fluent* (Welty et al. (2006)), notre modèle spécialise la classe *Timeslices* par réplication de l'ontologie de domaine (voir figure 3 : *partie dynamique des entités*). Ainsi, chaque classe de l'ontologie de domaine est associée à une sous-classe de la classe *Timeslices* représentant la partie dynamique des entités. L'avantage de cette méthode est qu'elle permet de conserver le *Domain* et le *Range* des propriétés spécifiques aux différents concepts. Par exemple, la propriété *EstColoniséPar* que nous introduisons dans notre exemple s'applique entre un *Territoire Colonisé* (*Domain*) et un *Etat Colonisateur* (*Range*). Or l'approche 4D-Fluent établit toutes propriétés dynamiques (fluents) entre des individus de la classe *Timeslices*. En d'autres termes, cette approche définie la classe *Timeslices* comme *Domain* et *Range* de toutes propriétés dynamiques. Par conséquent, toutes les propriétés dynamiques définies dans l'ontologie peuvent être appliquées à n'importe quel *timeslice* engendrant une sémantique floue et des risques d'incohérences élevés.

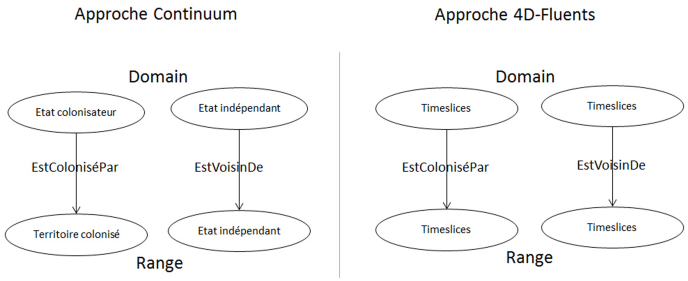

Figure 5 : Comparaison de la représentation des propriétés dynamiques dans l'approche Continuum et dans l'approche 4D-Fluents

Dans l'étude de la décolonisation, on assiste à plusieurs évolutions spatiales au niveau des frontières entre les différentes parcelles. L'identité des entités est amenée à varier lorsqu'un territoire colonisé est cédé à un autre état colonisateur ou bien lorsqu'un état accède à son indépendance. Le suivi des entités à des instants de temps consécutifs est assuré à l'aide de la relation de filiation. Lorsque l'identité est maintenue entre deux instants de temps consécutifs, la relation de continuation peut être établie. A l'inverse, si l'identité varie, alors la relation de dérivation devra être stockée.

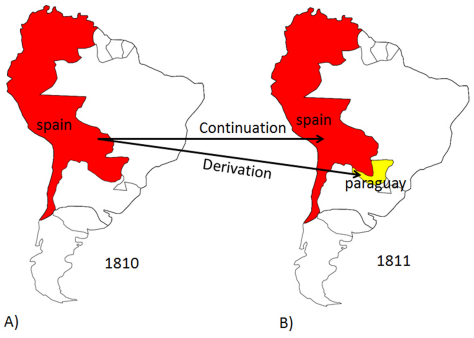

Figure 6 : Exemple de "Séparation" dans le cadre la décolonisation en Amérique du Sud

Les motifs définis dans notre modèle peuvent également être détectés au sein de l'espace modélisé et stockés dans l'ontologie sous forme d'une relation. La figure 5 représente un exemple de *séparation* intervenu entre 1810 et 1811. La *séparation* est le motif décrivant la division d'une entité spatiale avec une des parties qui conserve son identité (*Continuation*) et l'autre partie dont l'identité varie (*Dérivation*). Dans cet exemple, nous avons représenté en rouge un territoire occupé par l'espagne tandis que le jaune représente un état indépendant issu de la séparation du territoire Espagnol. Une partie du territoire reste occupé par l'Espagne induisant une relation de continuation tandis que l'autre partie du territoire a donné naissance à un état indépendant : le Paraguay. Les relations de *Continuation* ainsi que de *Dérivation* peuvent être automatiquement détectées par le modèle. En ce basant sur ces relations et par l'ajout de contraintes spatiales, il est également possible de détecter le motif de *Séparation* et de le stocker sous forme de relation. le motif de Séparation décrit par la figure représente en fait une partie du phénomène de *décolonisation* de l'Espagne qui s'accompagne ici d'un phénomène

d'*indépendance* pour le Paraguay. En utilisant les informations contextuelles présentes dans l'ontologie et en se basant sur les motifs préalablement détectés par le modèle, il est possible d'établir des règles permettant détecter et de suivre le processus complet de décolonisation d'un état.

```
INSERT {?ts1 ex:decolonisation ?ts2 }
WHERE {
   ?ts1 ex:hasContiSeparation?ts2  .
   ?ts1 ex:isTimesliceOf ?o1  .
   ?ts2 ex:isTimesliceOf ?o2  .
   ?o1 rdf:type ex:ColonizedTerritory .
   ?o2 rdf:type ex:ColonizedTerritory .
   ?ts1 ex:isColonizedBy ?tsColoStates1 .
   ?ts2 ex:isColonizedBy ?tsColoStates2 .
   ?tsColoStates1 ex:isTimesliceOf ?ColoStates .
   ?tsColoStates2 ex:isTimesliceOf ?ColoStates .
   ?ColoStates rdf:type ex:ColonialState .
}
```

Algorithme 1 : Règle SPARQL pour inférer un processus de décolonisation

```
INSERT {?ts1 ex:independance ?ts2 }
WHERE {
   ?ts1 ex:hasDerivSeparation?ts2  .
   ?ts1 ex:isTimesliceOf ?o1  .
   ?ts2 ex:isTimesliceOf ?o2  .
   ?o1 rdf:type ex:TSColonizedTerritory .
   ?o2 rdf:type ex:TSIndependantState .
}
```

Algorithme 2 : Règle SPARQL pour inférer un processus d'indépendance d'un territoire

Les algorithmes 1 et 2 présentés ci-dessus permettent d'inférer un processus de *décolonisation* ainsi qu'un processus de d'*indépendance* à l'aide du langage SPARQL. Il est également possible de fournir une analyse du système à l'aide de règle SWRL cependant cet outil n'étant que très rarement disponible sur les triplestores actuels, nous avons laissé cette option de côté pour le moment.

7 Conclusions et travaux futurs

Fort du constat qu'une représentation accrue de la connaissance liée aux entités spatiales offrent des capacités d'analyses supérieures aux SIG traditionnels, nous présentons dans ces travaux un modèle basé sur une ontologie destiné à représenter des entités spatiales ainsi que leur contexte d'évolution afin de mieux appréhender des phénomènes du monde réel. Le modèle s'articule autour de 4 composantes que sont l'espace, le temps, la sémantique et l'identité. Nous avons vu en quoi l'ontologie semblait être un candidat adéquate pour répondre aux besoins de la modélisation spatio-temporelle, c'est pourquoi nous avons rapproché nos travaux

des technologies du Web Sémantique pour définir notre modèle. En sus, nous utilisons la logique de description pour représenter notre ontologie à l'aide du langage OWL-DL. Le langage OWL-DL propose également plusieurs moteurs d'inférences pour raisonner sur des logiques de description afin de détecter des incohérences ou encore inférer des connaissances à partir de celles déjà existantes. Notre modèle peut également être interroger au travers du langage de requête SPARQL. Les entités spatiales peuvent être représentées de façons multiples suivant les besoins d'utilisations. Par exemple, une ville peut être représentée par un simple point dans le cadre d'une application visant à situer les villes au sein d'une région, ou bien, une ville peut être définie plus précisément à l'aide d'une empreinte spatiale (généralement un polygone). De même, les frontières d'un territoire comme la France peuvent être vue au travers de celle définie pour le pays, les régions qui le composent, les départements, les villes, etc... Dès lors, l'environnement spatiale peut être soumis à des contraintes de granularités spatiales permettant de visualiser et analyser l'espace géographique à différents niveaux de détails. Nos futurs travaux s'orienteront vers la prise en compte de cette granularité spatiale afin d'étendre les fonctionnalités de notre modèle.

Remerciements

Ces travaux sont financés par 1) la Direction Générale de l'Armement,
voir http ://www.defense.gouv.fr/dga/ 2) le Conseil régional de Bourgogne.

Références

Al-Debei, M. M., M. M. Al Asswad, S. de Cesare, et M. Lycett (2012). Conceptual modelling and the quality of ontologies : Endurantism vs. perdurantism. *International Journal of Database Management Systems 4*(3), 1–19.

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM 26*(11), 832–843.

Armstrong, M. P. (1988). Temporality in spatial databases. In *Proceedings : GIS/LIS*, Volume 88, pp. 880–889.

Claramunt, C. et M. Theriault (1996). Toward semantics for modelling spatio-temporal processes within gis. *Advances in GIs Research I*, 27–43.

Consortium, O. G. et al. (2012). Ogc geosparql-a geographic query language for rdf data. *OGC Candidate Implementation Standard*.

Del Mondo, G., M. Rodríguez, C. Claramunt, L. Bravo, et R. Thibaud (2013). Modeling consistency of spatio-temporal graphs. *Data & Knowledge Engineering 84*, 59–80.

Del Mondo, G., J. G. Stell, C. Claramunt, et R. my Thibaud (2010). A graph model for spatio-temporal evolution. *Journal of Universal Computer Science 16*(11), 1452–1477.

Egenhofer, M. J. et K. K. Al-Taha (1992). Reasoning about gradual changes of topological relationships. In *Theories and methods of spatio-temporal reasoning in geographic space*, pp. 196–219. Springer.

Harbelot, B., H. Arenas, C. Cruz, et al. (2013a). Continuum : A spatiotemporal data model to represent and qualify filiation relationships. In *4th ACM SIGSPATIAL International Workshop on GeoStreaming (IWGS) 2013*.

Harbelot, B., H. Arenas, C. Cruz, et al. (2013b). A semantic model to query spatial-temporal data. In *The 6th International Workshop on Information Fusion and Geographic Information Systems : Environmental and Urban Challenges*.

Harbelot, B., A. Helbert, C. Christophe, et al. (2013c). The spatio-temporal semantics from a perdurantism perspective. In *In Proceedings of the Fifth International Conference on Advanced Geographic Information Systems, Applications, and Services GEOProcessing*.

Hornsby, K. et M. J. Egenhofer (2000). Identity-based change : a foundation for spatio-temporal knowledge representation. *International Journal of Geographical Information Science 14*(3), 207–224.

Jiang, B. et C. Claramunt (2004a). A structural approach to the model generalization of an urban street network*. *GeoInformatica 8*(2), 157–171.

Jiang, B. et C. Claramunt (2004b). Topological analysis of urban street networks. *Environment and Planning B 31*(1), 151–162.

Jiang, B., C. Claramunt, et B. Klarqvist (2000). Integration of space syntax into gis for modelling urban spaces. *International Journal of Applied Earth Observation and Geoinformation 2*(3), 161–171.

Jiang, J. et M. Worboys (2009). Event-based topology for dynamic planar areal objects. *International Journal of Geographical Information Science 23*(1), 33–60.

Langran, G. et N. R. Chrisman (1988). A framework for temporal geographic information. *Cartographica : The International Journal for Geographic Information and Geovisualization 25*(3), 1–14.

Milea, V., F. Frasincar, U. Kaymak, et T. Noia (2007). An owl-based approach towards representing time in web information systems. In *4th International Workshop of Web Information Systems Modeling (WISM 2007)*, pp. 791–802.

Muller, P. (2002). Topological spatio–temporal reasoning and representation. *Computational Intelligence 18*(3), 420–450.

Peuquet, D. J. et N. Duan (1995). An event-based spatiotemporal data model (estdm) for temporal analysis of geographical data. *International journal of geographical information systems 9*(1), 7–24.

Stell, J., G. Del Mondo, R. Thibaud, et C. Claramunt (2011). Spatio-temporal evolution as bigraph dynamics. In *Spatial Information Theory*, pp. 148–167. Springer.

Stumme, G., M. Ehrig, S. Handschuh, A. Hotho, A. Maedche, B. Motik, D. Oberle, C. Schmitz, S. Staab, L. Stojanovic, N. Stojanovic, R. Studer, Y. Sure, R. Volz, et V. Zacharias (2003). The Karlsruhe view on ontologies. Technical report, University of Karlsruhe, Institute AIFB.

Welty, C., R. Fikes, et S. Makarios (2006). A reusable ontology for fluents in owl. In *FOIS*, Volume 150, pp. 226–236.

Worboys, M. F. (1994). A unified model for spatial and temporal information. *The Computer Journal 37*(1), 26–34.

Un modèle sémantique spatiotemporel pour capturer la dynamique des environnements

Summary

Geographic Information Systems have gradually emerged as powerful tools to organize, represent, analyze and visualize geographic data, however, the integration of a temporal dimension in GIS remains a challenge for research in sciences geographic information. Therefore, the development of spatio-temporal models suited to the study of real geographic phenomena becomes a major issue in the design of information systems dedicated to the evolution of spatial entities. In this work, we propose a new spatio-temporal model based on ontology incorporating expert knowledge on spatial data represented. Ultimately, this model allows experts to be assisted in the representation and analysis of spatio-temporal phenomenon.

Trajectoires médianes

Laurent Etienne*, Thomas Devogele**

*Institut de Recherche de l'Ecole navale, Brest, France
laurent.etienne@ecole-navale.fr
http://www.letienne.net
**Université de Tours, Laboratoire d'informatique, Blois, France
thomas.devogele@univ-tours.fr
http://thomas.devogele.free.fr

Résumé. Des masses de données représentant des trajectoires sont aujourd'hui disponibles. Ces trajectoires sont nombreuses et très diverses (piétons, véhicules, navires, avions, animaux...). Cet article présente un algorithme générique et robuste pour obtenir une trajectoire médiane. Ce motif permet de synthétiser un grand nombre de trajectoires partant d'une zone d'intérêt et se rendant dans une autre zone. Il permet de déduire des connaissances sur les comportements d'objets mobiles de même type. Il est basé sur l'appariement des positions des trajectoires à l'aide de la distance de Fréchet discrète. Deux jeux de trajectoires volumineux permettent d'illustrer cet algorithme.

1 Introduction

Des masses de données de localisation sont aujourd'hui collectées. En particulier, la technologie GPS permet de disposer de volumes importants de trajectoires d'objets mobiles. Pour comprendre le mouvement de ces objets, il est nécessaire de définir des motifs synthétisant ces masses de trajectoires et d'en extraire des connaissances. Dans cet article nous nous focalisons sur des groupes de trajectoires partant d'une même zone initiale et arrivant à une même destination. L'objectif est de synthétiser un groupe de trajectoires d'objets de même type suivant un même chemin à l'aide d'une trajectoire centrale. Plusieurs solutions ont été proposées (Buchin et al. (2010); Freeman et al. (2011); Vermeulen (2013)). Hélas, ces solutions ne sont pas robustes. Elles sont sensibles à la présence de trajectoires aberrantes (outliers) et traitent difficilement les trajectoires réalisant des boucles.

L'utilisation d'algorithmes basés sur la technique du Dynamic Time Warping (DTW) permet de s'affranchir de ces problématiques car il respecte l'ordonnancement temporel des positions (Petitjean et al. (2011)). Différentes techniques d'alignement de séries temporelles appliquées aux trajectoires produisent des suites de couples de positions pouvant être utilisées pour définir le concept de trajectoire médiane proposé dans cet article.

Ces travaux, issus de la thèse de Etienne (2011), étendent ceux de Devogele (2002) qui définissent des lignes centrales par appariement de points homologues. L'approche proposée est similaire à celle de Petitjean et al. (2011) qui propose une méthode pour définir une séquence moyenne à partir d'un ensemble de séquences temporelles.

La structure de cet article est la suivante. Dans le chapitre suivant, l'algorithme permettant de définir la trajectoire médiane est détaillé. Dans le chapitre 3, deux exemples de motifs s'appuyant sur 500 trajectoires de navires et 800 trajectoires de pigeons permettent d'illustrer cet algorithme. Le chapitre 4 conclut en rappelant les apports de cette technique et les principales perspectives.

2 Trajectoires médianes

2.1 Approche générale

L'étude d'un grand nombre de jeu de trajectoires montre que les clusters de trajectoires ne sont pas aisément modélisables par une loi statistique de distribution connue. Par exemple, une loi symétrique n'est pas adaptée pour modéliser des motifs de trajectoires (Etienne (2011)). En effet, un objet mobile se déplaçant entre deux points a tendance à optimiser son déplacement en termes de temps, de sécurité et de coûts. Il est par ailleurs contraint dans son déplacement par certains obstacles (fixes ou mobiles) ainsi que par son environnement (pluie, vent, courant, ...). L'ensemble de ces critères impactent son déplacement et sont difficilement modélisables par une loi statistique.

D'un point de vue statistique descriptive, il donc nécessaire de se tourner vers des approches de type médianes. De plus, la médiane est plus robuste à la présence de données aberrantes. Dans cet objectif, pour un ensemble de points, il est possible de définir un point médian. Il existe un grand nombre d'extensions de la médiane aux données en deux dimensions (Small (1990)). Le choix dépend essentiellement des propriétés recherchées pour ce point et de la complexité algorithmique pour l'obtenir. De même, pour une position (point ayant une estampille temporelle), une position médiane peut être choisie. Une autre contrainte forte à prendre en compte est qu'il ne faut pas que cette valeur centrale soit positionnée à l'intérieur d'un obstacle. Il faut donc mieux s'appuyer sur des points existants dans le jeu de données.

Pour définir une trajectoire centrale, il a donc été décidé que le processus doit :

- étendre une approche médiane,
- utiliser un processus d'appariement robuste entre les points de lignes,
- définir une trajectoire centrale (la trajectoire médiane) comme une suite ordonnée de positions où chaque point associé est un point réel et où l'estampille temporelle est le temps médian pour arriver à ce point depuis la zone de départ.

Dans cet article, l'approche consiste à apparier les positions des trajectoires afin de définir des clusters de positions. Pour chaque cluster, une position médiane peut alors être définie.

Pour apparier les données, une mesure de similarité basée sur une distance spatiale est utilisée. Pour les lignes ou trajectoires, différentes mesures sont envisageables :

- Dynamic Time Warping (DTW) (Berndt et Clifford (1994)) qui est un algorithme de mesure de similarité entre deux séquences temporelles et peut être étendu à des trajectoires. DTW met l'accent sur les différences locales entre les points correspondants de lignes.
- La distance de Fréchet (Alt et Godau (1995)) entre deux courbes qui tient compte de l'ordre des points le long des courbes. Cette mesure s'appuie sur la distance maximale entre les points appariés. Une programmation dynamique discrète a été proposée dans (Devogele (2002)) pour deux polylignes. Cette version discrète calcule une approxima-

tion de la distance de Fréchet maximale, une distance moyenne proche de celle renvoyée par DTW et des appariements entre les points.
– La technique du Time Warp Edit Distance (Marteau (2009)) étend le concept de la distance d'édition (Chen et Ng (2004)) afin de définir la similarité entre deux séries de données temporelles (Marteau (2009)).

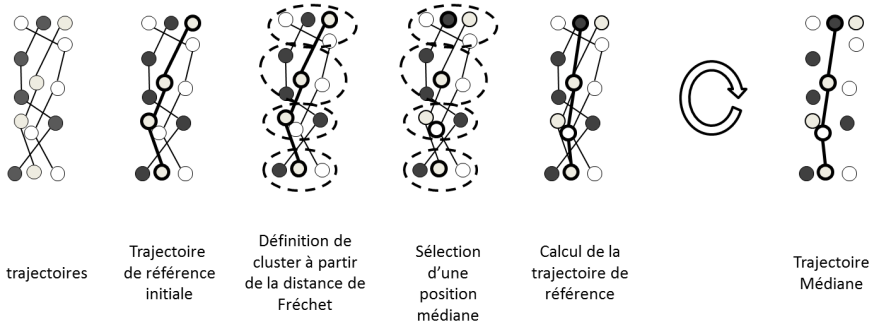

FIG. 1 – *Les différentes étapes de l'algorithme de définition de la trajectoire médiane.*

Ces méthodes sont robustes. Elles tiennent compte de l'ordre des points des lignes. De cette manière, les lignes avec des boucles ou des lignes sinueuses avec des décalages peuvent être appariées. Les étapes principales de l'algorithme de définition de la trajectoire centrale sont les suivantes :

1. Une trajectoire de référence est choisie ;
2. toutes les positions de chaque trajectoire sont appariées avec la position de la trajectoire de référence, ce processus génère une suite ordonnée de clusters de positions ;
3. pour chaque cluster, une position médiane est définie (cette position doit être une position réelle d'une trajectoire) ;
4. une nouvelle trajectoire de référence est construite à partir des positions médianes ;
5. les étapes 2 à 4 sont répétées jusqu'à ce qu'il n'y ait plus de changement dans les positions médianes ;
6. pour chaque point de la trajectoire de référence, une estampille temporelle médiane est calculée à partir du cluster de positions appariées à ce point.

La figure 1 illustre ces principales étapes avec un exemple très simple. Une trajectoire centrale est calculée pour un groupe de trois trajectoires différentes. Les positions de référence de la trajectoire de référence sont cerclées de gras. Les clusters de position sont délimités par des ellipses en pointillés. Plusieurs positions d'une même trajectoire peuvent appartenir à un même cluster. Le déroulement de cet algorithme est très proche de l'algorithme de clustering k-medoids (Kaufman et Rousseeuw (1987)).

Néanmoins, pour traiter les trajectoires de nouvelles contraintes doivent être prises en compte.

– Pour chaque cluster, au moins une position de chaque trajectoire doit être associée ;
– Toutes les positions p_{i_j} (position j de trajectoire i) sont dans au moins un cluster C_k ;
– l'ordre des positions doit être pris en compte. Si une position p_{i_j} est dans le cluster C_k, la position $p_{i_{j+1}}$ doit être dans le cluster C_k ou dans C_{k+1}.

Le processus d'appariement choisi doit répondre à toutes ces contraintes. Ainsi, la trajectoire médiane obtenue est composée d'une suite ordonnée temporellement de positions médianes minimisant la somme des écarts (Fréchet, DTW ou TWED) avec toutes les autres trajectoires du groupe. Il est important de noter que les positions qui composent cette trajectoire sont des positions réelles appartenant à différentes trajectoires du groupe.

2.2 Phase de prétraitement

La trajectoire médiane est calculée pour définir la trajectoire typique entre deux zones d'intérêt pour un même type d'objet mobile. Les trajectoires suivant ce chemin entre ces deux zones doivent donc au préalable être extraites.

L'algorithme de calcul de la trajectoire médiane proposé est basé sur des techniques d'alignement de série temporelles (DTW, Fréchet, TWED) dont la complexité dépend du nombre T de positions d'une trajectoire. L'ordre de complexité de ces algorithmes d'alignement est $\Theta(T^2)$. L'algorithme de calcul de la trajectoire médiane est itératif et réalise à chaque itération un nombre N d'appariements entre la trajectoire de référence et les N trajectoires du groupe. Ainsi pour un nombre I d'itérations réalisées, la complexité de l'algorithme de calcul de la trajectoire médiane est de l'ordre de $\Theta(I.N.T^2)$ (Petitjean et al. (2011)).

La complexité de l'algorithme de calcul de la trajectoire médiane étant directement liées au nombre T de positions des trajectoires du groupe, un filtrage et un ré-échantillonnage permet de réduire le temps de calcul. Un filtre spatio-temporel de Douglas et Peucker (Meratnia et de By (2004)) est utilisé. Le but de ce filtre est de ne retenir que les positions significatives de chaque trajectoire. Il faut donc conserver les positions qui permettent de conserver les informations sur les changements de directions ou de vitesse. Dans un deuxième temps, afin d'améliorer le processus d'appariement, si certains segments des trajectoires sont trop longs, elles sont ré-échantillonnées. Ces deux pré-traitements permettent de diminuer le temps CPU nécessaire au calcul de la trajectoire médiane.

2.3 Sélection de la trajectoire de référence

Le choix de la trajectoire initiale de référence a également un impact sur le nombre d'itérations nécessaires à l'algorithme avant de se stabiliser. Un choix judicieux de la trajectoire initiale de référence permet d'optimiser le processus.

En effet, en cas de sélection aléatoire, la trajectoire de référence peut potentiellement être une trajectoire aberrante du groupe. Ceci engendrera un nombre important d'itérations de l'algorithme et de répétitions des étapes 2 à 4.

Afin de sélectionner une trajectoire de référence proche de la trajectoire médiane, une heuristique est proposée. Cette heuristique consiste à sélectionner comme trajectoire initiale de référence celle dont la longueur est la plus proche de la valeur médiane des longueurs de toutes les trajectoires du groupe.

La complexité du calcul de la longueur d'une trajectoire est proportionnelle au nombre de segments composant cette trajectoire ($\Theta(T)$). Ainsi la complexité algorithmique de cette étape

d'optimisation est de ($\Theta(N.T)$) et n'impacte donc pas la complexité globale de l'algorithme $\Theta(I.N.T^2)$.

2.4 Processus de clustering des positions

Une fois la trajectoire de référence sélectionnée, les positions de chaque trajectoire peuvent être appariées aux positions de référence. Les différentes méthodes (DTW, distance de Fréchet et TWED) peuvent être employées. Les résultats de ces processus d'appariement sont proches et la complexité reste la même. Dans cet article, la distance de Fréchet discrète a été retenue et va être décrite. L'illustration classique de la distance de Fréchet est la suivante : un homme marche avec un chien en laisse. Il marche sur une courbe, le chien sur l'autre. Les deux peuvent faire varier leur vitesse, mais le retour en arrière n'est pas permis.

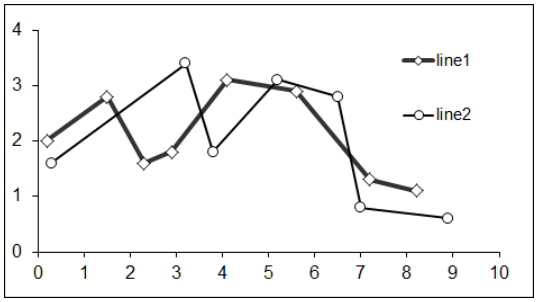

FIG. 2 – *Exemple de couple de lignes.*

La distance Fréchet entre ces deux courbes est la longueur minimale de la laisse qui est nécessaire pour assurer ce cheminement. La distance de Fréchet discrète entre L_a et L_b (d_{Fd}) est calculée de manière récursive :

$$d_{Fd}(L_a, L_b) = max \left(\begin{array}{l} d_E(L_{a_n}, L_{b_m}) \\ min \left(\begin{array}{ll} d_{Fd}(\{L_{a_1}...L_{a_{n-1}}\}, \{L_{b_1}...L_{b_m}\}) & \forall n > 1 \\ d_{Fd}(\{L_{a_1}...L_{a_n}\}, \{L_{b_1}...L_{b_{m-1}}\}) & \forall m > 1 \\ d_{Fd}(\{L_{a_1}...L_{a_{n-1}}\}, \{L_{b_1}...L_{b_{m-1}}\}) & \forall n > 1 \\ & \forall m > 1 \end{array} \right) \end{array} \right) \quad (1)$$

Où L_{i_j} est le point j de la ligne i et d_E est la distance euclidienne. A l'aide d'une méthode de programmation dynamique, ce problème récursif complexe peut être résolu en calculant une matrice intermédiaire. Ainsi, deux matrices sont employées :

- La matrice de distance euclidienne (MD). La valeur de la cellule $MD_{i,j}$ correspond à la distance euclidienne entre L_{a_i} et L_{b_j}.
- La matrice Fréchet (MF) qui permet de calculer de manière itérative la distance de Fréchet (Eiter et Mannila (1994)).

La formule utilisée pour calculer $MF_{i,j}$ est :

$$MF_{i,j} = max(d_E(L_{a_i}, L_{b_j}), min(MF_{i-1,j}, MF_{i,j-1}, MF_{i-1,j-1})) \quad (2)$$

L'exemple présenté dans la figure 2, illustre le calcul de la distance de Fréchet discrète. La ligne 1 (L_a) se compose de 8 points (L_{a_1} à L_{a_8}) et la ligne 2 (L_b) est composée de 7 points (L_{b_1} à L_{b_7}). La distance de Fréchet discrète est contenue dans la dernière case de la matrice de Fréchet discrète ($MF_{n,m}$). Le tableau 1 donne la distance de Fréchet discrète entre les deux lignes de la figure 2 (1,80).

	i	1	2	3	4	5	6	7	8
La	x	0,20	1,50	2,30	2,90	4,10	5,60	7,20	8,20
	y	2,00	2,80	1,60	1,80	3,10	2,90	1,30	1,10

Lb

j	x	y
1	0,30	1,60
2	3,20	3,40
3	3,80	1,80
4	5,20	3,10
5	6,50	2,80
6	7,00	0,80
7	8,90	0,60

Matrice de distances

MD	1	2	3	4	5	6	7	8
1	0,41	1,70	2,00	2,61	4,09	5,46	6,91	7,92
2	3,31	1,80	2,01	1,63	0,95	2,45	4,52	5,50
3	3,61	2,51	1,51	0,90	1,33	2,11	3,44	4,46
4	5,12	3,71	3,26	2,64	1,10	0,45	2,69	3,61
5	6,35	5,00	4,37	3,74	2,42	0,91	1,66	2,40
6	6,91	5,85	4,77	4,22	3,70	2,52	0,54	1,24
7	8,81	7,72	6,68	6,12	5,41	4,02	1,84	0,86

Matrice de Fréchet

MF	1	2	3	4	5	6	7	8
1	0,41	1,70	2,00	2,61	4,09	5,46	6,91	7,92
2	3,31	1,80	2,01	2,00	2,00	2,45	4,52	5,50
3	3,61	2,51	1,80	1,80	1,80	2,11	3,44	4,46
4	5,12	3,71	3,26	2,64	1,80	1,80	2,69	3,61
5	6,35	5,00	4,37	3,74	2,42	1,80	1,80	2,40
6	6,91	5,85	4,77	4,22	3,70	2,52	1,80	1,80
7	8,81	7,72	6,68	6,12	5,41	4,02	1,84	1,80

TAB. 1 – *Matrice euclidienne et Matrice de Fréchet.*

L'ensemble des couples de points est défini par back-tracking à l'aide des deux matrices. Premièrement, le couple L_{a_n} et L_{b_m} est sélectionné. Un nouveau couple est choisi parmi les trois suivants : ($L_{a_{n-1}}$, L_{b_m}), (L_{a_n}, $L_{b_{m-1}}$) et ($L_{a_{n-1}}$, $L_{b_{m-1}}$). Les valeurs $MF_{n-1,m-1}$, $MF_{n-1,m}$ et $MF_{n,m-1}$ sont utilisées en premier. Ces trois valeurs permettent de sélectionner le couple de positions à apparier. Le couple sélectionné est celui ayant la plus petite valeur dans la matrice de Fréchet. Si plusieurs cellules ont des valeurs minimales identiques, alors la distance euclidienne est utilisée. Enfin, si les distances euclidiennes sont aussi identiques, le parcours en diagonale est privilégié (couple $n-1, m-1$).

Cette étape est répétée jusqu'au couple de positions de départ des deux lignes L_{a_1}, L_{b_1}. Pour cet exemple, les valeurs $MF_{6,7}$ et $MF_{6,8}$ sont identiques (1,80), nous avons donc besoin de comparer $ME_{7,6}$ et $ME_{8,6}$. La distance $ME_{7,6}$ est inférieure à $ME_{8,6}$, le couple (L_{a_7}, L_{b_6}) est donc ajouté au début de la liste des couples de points appariés. L'étape suivante consiste à sélectionner le prochain couple de points appariés entre (L_{a_6}, L_{b_6}), (L_{a_6}, L_{b_5}) et (L_{a_7}, L_{b_5}). Le couple de points (L_{a_6}, L_{b_5}) est sélectionné en appliquant le même algorithme de sélection.

La figure 3 montre les 9 couples de points appariés correspondant aux cellules surlignées dans la matrice de la table 1. Nous pouvons remarquer que dans certains cas, plusieurs points de la ligne L_a sont appariés à un même point de la ligne L_b et vice versa. Par exemple le point L_{a_6} est assorti avec L_{b_4} et L_{b_5}.

La figure 4, illustre la robustesse de l'algorithme aux lignes ayant des boucles. Dans les figures 3 et 4, nous pouvons également remarquer que les points ne sont pas toujours appariés

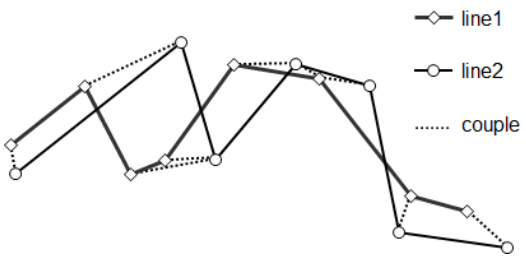

FIG. 3 – *Couples de points appariés à l'aide de la distance de Fréchet discrète.*

avec le point le plus proche. Les clusters de positions appariés sont définis à partir de ces couples de positions appariées entre la ligne de la trajectoire de référence (L_r) et les autres trajectoires. Pour chaque L_{r_i} de la trajectoire de référence, un cluster de points appariés est construit. Une position peut être dans plusieurs clusters. Par exemple, si les couples sont les suivants : (L_{r_1},L_{1_1}), (L_{r_1},L_{1_2}), (L_{r_2},L_{1_2}), (L_{r_2},L_{1_3}), (L_{r_3},L_{1_3}) et (L_{r_1},L_{2_1}), (L_{r_2},L_{2_2}), (L_{r_3},L_{2_3}), trois clusters sont définis :

- Cl_1 : Positions appariées à L_{r_1} : L_{1_1} L_{1_2} L_{2_1}
- Cl_2 : Positions appariées à L_{r_2} : L_{1_2} L_{1_3} L_{2_2}
- Cl_3 : Positions appariées à L_{r_3} : L_{1_3} L_{2_3}

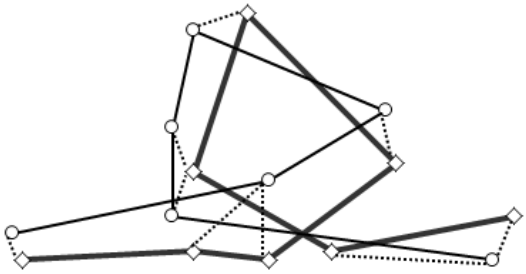

FIG. 4 – *Couples de points appariés de lignes ayant une boucle.*

2.5 Définition de la nouvelle trajectoire de référence

Pour chaque cluster une position médiane peut alors être définie. Cette position doit être une position réelle d'une trajectoire. Plusieurs médianes multidimensionnelles (Small (1990)) existent. La médiane marginale définie par (Puri et Sen (1971)) est utilisé. Cette position centrale est simple à calculer : X et Y médians sont calculés indépendamment. Pour éviter une position irréaliste, une position réelle proche de la médiane marginale doit être choisie. En effet, une position centrale synthétique pourrait être localisée dans une zone impossible (par

exemple l'emprise d'une rivière pour une trajectoire de piétons). La position la plus proche de la médiane marginale selon la distance euclidienne est choisie comme position centrale.

Des médianes multidimensionnelles ayant des meilleurs caractéristiques auraient pu être employées, mais le temps CPU aurait été plus important et les résultats sont très similaires. De même, le médoïde (Struyf et al. (1997)) n'est pas employé. Ce point est souvent utilisé comme position centrale. Malheureusement, il n'est pas robuste dès que le nombre de points aberrants est important. Une fois la position médiane de chaque cluster défini, la trajectoire de référence peut être construite en réutilisant l'ordre des clusters pour ordonner les positions médianes.

2.6 Réitération des étapes

Les étapes de constructions des clusters, de définition des positions médianes et de la trajectoire de référence sont réitérées jusqu'à ce que la trajectoire de référence converge. Cette dernière est alors considérée comme la trajectoire médiane.

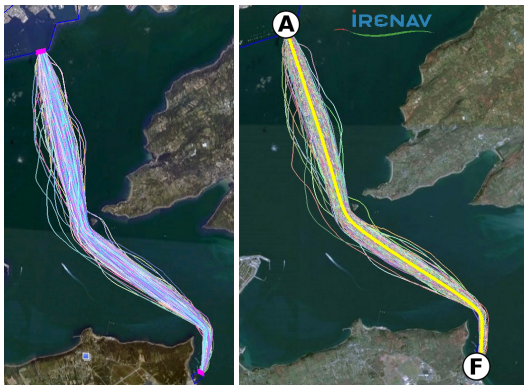

FIG. 5 – *554 trajectoires entre Brest et l'Ecole navale et la trajectoire médiane associée.*

2.7 Estampille temporelle

À l'issu de cette phase, nous obtenons une médiane qui prend en compte uniquement la dimension spatiale de la trajectoire. Pour chaque position médiane, le temps médian nécessaire pour aller de la zone de départ à cette position est calculé.

Par construction cette médiane est robuste. Les outils employés ne sont pas sensibles aux données aberrantes et tiennent compte de l'ordre des points. De même, les trajectoires peuvent avoir des boucles. En revanche, la complexité de cet algorithme est importante et demande une phase d'optimisation préalable.

3 Applications à des trajectoires réelles

3.1 Trajectoire médiane de navires

Le premier jeu de données décrit 554 trajectoires de navires à passagers entre le port de Brest et l'Ecole navale (Etienne (2011)). La durée entre deux positions varie entre 1 à 6 secondes. La figure 5 (partie gauche) illustre ce jeux de trajectoires. Ce jeu de données peut être téléchargé librement sur internet (fichier AIS Brest, www.chorochronos.org).

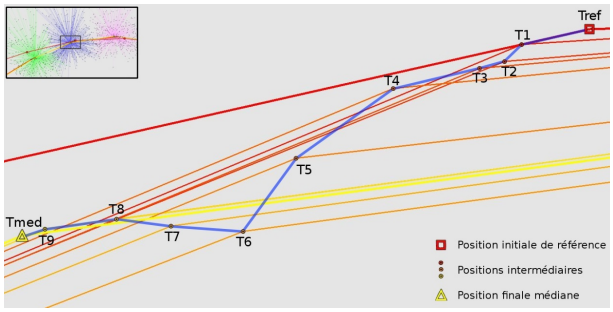

FIG. 6 – *Evolution d'une position de la trajectoire de référence.*

La figure 6 montre l'évolution du calcul d'une position de référence. 10 itérations ont été nécessaires pour définir cette trajectoire médiane. Sur la superposition du jeu de trajectoire et de la trajectoire médiane (figure 5 partie droite), nous pouvons constater que la trajectoire médiane n'est pas centrée sur le virage au milieu de la figure. En effet, la trajectoire passe plus près de la petite île. De même au niveau du virage du bas, la trajectoire médiane est proche de la rive.

3.2 Trajectoire médiane de pigeons

Le deuxième jeu de données concerne les déplacements de pigeons voyageurs. Ce jeu de données est présenté en détail dans Pettit et al. (2013). Il est composé de trajectoires de 23 pigeons différents équipés de balises GPS enregistrant leurs positions toutes les 200 ms (UTM Zone 30U). Les pigeons étudiés ont tous été entraînés de la même façon préalablement à l'expérience et connaissent déjà l'environnement aux alentours du pigeonnier. Lors de l'expérience, ces pigeons sont relâchés, un par un ou en couple, à partir d'un même endroit puis rentrent ensuite à leur pigeonnier. Ce trajet est réalisé 21 fois par pigeon afin de lui permettre de reconnaître l'environnement. Seul les 5 derniers trajets d'un pigeon sont enregistrés à l'aide du GPS.

L'étude de Pettit et al. (2013) montre que la distance moyenne entre les points appariés les plus proches des trajectoires consécutives diminue lorsque le pigeon a réalisé un nombre important de trajet identiques. Cela signifie que plus le pigeon réalise un trajet, plus il a tendance à suivre une route proche de son précédent trajet. L'article de Pettit et al. (2013) s'intéresse à la comparaison des déplacements des pigeons lorsqu'ils sont relâchés seuls ou en couple.

Trajectoires médianes

L'algorithme de Freeman et al. (2011) est utilisé pour synthétiser la trajectoire moyenne des 5 trajectoires d'un même pigeon avec un échantillonnage de 1000 points.

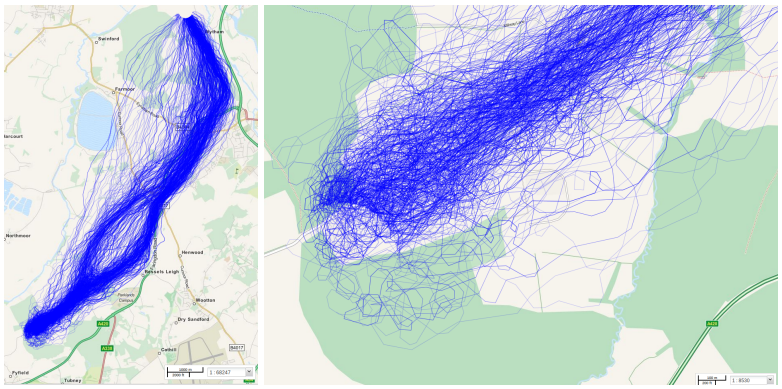

FIG. 7 – *Cluster de trajectoires de pigeons voyageurs.*

Comme indiqué précédemment, l'inconvénient de cette technique réside dans le fait qu'elle se base principalement sur une analyse purement spatiale en minimisant la distance aux plus proches voisins sans tenir compte de l'ordre de relation temporel entre positions consécutives. Ceci pose problème en cas de boucles dans les trajectoires. Dans ce jeu de données, nous avons pu observer de nombreuses petites boucles locales réalisées le long du parcours par certains pigeons (figure 7). Ces boucles sont problématiques pour l'analyse et la génération d'une trajectoire médiane car elles introduisent de nombreuses positions appariées à une même position de référence.

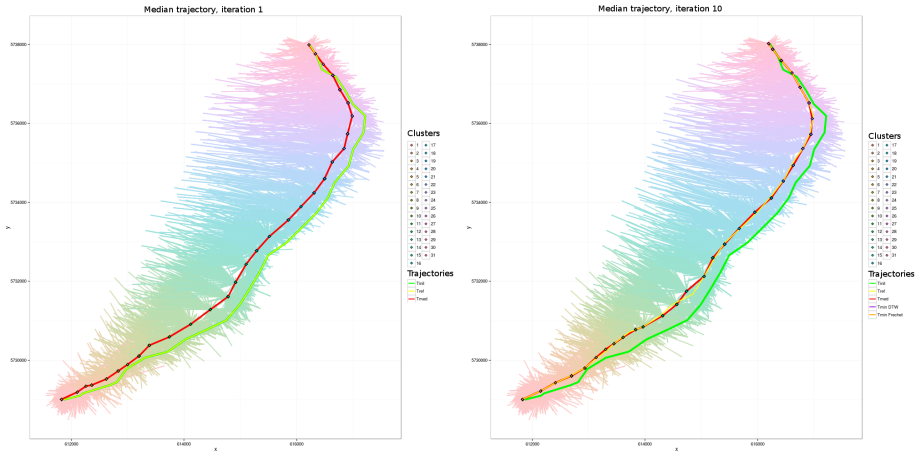

FIG. 8 – *Trajectoires médianes obtenues après 1 et 10 itérations.*

L'algorithme de calcul de la trajectoire médiane a été réalisé sur une version sous échantillonnée (une position toute les 10 s) du cluster de trajectoires de pigeons pour des raisons de temps de calculs. La trajectoire réelle initiale de référence sélectionnée dans le cluster est présentée en vert sur la figure 8. Les positions centrales des clusters de positions appariées sont représentées par des losanges colorés. Les résultats des itérations 1 et 10 sont présentés sur la figure 8. La trajectoire rouge correspond à la trajectoire médiane calculée à partir de la précédente trajectoire de référence représentée en jaune.

L'inconvénient du sous-échantillonnage est qu'il a tendance à supprimer les petites boucles locales dans les trajectoires. L'application d'un filtre spatio-temporel de Douglas et Peucker (Meratnia et de By (2004)) pourrait également minimiser les temps de calculs. En revanche, le nombre de positions multiples d'une même trajectoire affectées à un cluster serait plus important.

4 Conclusion

Cet article a proposé un algorithme robuste pour définir une trajectoire médiane. Ce motif permet d'obtenir une trajectoire centrale pour un groupe de trajectoires entre deux zones d'intérêt. Les deux exemples illustrent le potentiel et les avantages de cet algorithme sur des larges jeux de données. Cet algorithme est polynomial du fait qu'il tient compte de l'ordre des positions. Néanmoins, il peut être optimisé grâce à l'utilisation d'algorithmes de filtrage et il est également parallélisable. Les temps de calcul pour des jeux de données réels très conséquents se mesurent alors en heures. Ce motif ne devant pas être défini fréquemment, cela reste acceptable. La trajectoire médiane est un bon estimateur d'une trajectoire centrale. Afin de pouvoir distinguer les trajectoires inhabituelles, la trajectoire médiane doit être étendue à des représentations statistiques intégrant les moments d'ordre supérieurs. Des réflexions portant sur l'extension spatio-temporelle des boîtes à moustache ont été avancées dans (Etienne (2011)). Elles doivent encore être affinées. De même, pour des clusters de trajectoires qui se séparent en deux groupes, tel l'exemple des trajectoires des pigeons, des outils automatiques de détection d'une distribution bimodale des groupes de trajectoires doivent être mis en place. Finalement, une étude comparative approfondie de l'impact des différentes techniques d'alignement de trajectoires sur l'algorithme de génération de la trajectoire médiane reste à réaliser.

Références

Alt, H. et M. Godau (1995). Computing the fréchet distance between two polygonal curves. *International Journal of Computational Geometry and Applications 5*(1), 75–91.

Berndt, D. et J. Clifford (1994). Using dynamic time warping to find patterns in time series. In *AAAI94 workshop on knowledge discovery in databases*, pp. 359–370.

Buchin, K., M. Buchin, M. van Kreveld, M. Löffler, R. Silveira, C. Wenk, et L. Wiratma (2010). Median trajectories. *Algorithms–ESA 2010*, 463–474.

Chen, L. et R. Ng (2004). On the marriage of lp-norms and edit distance. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pp. 792–803. VLDB Endowment.

Devogele, T. (2002). A new merging process for data integration based on the discrete fréchet distance. In *Advances in Spatial Data Handling : 10th International Symposium on Spatial Data Handling*, pp. 167–181. Springer Verlag.

Eiter, T. et H. Mannila (1994). Computing discrete fréchet distance. Technical report, Technische Universitat Wien.

Etienne, L. (2011). *Motifs spatio-temporels de trajectoires d'objets mobiles, de l'extraction à la détection de comportements inhabituels. Application au trafic maritime.* Ph. D. thesis, Université de Bretagne Occidentale.

Freeman, R., R. Mann, T. Guilford, et D. Biro (2011). Group decisions and individual differences : route fidelity predicts flight leadership in homing pigeons. *Biology letters 7.1*, 63–66.

Kaufman, L. et P. Rousseeuw (1987). *Clustering by means of medoids*, pp. 405–416.

Marteau, P.-F. (2009). Time warp edit distance with stiffness adjustment for time series matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence 31*(2), 306–318.

Meratnia, N. et R. A. de By (2004). Spatiotemporal compression techniques for moving point objects. In *Advances in Database Technology - EDBT 2004*, Volume 2992 of *Lecture Notes in Computer Science*, pp. 561–562. Springer Berlin / Heidelberg.

Petitjean, F., A. Ketterlin, et P. Gançarski (2011). A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition 44*(3), 678–693.

Pettit, B., A. Perna, D. Biro, et D. J. Sumpter (2013). Interaction rules underlying group decisions in homing pigeons. *Journal of The Royal Society Interface 10*(89).

Puri, M. L. et P. K. Sen (1971). *Nonparametric methods in multivariate analysis*. John Wiley & Sons.

Small, C. G. (1990). A survey of multidimensional medians. *International Statistical Review/Revue Internationale de Statistique 58*, 263–277.

Struyf, A., M. Hubert, et P. Rousseeuw (1997). Clustering in an object-oriented environment. *Journal of Statistical Software 1*(4), 1–30.

Vermeulen, T. (2013). Algorithms for finding a middle trajectory. Master's thesis, Eindhoven University of Technology.

Summary

Many different huge trajectories database are nowadays available. These trajectories are collected in different contexts (pedestrians, vehicules, ships, planes, animals...). This article introduce a robust algorithm to compute median trajectory that can be applied to these different contexts. This median trajectory depicts the spatio-temporal behaviour of a cluster of trajectory moving between two zones of interest. The algorithm rely on trajectories positions matching using discrete Fréchet distance. Two important trajectories datasets illustrates how this algorithm perform on real world examples.

Caractérisation d'un parcellaire agricole : comparaison des sacs de nœuds obtenus par un chemin de Hilbert adaptatif et un graphe de voisinage

Thomas Guyet*, Sébastien da Silva**,***, Claire Lavigne***, Florence Le Ber****

* Agrocampus Ouest, Rennes
** LORIA – INRIA Grand Est, Villers-lès-Nancy
*** INRA PSH, Avignon
**** ICUBE, Université de Strasbourg/ENGEES - CNRS, Strasbourg

Résumé. Cet article porte sur la comparaison d'approches pour la fouille de motifs spatiaux. Il concerne des données de parcellaires agricoles, qui sont explorées de deux façons, par une méthode de linéarisation de l'espace, qui permet d'obtenir une séquence, et par la construction d'un graphe de voisinage. Ces représentations sont ensuite exploitées par des algorithmes d'énumération de « sacs de nœuds », c'est-à-dire un vecteur de présence / absence d'un type de nœuds dans les sous-structures de la représentation de l'espace (séquence ou graphe). Les sacs de nœuds sont ensuite comparés pour mettre en évidence la capacité qu'ont les deux méthodes proposées à caractériser le parcellaire agricole. Les résultats tendent à montrer que la linéarisation de l'espace capte la majorité de l'information – à l'exception des éléments rares – sur l'organisation du parcellaire agricole et que des outils de fouille de données utilisant ces représentations linéaires peuvent être utilisés à la place d'outils moins efficaces tels que les méthodes de fouille de graphes spatiaux.

1 Introduction

Les évolutions actuelles de l'agriculture, liées entre autres aux contraintes environnementales grandissantes et à la mise en concurrence des terres pour la production alimentaire et la production énergétique, mettent en évidence le besoin de suivre et de comprendre les différents usages des sols. Pour cela différentes sources de données sont disponibles, images satellitaires ou enquêtes de terrain, et différentes méthodes d'analyse peuvent être mises en œuvre, analyses statistiques, méthodes markoviennes, ou recherche de motifs fréquents (par ex. Castellazzi et al. (2007); Le Ber et al. (2006); Lazrak et al. (2010)). Ces études conduisent ensuite à dessiner des scénarios d'évolution ou d'optimisation, qui peuvent s'appuyer sur différentes techniques (par ex. Castellazzi et al. (2010); Sorel et al. (2010)). Ces études se basent sur des caractérisations spatiales et temporelles réalistes de l'organisation des usages des sols à l'échelle d'un territoire. Dans cet article, nous exploitons des données d'enquête sur la Zone Atelier Armorique (ZAA) [1], représentant un parcellaire agricole continu constitué de plus de

1. Une Zone Atelier est un dispositif scientifique qui implique des observations et des travaux sur plusieurs années.

7000 parcelles. Ce secteur d'étude est particulièrement intéressant pour les écologues et les agronomes car il est structuré en zones aux caractéristiques différentes allant des grandes cultures de plaine aux prairies du bocage. L'objectif de notre travail est d'étudier et de développer des méthodes de fouille de données pour caractériser les assolements (*i.e.* les allocations d'occupations du sol à des parcelles agricoles une année donnée) au travers des motifs spatiaux fréquents qui peuvent en être extraits.

La recherche de motifs spatiaux et plus généralement de motifs spatio-temporels est un domaine de recherche actif et qui trouve de nombreuses applications. On trouvera un panorama du domaine, concernant plus particulièrement les applications agro-environnementales par Selmaoui-Folcher et al. (2013). Toutes ces méthodes de fouille de données spatiales s'appuient sur une représentation de l'espace dans des structures de données dont les propriétés formelles peuvent être utilisées pour proposer des algorithmes efficaces. Il s'agit classiquement de graphes, d'arbres ou de chemins. Pour chacune de ces représentations, plusieurs constructions alternatives sont possibles. Les graphes peuvent être construits à partir des informations d'adjacence ou de proximité entre parcelles, les arbres peuvent être construits par exemple à partir des graphes par la recherche du *minimum spanning tree*, les chemins peuvent être obtenus par des marches aléatoires dans les graphes ou en parcourant l'espace par des courbes fractales, telles que la courbe de Hilbert-Peano (voir Peano (1890)).

L'utilisation de méthodes de fouilles de sous-graphes dans le graphe de voisinage présente *a priori* un intérêt supérieur à d'autres méthodes de fouille de données plus simples, notamment sur des chemins ou séquences. Mais relativement aux temps de calcul importants qu'impliquent les méthodes de fouille de graphe, il pourrait tout de même être plus intéressant de travailler sur des représentations plus simples de l'espace.

Outre les aspects calculatoires, Il faut également considérer les apports des différentes représentations en terme de résultats. En effet, les possibilités de caractérisation de l'espace par des méthodes de fouille de motifs sont dépendantes, d'une part, de l'algorithme de fouille de motifs et, d'autre part, du choix de la représentation de l'espace. Les propriétés sur les algorithmes des fouilles de motifs sont bien connues, mais peu d'intérêt a été porté au choix de la représentation de l'espace dans le cadre de la mise en place d'une extraction de motifs fréquents. En première approche, il semble que le choix de la représentation influe fortement sur les caractérisations qui peuvent être extraites des données, mais ceci mérite d'être exploré plus avant. C'est pourquoi, nous proposons une méthode pour comparer des représentations de l'espace, indépendamment de l'algorithme de fouille de données qui sera utilisé. Nous comparons ici plus particulièrement deux représentations de l'espace : un chemin de Hilbert adaptatif et un graphe spatial.

Pour comparer les caractérisations obtenues par les graphes et les chemins, l'approche classique consiste à établir des corrélations entre les localisations des motifs dans les chemins et les motifs dans les graphes. Deux motifs seront d'autant plus corrélés qu'il se situent fréquemment et exclusivement co-localisés. Nous sommes alors confrontés à la difficulté de comparer (par exemple, par une matrice de confusion) des caractérisations de natures différentes, sous-graphes et sous-chemins.

Dans le cadre de ce travail, nous introduisons les « sacs de nœuds » comme outil de caractérisation de l'organisation spatiale. Cette représentation reprend l'idée des *bags of word* initialement introduits dans le contexte de l'analyse de texte (Salton et al. (1975)) puis appliqués en analyse d'images (Weber et al. (2000)) ou de séries temporelles (Megalooikonomou

et al. (2005)). Ce modèle représente une information complexe tel qu'un texte ou une image au moyen d'une liste de termes référencés dans un dictionnaire. On dispose alors d'une représentation vectorielle d'un document qui facilite sa manipulation (apprentissage supervisé, classification, etc.). Le succès de la méthode tient dans la simplicité de cette représentation et dans son efficacité pour répondre aux tâches de classification et recherche d'information.

L'article est organisé comme suit : dans une deuxième section, nous recensons quelques méthodes de comparaison de caractéristiques d'un document ou d'un jeu de données. Dans la section 3, nous présentons les données et les deux méthodes de représentation sur lesquelles nous nous focalisons. Nous introduisons également les « sacs de nœuds » comme outil de caractérisation de l'organisation spatiale qui peut être construit à partir des chemins et des graphes. La section 4 s'attache à la comparaison des résultats obtenus. La dernière section offre une discussion suivie d'une conclusion.

2 Travaux connexes

Peu de travaux se sont portés sur la comparaison de différentes représentations structurelles de l'espace. Dans la plupart des travaux, cette comparaison s'appuie sur des résultats de performance dans une tâche de classification supervisée. Par exemple, des résultats de classification de SVM ont été utilisés pour comparer des descriptions de l'information spatiale encodées dans des noyaux SVM, pour représenter des molécules (Wale et al. (2008)), ou des données multimédia (Gosselin (2011)). Pour une tâche d'indexation de graphes, Zhao et al. (2007) comparent l'utilisation de chemins, d'arbres et de sous-graphes sur les performances et l'efficacité de l'indexation. Les approches d'évaluation en cascade (Candillier et al. (2006)) utilisent des schémas d'apprentissage supervisé pour évaluer l'apport d'une caractérisation en ajoutant cette caractérisation aux données d'entrée des données d'un classifieur. Dans le cas de la caractérisation des parcellaires agricoles, nous ne disposons pas d'étiquetage permettant d'évaluer les performances en classification. Nous restons donc dans un schéma d'évaluation fondé sur la comparaison de caractéristiques extraites, dans un contexte non-supervisé.

Parmi les approches non-supervisées, le système PANDA (*PAtterns for Next-generation DAtabase systems*) propose un cadre pour la comparaison de motifs simples et complexes (cf. Bartolini et al. (2009, 2004)). La comparaison de motifs s'appuie sur une définition unifiée des motifs par un ensemble de composants (définis en fonction des types de motifs). Cette comparaison est réalisée en recherchant les appariements possibles entre les ensembles de composants. Les exemples proposés par les auteurs ne concernent que la comparaison de motifs de même nature (par ex. *itemsets* avec *itemsets*). Or, dans notre travail, nous cherchons à mettre en regard d'une part des structures linéaires, d'autre part des structures prenant la forme de graphes. Dans les tentatives d'unification des méthodes de fouille de données structurées (cf. Nourine et Petit (2012); Négrevergne et al. (2013)), ces types de motifs n'ont actuellement pas été pris en compte.

3 Matériel et méthodes

3.1 Présentation des données

Les données sur lesquelles nous travaillons ont été collectées dans le cadre de la Zone Atelier Armorique, autour de la commune de Pleine-Fougères (35) [2]. Ces données sont sous une forme vectorielle : chaque parcelle agricole est définie par une forme géométrique (un polygone) associée à un attribut catégoriel qui donne l'occupation du sol en été (céréale, prairie, etc.). On note $\mathcal{O} \subset \mathbb{N}$ l'ensemble des identifiants des types d'occupation du sol. La figure 1 illustre la répartition des types d'occupation du sol, en nombre de parcelles, pour la zone qui nous intéresse. On remarque un très fort déséquilibre entre les classes, les prairies semées, puis les bois, maïs et céréales étant largement majoritaires.

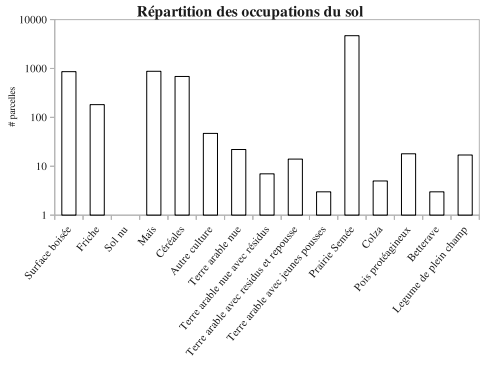

FIG. 1 – *Répartition des types d'occupation du sol en nombre de parcelles dans la zone d'étude (échelle logarithmique)*

Pour mieux mettre en avant des caractéristiques agricoles dans les motifs, nous nous limitons à la matrice agricole : seules les parcelles cultivées sont conservées. Toutes les parcelles de bâti, de route et de chemin ont été supprimées pour éviter d'extraire des motifs qui les impliqueraient. La matrice agricole comporte 7421 parcelles (cf. figure 2).

La matrice agricole a ensuite été décomposée en 14 sous-zones, numérotées selon leur position dans la grille de découpage. Ces différentes sous-zones n'ont pas les mêmes caractéristiques, en particulier en nombre de parcelles (de 189 à 1231). Les zones les plus au nord (par ex. zones 42 et 52) sont caractérisées par des grands champs de céréales, tandis que les zones au sud (par ex. zones 11 et 21) correspondent à un secteur bocager constitué de petites parcelles de prairies.

2. Les données ont été acquises par le laboratoire COSTEL et l'unité SAD-Paysage de l'INRA.

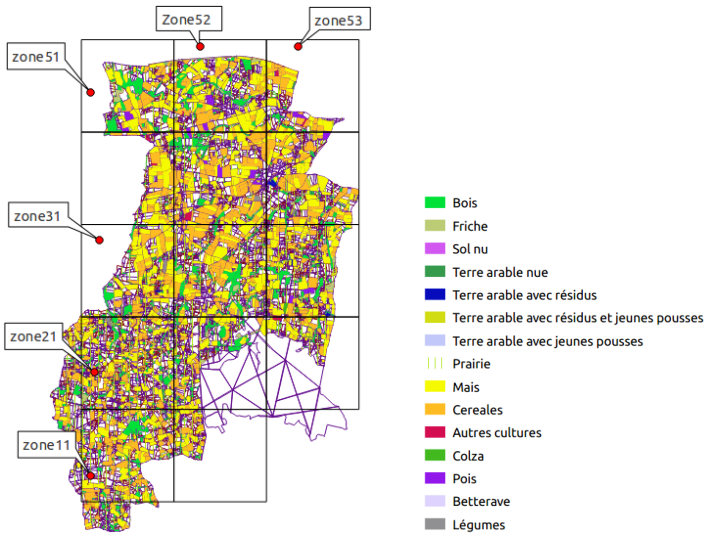

FIG. 2 – *Matrice agricole et découpage du parcellaire en sous-zones : la couleur de chaque polygone correspond à son occupation du sol*

3.2 Représentations de l'espace

3.2.1 Représentation par un graphe spatial

Un graphe $S_G = \langle V, E_G, \sigma \rangle$ est construit à partir des données au format vectoriel selon la méthode décrite par Guyet (2010). Un nœud $v \in V$ est construit pour chaque parcelle, représentée par son barycentre (cf. figure 3). Chaque nœud $v \in V$ est associé à une occupation du sol $\sigma(v) \in \mathcal{O}$ (o, j, v pour orange, jaune et vert sur la figure). Un arc $e \in E_G \subset V \times V$ lie deux parcelles voisines, c'est-à-dire connexes ou séparées par un faible espace (séparation par une route ou imprécision géométrique des données). Les arcs ne sont pas étiquetés.

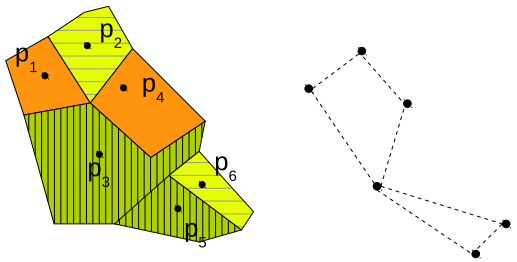

FIG. 3 – *À gauche, exemple de parcellaire avec les barycentres de parcelles et les occupations du sol (vert : hachures verticales, orange : sans hachures, jaune : hachures horizontales) ; à droite, graphe de voisinage obtenu pour l'exemple*

3.2.2 Représentation par un chemin de Hilbert adaptatif

Les courbes remplissant l'espace sont utilisées pour linéariser une information spatiale et sont couramment employées en traitement d'images mais aussi pour l'étude de phénomènes spatiaux. On utilise en particulier la courbe de Hilbert-Peano qui s'appuie sur un motif simple. Cependant, si le phénomène spatial étudié présente des irrégularités, des zones très denses et des zones vides d'information, la courbe classique, régulière, est peu appropriée. Dans ce cas, il faut soit choisir une échelle moyenne, qui délaisse une partie de l'information, comme cela a été fait par Lazrak et al. (2010), soit utiliser une méthode qui s'adapte localement à l'information : c'est ce que fait la courbe adaptative de Hilbert (CHA), dont l'algorithme est décrit par Quinqueton et Berthod (1981). Cette dernière méthode a été utilisée par exemple pour extraire les structures spatiales de linéaires agricoles (Da Silva (2013)).

La méthode de construction d'un CHA utilise un ensemble de points spatialement distribués, qui sont parcourus de manière déterministe. Sur la figure 4, les parcelles de la figure 3 ont été représentées par leurs barycentres. Le point de départ de l'algorithme est la case englobant tous les barycentres ainsi qu'une direction principale (haut-droite-bas sur l'illustration). À une étape de l'algorithme, on considère le nombre de barycentres dans la case courante : si ce nombre est égal à 1 ou 0, on passe à la case suivante ; sinon on découpe la case en quatre sous-cases.

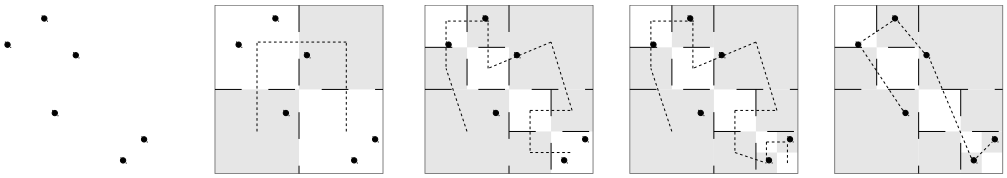

FIG. 4 – *Illustration de la construction du chemin de Hilbert Adaptatitif (CHA)*

Le chemin est ensuite simplifié par suppression des nœuds qui ne correspondent à aucun barycentre. Par cette opération, on perd donc la structure originelle du CHA et on se ramène à une simple succession de parcelles, qui est ensuite transcrite sous la forme d'une séquence d'occupations du sol. Dans l'exemple de la figure 4, on obtient la séquence suivante : $v > o > j > o > v > j$.

Pour l'unification des notations, un CHA peut être décrit de la même façon qu'un graphe, par $S_{CHA} = \langle V, E_{CHA}, \sigma \rangle$, avec un nœud par *item* de la séquence et un arc pour deux *items* successifs. L'ensemble de nœuds V est le même pour les représentations avec un graphe ou avec un chemin. Toutefois, on peut remarquer que la séquence S_{CHA} obtenue sur les mêmes données n'est pas un sous-graphe de S_G. En effet la construction de S_{CHA} ne s'appuie pas sur les informations de voisinage, mais simplement sur les positions des barycentres dans l'espace des parcelles. Certains arcs ne sont donc pas présents (par ex. entre p_3 et p_6) tandis que d'autres s'ajoutent entre les parcelles proches mais non connexes (par ex. entre p_4 et p_5).

Par ailleurs la méthode de construction d'un CHA est sensible aux deux choix initiaux de la construction du chemin ; la définition des limites de la case initiale et le choix de la direction principale (quatre possibilités). Les chemins qui sont construits peuvent varier en fonction de ces choix initiaux :

– si la direction principale est modifiée, l'ordre de décomposition des zones change et les chemins sont modifiés ;
– si les limites de la case initiale sont déplacées, la médiane de chaque case l'est également et deux parcelles – spatialement proches – dont les barycentres se situaient de part et d'autre de la médiane (horizontale ou verticale) d'une case peuvent se retrouver dans la même case. La décomposition se trouve ainsi modifiée.

Pour atténuer l'influence de ces paramètres de construction des CHA, nous construisons plusieurs chemins pour une zone d'étude en faisant varier les limites de la case initiale aléatoirement autour des limites géographiques qui définissent une zone (+- 10%) et en générant les chemins pour les 4 directions principales. Concrètement, dans l'expérience décrite ici, douze chemins ont été construits pour chacune des 14 sous-zones (3 cadrages aléatoires initiaux parcourus selon les 4 directions principales).

3.2.3 Exemple

La figure 5 illustre, à droite, un chemin de Hilbert adaptatif généré sur une sous-zone, et réduit aux barycentres des parcelles. Il comporte autant de points que de parcelles, soit 674 nœuds. La figure de gauche illustre le graphe sur la même zone, il comporte 1693 arcs.

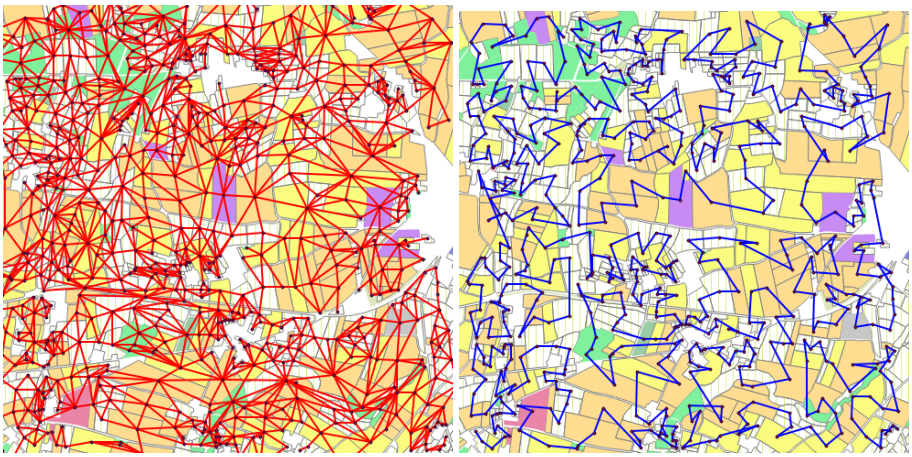

FIG. 5 – *À gauche : graphe de connexité d'une sous-zone ; à droite : chemin de Hilbert Adaptatif sur la même sous-zone*

3.3 Caractérisation d'une région par un « sac de nœuds » (SdN)

La caractérisation des régions par SdN débute par l'énumération de toutes les sous-structures d'une représentation de l'espace. Les sous-structures organisent le jeu de données initial au travers d'une vue spécifique des méthodes de fouille de motifs. Ces méthodes s'intéressent à dénombrer des motifs parmis ces sous-structures.

Définition 1 (Sous-structure)**.** *Soit $S = \langle V, E, \sigma \rangle$ une représentation de l'espace, chemin de Hilbert adaptatif ou graphe. Une sous-structure de S' est un triplet $\langle V', E', \sigma \rangle$ où $V' \subset V$, $E' \subset E$ tel que $\forall u, v \in V'$, il existe $e \in E'$ tq e soit un arc entre v et u.*

Dans le cas des CHA, on s'intéresse à des sous-séquences du CHA de taille fixe w_s. Dans le cas des graphes de voisinage, on s'intéresse à des sous-graphes contenant exactement w_g arcs. En reprenant l'exemple de la figure 3, pour $w_s = 3, w_g = 2$, les sous-structures construites sont les suivantes :

- pour le CHA : $\{p_3, p_1, p_2\}$, $\{p_1, p_2, p_4\}$, $\{p_2, p_4, p_5\}$, $\{p_4, p_5, p_6\}$
- pour le graphe : $\{p_1, p_2, p_3\}$, $\{p_4, p_2, p_3\}$, $\{p_1, p_4, p_3\}$, $\{p_4, p_6, p_3\}$, $\{p_4, p_5, p_3\}$, $\{p_1, p_6, p_3\}$, $\{p_1, p_5, p_3\}$, $\{p_6, p_5, p_3\}$.

L'énumération de toutes les structures est possible en temps output-polynomial (Bonzini et Pozzi (2007)). Contrairement à des approches de fouille de sous-séquences ou de fouille de graphes, il n'y a pas de seuil de fréquence à fixer. Pour l'énumération des sous-graphes, nous utilisons l'outil TGE de Uno (2005).

Définition 2 (« Sac de nœuds » (SdN))**.** *Soit S une représentation de l'espace, s_S une sous-structure. On appelle sac de nœuds, noté $SdN(s_S) \in 2^{\mathcal{O}}$, un vecteur de présence / absence des types d'occupations du sol dans la sous-structure s_S.*

Il faut noter qu'un sac de nœuds ne conserve que l'information de présence / absence. Cette solution a été préférée à celle consistant à compter le nombre de nœuds pour chaque type d'occupation du sol. Ce choix évite la multiplication combinatoire des sacs de nœuds, par ex. trois nœuds voisins, représentant deux parcelles vertes et une orange, seront regroupés dans le mêmes sac que trois nœuds représentant deux parcelles oranges et une verte. Sur l'exemple, les SdN construits sont donnés ci-dessous. On liste dans chaque ensemble uniquement les éléments de $\mathcal{O}$ qui sont présents ; selon des occupations répétées, les ensembles n'ont pas nécessairement tous la même cardinalité :

- pour le CHA : $\{v, o, j\}$, $\{o, j\}$, $\{j, o, v\}$, $\{o, v, j\}$
- pour le graphe : $\{o, j, v\}$, $\{o, j, v\}$, $\{o, v\}$, $\{o, j, v\}$, $\{o, v\}$, $\{o, j, v\}$, $\{o, v\}$, $\{j, v\}$.

Finalement, on dénombre trois occurrences du SdN $\{v, o, j\}$ et une occurrence de $\{o, j\}$ pour le CHA. Pour le graphe, on dénombre quatre occurrences du SdN $\{v, o, j\}$, trois du SdN $\{v, o\}$ et une du SdN $\{j, v\}$. On constate sur cet exemple que les deux méthodes extraient presque également le SdN le plus fréquent mais différemment les SdN moins fréquents : ainsi les SdN $\{v, o\}$ et $\{j, v\}$ n'ont pas été captés par le CHA ; inversement le CHA a capté un sac de nœuds $\{o, j\}$ inexistant pour le graphe. Notons qu'ici un seul chemin a été construit.

4 Comparaison des méthodes sur les données complètes

4.1 Principes

La comparaison s'effectue selon deux étapes. Lors de la première étape, on évalue la corrélation globale entre les sacs de nœuds obtenus par l'approche graphe et par l'approche chemin sur un ensemble de zones issues de la région étudiée. On utilise pour cela la mesure de corrélation de Spearman, suivie d'une mesure du χ^2 (Saporta (1990)). Dans la seconde étape, on examine les différences entre les sacs de nœuds obtenus par le graphe et par le chemin, en

particulier les sacs oubliés par la méthode CHA et la part qu'ils représentent dans l'ensemble des sacs obtenus par la recherche par graphe.

La corrélation globale est évaluée en comparant les décomptes des différents SdN pour le graphe et pour le chemin sur la zone étudiée. Le tableau 4.1 ci-dessous illustre les décomptes obtenus pour l'exemple de la figure 3. Outre le calcul de corrélation, ce tableau permet de mettre en évidence les SdN trouvés conjointement par les deux méthodes et ceux qui sont oubliés par l'une ou l'autre.

	graphe	chemin
{o, j, v}	4	3
{o, j}	0	1
{j, v}	1	0
{o, v}	3	0

TAB. 1 – *Décompte des SdN pour le graphe et le chemin de l'exemple de la figure 3*

Pour une grande zone telle que celle que nous voulons étudier (cf. figure 2), le parcellaire est découpé en sous-zones analysées séparément pour réduire le nombre de parcelles à prendre en compte et donc obtenir des résultats plus rapidement. Cela permet également d'étudier le caractère spécifique des motifs d'une zone par rapport à une autre et le comportement des deux méthodes en relation avec les caractéristiques des zones.

4.2 Résultats

L'analyse des zones par les deux méthodes a permis d'extraire pour chacune un certain nombre de sacs de nœuds distincts dont on a dénombré les répétitions.

Le nombre total de sacs (de 1 à 4 occupations du sol) distincts s'élève à 475 pour les graphes (toutes zones comprises), 206 pour les chemins : le rapport moyen du nombre de sacs différents trouvés par les graphes et par les chemins dans les différentes zones s'élève à 1,9. Toutefois, on observe deux zones remarquables, l'une où le nombre de sacs chemins est presque équivalent au nombre de sacs graphes (zone 11, rapport 1,1), l'autre au contraire où le rapport est fortement plus élevé (zone 33, rapport 2,9). Soulignons que la zone 11 est celle qui compte le moins de sacs (29 sacs extraits des graphes), c'est en effet une zone très homogène (bocage avec une large majorité de prairies semées), tandis que la zone 33, très hétérogène, est celle qui en compte le plus (301 sacs distincts extraits des graphes). Si on exclut ces deux zones de l'analyse, le rapport ne bouge pas.

Pour les graphes, le nombre d'éléments dans chaque sac varie entre 1 et 237630 (zone 32, sac {bois, prairie semée}) sur les différentes zones. Pour les sacs issus des chemins, le nombre maximum d'éléments s'élève à 3006 (zone 21, sac {prairie semée}). Le nombre moyen d'éléments dans les sacs issus des graphes s'élève à 879 (toutes zones confondues), et à 28 pour les sacs issus des chemins (soit un rapport de 1/31). En considérant ensemble tous les sacs obtenus sur les 14 zones par le CHA d'une part et par le graphe d'autre part, on obtient un indice de corrélation (Spearman) I_r(graphes, chemins) = 0,712, ce qui indique une forte corrélation ; pour le test du χ^2 (avec distribution sous H_0 simulée à cause des faibles valeurs), on obtient I_χ(graphes, chemins) = 30094, valeur de p < 0,0005, soit des distributions très

différentes. On observe les mêmes résultats sur les zones prises séparément. Les proportions des sacs estimées par les deux méthodes sont donc différentes mais corrélées.

Pour expliciter ces distributions différentes, on s'intéresse maintenant aux sacs oubliés par les CHA, on peut calculer le nombre d'éléments par sac à partir duquel un sac présent dans le graphe disparaît dans le chemin (moyenne sur les 14 zones $1215,6 \pm 855,9$) et le rapporter au nombre maximum d'éléments dans les sacs issus des graphes ($81307,0 \pm 64722,3$) ou au nombre moyen ($5891 \pm 6322,74$). Le premier taux s'élève à 2,1%, le second à 44,6%. On remarque que ces chiffres sont très variables, ils dépendent fortement du nombre de parcelles dans chaque zone et de la diversité des occupations représentées.

Si on regarde plus précisément la répartition des sacs perdus, on observe qu'il s'agit la plupart du temps de « petits » sacs, comptant moins de 100 éléments (voir tableau 2). On retrouve les comportements atypiques des zones 11 et 33. Au delà on peut séparer des zones plutôt homogènes – peu de sacs perdus et peu remplis – (zones 22, 12, 31, 21, 41) et des zones plutôt hétérogènes – sacs plus nombreux et plus remplis (zones 51, 42, 23, 52, 43). Les premières rassemblent des zones de petites parcelles de bocage (prairies très majoritaires) et des zones de grands parcelles cultivées (céréales et maïs majoritaires) : les sacs perdus par la méthode CHA représentent des occupations et des voisinages très minoritaires. Les deuxièmes sont plus diversifiées en taille de parcelles et occupations du sol : les sacs perdus peuvent correspondre à des types de voisinages relativement fréquents même si non majoritaires.

nombre	numéro de zone													
d'éléments	22	51	53	11	33	12	31	32	42	23	21	52	43	41
1-10	8	19	16	0	76	4	11	14	28	16	12	13	36	12
11-100	5	35	19	2	79	1	7	15	35	23	8	18	36	7
101-500	2	12	8	0	32	5	5	7	9	15	4	24	15	2
> 500	0	4	3	1	11	1	2	3	5	1	4	6	1	0

TAB. 2 – *Répartition des sacs perdus dans les CHA selon le nombre d'éléments dans les sacs correspondants des graphes sur les différentes zones*

5 Discussion et conclusion

La comparaison d'une analyse fondée sur les données extraites par un chemin et d'une analyse sur les données complètes a été réalisée dans le cadre des modèles de Markov en analyse d'images (Benmiloud et Pieczynski (1995)). Pour ces modèles, l'analyse fondée sur un chemin, bien que moins ajustée à la réalité des données, s'est montrée acceptable et pertinente en termes de rapidité et d'efficacité.

Pour l'étude que nous avons menée, nous aboutissons à une conclusion similaire tout en mettant en évidence certains manquements de la méthode fondée sur le CHA, qui conduit à oublier les motifs les plus rares mis en évidence par la méthode appuyée sur le graphe de voisinage. En effet, du fait de sa structure le CHA perd certains voisinages : un motif fréquent dans l'espace ne l'est pas forcément sur le CHA alors que l'inverse est vrai. La méthode est donc incomplète. De plus il est nécessaire de répéter les chemins sur une zone selon différents

cadrages afin d'obtenir une certaine « robustesse » des motifs extraits. Cependant, nous avons montré que cet inconvénient est variable selon l'hétérogénéité des zones étudiées.

Finalement, si on s'intéresse à des voisinages fréquents et à des zones relativement homogènes, la recherche de motifs par linéarisation de l'espace s'avère pertinente et efficace.

Le fait que les différences se trouvent principalement sur les motifs rares laisse également espérer des résultats de caractérisation des parcellaires intéressants par des méthodes de fouille de données. En effet, les méthodes de fouille de données ne s'intéressant qu'aux motifs fréquents, on peut s'attendre à ce que l'information extraite sur les séquences soit très similaire à celle obtenue sur des graphes tout en réduisant considérablement les temps de calcul.

Dans le futur et pour l'exemple traité, il reste à approfondir l'étude en regardant de plus près à quels ensembles d'occupation correspondent les sacs trouvés et oubliés par la méthode CHA. L'analyse peut être développée en considérant des voisinages plus éloignés, au prix d'une complexité calculatoire plus élevée, en particulier pour les graphes. Une extension du travail portera aussi sur la recherche de motifs ordonnés afin de mieux spécifier les voisinages, et d'en garder partiellement la structure spatiale, en lien avec des problématiques agro-écologiques.

Références

Bartolini, I., P. Ciaccia, I. Ntoutsi, M. Patella, et Y. Theodoridis (2004). A unified and flexible framework for comparing simple and complex patterns. In *Knowledge Discovery in Databases, PKDD 2004*, pp. 496–499. Springer.

Bartolini, I., P. Ciaccia, I. Ntoutsi, M. Patella, et Y. Theodoridis (2009). The Panda framework for comparing patterns. *Data Knowl. Eng. 68*(2), 244–260.

Benmiloud, B. et W. Pieczynski (1995). Estimation des paramètres dans les chaînes de Markov cachées et segmentation d'images. *Traitement du signal 12*(5), 433–454.

Bonzini, P. et L. Pozzi (2007). Polynomial-time subgraph enumeration for automated instruction set extension. In *Design, Automation Test in Europe Conference Exhibition*, pp. 1–6.

Candillier, L., I. Tellier, F. Torre, et O. Bousquet (2006). Cascade evaluation of clustering algorithms. In *Machine Learning : ECML 2006*, pp. 574–581. Springer.

Castellazzi, M., J. Matthews, F. Angevin, C. Sausse, G. Wood, P. Burgess, I. Brown, K. Conrad, et J. Perry (2010). Simulation scenarios of spatio-temporal arrangement of crops at the landscape scale. *Environmental Modelling & Software 25*(12), 1881–1889.

Castellazzi, M., J. Perry, N. Colbach, H. Monod, K. Adamczyk, V. Viaud, et K. Conrad (2007). New measures and tests of temporal and spatial pattern of crops in agricultural landscapes. *Agriculture, Ecosystems & Environment 118*, 339–349.

Da Silva, S. (2013). Fouille de données spatiales et modélisation de paysages. Rapport interne, INRA – INRIA Nancy Grand Est.

Gosselin, P.-H. (2011). *Apprentissage interactif pour la recherche par le contenu dans les bases multimédias*. Hdr, Université de Cergy Pontoise.

Guyet, T. (2010). Fouille de données spatiales pour la caractérisation spatiale de paysages en lien avec des fonctionnalités agro-écologiques. In *Spatial Analysis and GEOmatics (SAGEO'10)*, pp. 3.

Lazrak, E., J.-F. Mari, et M. Benoît (2010). Landscape regularity modelling for environmental challenges in agriculture. *Landscape Ecology 25*, 169–183.

Le Ber, F., M. Benoît, C. Schott, J.-F. Mari, et C. Mignolet (2006). Studying crop sequences with CARROTAGE, a HMM-based data mining software. *Ecological Modelling 191*(1), 170–185.

Megalooikonomou, V., Q. Wang, G. Li, et C. Faloutsos (2005). A multiresolution symbolic representation of time series. In *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*, pp. 668–679. IEEE.

Nourine, L. et J.-M. Petit (2012). Extending set-based dualization : Application to pattern mining. In *ECAI*, Volume 242 of *Frontiers in Artificial Intelligence and Applications*, pp. 630–635. IOS Press.

Négrevergne, B., A. Termier, M.-C. Rousset, et J.-F. Méhaut (2013). Paraminer : a generic pattern mining algorithm for multi-core architectures. *Data Mining and Knowledge Discovery 20*(3), 300–320.

Peano, G. (1890). Sur une courbe, qui remplit toute une aire plane. *Mathematische Annalen 36*(1), 157–160.

Quinqueton, J. et M. Berthod (1981). A Locally Adaptive Peano Scanning Algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-3*(4).

Salton, G., A. Wong, et C. S. Yang (1975). A vector space model for automatic indexing. *Communication ACM 18*(11), 613–620.

Saporta, G. (1990). *Probabilités, analyse des données et statistiques*. Édition Technip.

Selmaoui-Folcher, N., F. Flouvat, H. Alatrista Salas, et S. Bringay (2013). Motifs spatio-temporels – enjeux et applications à l'environnement. *Revue d'Intelligence Artificielle 2013*, 619–648.

Sorel, L., V. Viaud, P. Durand, et C. Walter (2010). Modeling spatio-temporal crop allocation patterns by a stochastic decision tree method, considering agronomic driving factors. *Agricultural Systems 103*(9), 647 – 655.

Uno, T. (2005). TGE : subtree/subgraph/connected components enumeration algorithm. URL : http ://research.nii.ac.jp/ uno/code/tge.html.

Wale, N., I. Watson, et G. Karypis (2008). Comparison of descriptor spaces for chemical compound retrieval and classification. *Knowledge and Information Systems 14*(3), 347–375.

Weber, M., M. Welling, et P. Perona (2000). Unsupervised learning of models for recognition. In *Computer Vision - ECCV 2000*, Volume 1842 of *Lecture Notes in Computer Science*, pp. 18–32. Springer Berlin Heidelberg.

Zhao, P., J. X. Yu, et P. S. Yu (2007). Graph indexing : tree + delta <= graph. In *Proceedings of the 33rd International Conference on Very Large Data Bases*, pp. 938–949.

Summary

This article focuses on the comparison of approaches for spatial patterns mining. It deals with agricultural fields, which are mined in two ways, 1) by a fractal linearization method

of space which provides a sequence of fields and 2) by the construction of a neighborhood graph. These representations are then used by enumeration algorithms to extract "bags of nodes" (BoN). A BoN is a vector representing the presence or absence of node types in the sub-structures of sequences or graph. BoNs are compared to highlight the ability of the two proposed methods to characterize the neighborhood of agricultural fields. The results suggest that the linearization of space captures most of the information – except some rare elements – about the organization of agricultural fields. Thus, data mining algorithms using these linear representations can be used instead of less efficient tools such as graphs mining algorithms.

Index

A

Abboute, Amayas 13
Alatrista-Salas, Hugo 5
Arenas, Helbert 39

B

Béchet, Nicolas 5
Boyer, Anne 27
Brun, Armelle 27

C

Cruz, Christophe 39

D

da Silva, Sébastien 67
Devogele, Thomas 55
Duck, Donald 1

E

Entringer, Gilles 13
Etienne, Laurent 55

F

Fahed, Lina 27

G

Guyet, Thomas 67

H

Harbelot, Benjamin 39

K

Kergosien, Eric 13

L

Lavigne, Claire 67
Le Ber, Florence 67

EGC 2014

Organisateurs

Sponsors