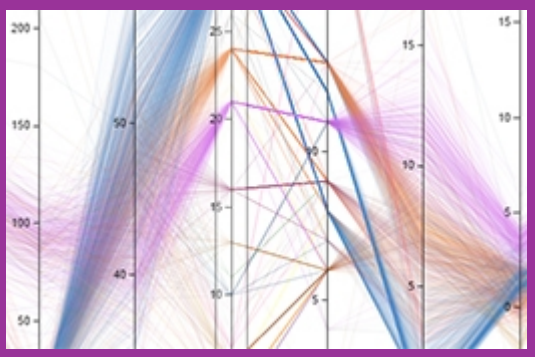
Actes de l'atelier GAST – **Gestion et Analyse de données Spatiales et Temporelles**



Thomas Guyet (AGROCAMPUS-OUEST/IRISA)
Éric Kergosien (GERiiCO, Université Lille 3)

http://evenements.univ-lille3.fr/gast-2015/

Mardi 27 janvier 2015
Luxembourg

PRÉFACE

Avec le développement des entrepôts de données de plus en plus massifs, les informations sont accumulées pour espérer que leur exploitation puisse apporter des informations profitables à terme. Dans cette perspective, la gestion des données massives de toute nature (documents textuels, images ou simple valeurs numériques, etc.) cherche de plus en plus à intégrer des estampilles temporelles ou spatiales pour enrichir leur indexation. Pour répondre aux besoins d'exploration approfondie des données temporelles et spatiales ; et d'exploitation des informations qu'elles contiennent, des méthodes et outils spécifiques sont requis. Ceux-ci doivent, en particulier, faciliter leur(s) extraction(s), leur(s) gestion(s), leur(s) représentation(s), leur(s) analyse(s) et leur(s) visualisation(s).

Les défis qui se posent sont de deux ordres : comment informer et compléter les estampilles temporelles ou spatiales et comment tenir compte de ces informations dans les processus d'extraction de connaissances. Le premier défis aborde la question de la construction, de l'évaluation, de l'enrichissement et de l'exploitation de sources d'informations variées (en particulier textuelles) pour indexer spatialement ou temporellement les données de manière efficace et précise. Ces problèmes se posent depuis longtemps pour des données statiques et se renouvellent dans un contexte de données liées et ouvertes. Le besoin d'outils de représentation et de manipulation de l'information temporelle et spatiale devient de plus en plus nécessaire pour automatiser l'enrichissement des données. Le second défi vise à exploiter la richesse de ces informations. également étudié depuis plusieurs dizaines d'années, le développement d'algorithmes dédiés à l'extraction d'information à partir de données temporelles ou spatiales posent toujours le défi de concilier l'efficacité des méthodes et la richesse de l'information qui peut être tirée des données.

Ces actes regroupent les soumissions acceptées à l'atelier GAST (Gestion et Analyse Spatio-temporelles) en 2015 dans le cadre de la conférence Extraction et Gestion des connaissances (EGC). Cet atelier s'inscrit dans une suite d'ateliers proposés à la conférence EGC depuis plusieurs années et vise à devenir un rendez-vous annuel fédérateur, convivial et scientifique riche de l'ensemble de la communauté s'intéressant à la gestion et à l'analyse de données spatiales et temporelles. Cette année, l'atelier a été organisé en trois temps : une présentation invitée de T. Devolgele sur les méthodes d'analyse de traces, puis un ensemble de présentations orales des articles retenus pour l'atelier et finalement un temps dédié à la présentation de réalisation/démonstration. L'ajout de ce temps de présentation/démonstration vise à favoriser les échanges entre les orateurs et l'auditoire et une plus-value pour toutes les personnes qui auront assistés à cette journée. Nous espérons néanmoins que le lecteur qui n'a pu y assisté trouvera toutes les informations dans les articles de ce volume.

Conférencier invité : Thomas Devolgele blablabla

L'article "Vers un système d'orchestration adaptatif et collaboratif des activités d'apprentissage en mobilité", de Nassim Dennouni, Yvan Peter, Luigi Lancieri et Zohra Slama apporte un éclairage sur l'organisation dynamique (orchestration) des

activités d'apprentissage pendant le déroulement de sorties pédagogiques. Dans la proposition, le chemin parcouru par l'apprenant est guidé par une technique de filtrage collaboratif inspirée de l'intelligence en essaim conciliant les contraintes pédagogiques avec l'autonomie des apprenants. Les propositions ont fait l'objet d'un démonstrateur opérationnel pour aider les nouveaux bacheliers à découvrir leur campus universitaire.

L'article "LC3 : un modèle spatial et sémantique pour découvrir la connaissance dans les jeux de données géospatiaux", de Benjamin Harbelot, Helbert Arenas et Christophe Cruz, s'inscrit dans le cadre de l'étude de la dynamique des territoires. L'article aborde le difficile problème de l'identification de motifs et l'extraction de la connaissance à partir d'une grande quantité d'informations liées à la couverture terrestre générées par les outils de télédétection. Dans ce sens, les auteurs proposent un modèle, appelé Land Cover Change Continuum (LC3), capable de découvrir la connaissance sur des données parcellaires et permettant l'analyse des phénomènes dynamiques à l'aide de données temporelles, spatiales et thématiques.

L'article "Exploitation de données spatiales provenant d'articles scientifiques pour le suivi des maladies tropicales", de Iana Atanassova, Marc Bertin et Tomi Kauppinen, s'intéresse à la représentation des informations géographiques extraites d'articles scientifiques. L'article propose de s'appuyer sur des outils de Traitement Automatique des Langues et de géo-codage pour extraire des données spatiales et temporelles représentant les zones du monde concernées par les maladies tropicales à différentes échelles. Le corpus utilisé dans les expérimentations est un ensemble d'articles de la revue PLOS Neglected Tropical Diseases et les données extraites sont ensuite mise à disposition sous forme de données ouvertes.

L'article "Sequential pattern mining for customer relationship management analysis", de Kiril Gashteovki, Thomas Guyet, René Quiniou, Alzennyr Gomes Da Silva et Véronique Masson, s'intéresse à l'étude des habitudes de clients afin d'améliorer le service qui leur est rendu et cela en s'appuyant sur les données issues de logiciels de gestion de la relation client (CRM). Dans ce cadre, les auteurs présentent un travail préliminaire sur le développement d'un outil interactif d'extraction de séquences fréquentes avec durées inter-événements. Les motifs extraits sont visualisés par un arbre représentant l'ensemble des séquences fréquentes afin d'en faciliter l'analyse.

L'article "Exploration de données temporelles avec des treillis relationnels", de Cristina Nica, Xavier Dolques, Agnès Braud, Marianne Huchard et Florence Le Ber, présente une méthode d'exploration de données temporelles à l'aide de treillis relationnels. Elle s'applique à un jeu de données composé de séquences de valeurs concernant des paramètres physico-chimiques et biologiques mesurés dans des cours d'eau. Les auteurs montrent sur un exemple que l'analyse relationnelle de concepts (ARC) permet de mettre en évidence l'influence dans le temps des paramètres physico-chimiques sur les paramètres biologiques.

L'article "Vers un Système d'Aide à la Décision Spatiotemporel et Multicritères pour la Surveillance Epidémiologique", de Zemri Farah, Hamdadou Djamila et Zeitouni Karine, s'inscrit dans le contexte d'une approche guidée données visant à détecter les facteurs réels et responsables de propagation des épidémies et à expliquer son émergence ou réémergence. Dans un premier temps, les auteurs proposent la mise en œu-

vre d'un système SOLAP (Spatial On Line Analysis Processing) pour la surveillance épidémiologique "EPISOLAP"permettant la localisation des foyers d'épidémies. Dans un second temps, les auteurs intègrent une approche de fouille de données spatiales et des méthodes d'aide à la décision multicritères afin d'estimer le risque sanitaire en présence d'aléa et connaissant la vulnérabilité de la population exposée.

Ces articles montrent une large étendue des recherches actuelles dans le domaine de la gestion et de l'analyse de l'information spatiale et temporelle. Nous y trouvons avec plaisir des thématiques aussi différentes que le traitement automatique des langues et la fouille de motifs temporels portant sur des données temporelles, spatiales ou spatio-temporelles. Tout ceci correspond aux intérêts premiers de GAST. Au travers de cet atelier, nous espérons que les orateurs, les auditeurs et les lecteurs constatent la complexité que continue de poser l'information temporelle et spatiale, qu'ils voient les défis qui se posent encore aux chercheurs et qu'ils s'y intéressent encore un moment ... pour que vive GAST quelques années, gast[1] !!

Nous tenons à remercier tous les auteurs pour leurs propositions d'articles ainsi que les membres du comité de lecture qui ont su respecter les contraintes imposées par le planning serré d'un atelier et dont les relectures ont été de qualité pour l'ensemble des articles. Nous remercions également chaleureusement Thomas Devogele, professeur à l'Université de Tours, pour son intervention en tant que conférencier invité à la journée GAST'2015.

En espérant que ces articles vous apporteront de nouvelles perspectives sur la gestion et l'analyse de données spatiales et temporelles, nous vous souhaitons une bonne lecture.

Thomas Guyet
AGROCAMPUS-OUEST/IRISA

Éric Kergosien
GERiiCO, Université Lille 3

[1] Interjection bretonne non-traduisible ici !

Membres du comité de lecture

Le Comité de Lecture est constitué de:

Peggy Cellier - IRISA, Rennes
Géraldine Del Mondo - INSA, Rouen
Thomas Devogele - LI, Tours
Frédéric Flouvat - PPME, NoumÃĺa
Mehdi Kaytoue - LIRIS, Lyon
Éric Kergisien - GERIICO, Lille
Florence Le Ber - ENGEES, Strasbourg
Mauro Gaio - LIUPPA, Pau
Thomas Guyet - AGROCAMPUS-OUEST/IRISA, Rennes
Simon Malinowski - FEMTO-ST, Besançon
René Quiniou - IRISA, Rennes
Jérôme Gensel - LIG, Grenoble
Mathieu Roche - CIRAD, Montpellier
Cyril de Runz - CRESTIC, Reims
Fatiha Saïs - LRI, Paris
Christian Sallaberry - LIUPPA, Pau
Nazha Selmaoui - PPME, Nouméa
Maguelonne Teisseire - IRSTEA, Montpellier

TABLE DES MATIÈRES

Présentation invitée

Analyse de trajectoires
Thomas Devogele 1

Articles sélectionnés

Vers un système d'orchestration adaptatif et collaboratif des activités d'apprentissage en mobilité
Nassim Dennouni, Yvan Peter, Luigi Lancieri, Zohra Slama 3

LC3: un modèle spatial et sémantique pour découvrir la connaissance dans les jeux de données géospatiaux
Benjamin Harbelot, Helbert Arenas, Christophe Cruz 9

Exploitation de données spatiales provenant d'articles scientifiques pour le suivi des maladies tropicales
Iana Atanassova, Marc Bertin, Tomi Kauppinen 21

Sequential pattern mining for customer relationship analysis
Kiril Gashteovski, Thomas Guyet, René Quiniou, Alzennyr Gomes Da Silva, Véronique Masson 33

Exploration de données temporelles avec des treillis relationnels
Cristina Nica, Xavier Dolques, Agnès Braud, Marianne Huchard, Florence Le Ber . . 45

Vers un Système d'Aide à la Décision Spatiotemporel et Multicritères pour la Surveillance Epidémiologique
Farah Amina Zemri, Djamila Hamdadou, Karine Zeitouni 57

Index des auteurs **73**

Analyse de trajectoires

Thomas Devogele*

* Laboratoire d'Informatique (EA 6300), Université de Tours

Résumé

De plus en plus de capteurs (GPS, GSM, WiFi, RFID), de systèmes de surveillance permettent de récupérer les trajectoires d'objets mobiles. L'analyse de ces masses de données est fondamentale pour des applications allant de la biologie à la gestion de flotte de véhicules. Elle nécessite de faire appel à des méthodes d'analyse spatiale ou de fouille de trajectoires. Cet exposé propose un état de l'art sur ces méthodes et les mesures de similarité entre trajectoires. Il se focalise sur les techniques récentes de clustering et de définition de motifs de trajectoires. Finalement, le Trajectory Box Plot (TBP) est détaillé. Il permettant de synthétiser un ensemble de trajectoires ayant le même itinéraire,. Le TBP étend les box plots classiques aux trajectoires. Il regroupe une trajectoire médiane et une boîte à moustache 3D. La trajectoire médiane décrit un déplacement typique et la boîte à moustache résume la dispersion spatiale et temporelle des trajectoires. Ces motifs sont primordiaux pour proposer une analyse visuelle des trajectoires, détecter des trajectoires inhabituelles (outliers), qualifier et prédire les déplacements.

Vers un système d'orchestration adaptatif et collaboratif des activités d'apprentissage en mobilité

Nassim DENNOUNI*,** Yvan PETER*, Luigi LANCIERI* et Zohra SLAMA**

*Équipe NOCE, Laboratoire LIFL, Université Lille 1, France
nassim.dennouni@ed.univ-lille1.fr
http://www.lifl.fr/
**Équipe ISIBA, Laboratoire EEDIS, Université Djilali Liabes, Algérie
slama@univ-sba.dz
http://www.univ-sba.dz/

Résumé. Cet article aborde le thème de l'organisation dynamique (orchestration) des activités d'apprentissage pendant le déroulement des sorties pédagogiques. L'apprentissage adaptatif offre une alternative à l'apprentissage traditionnel car il permet de mieux prendre en compte les paramètres spatio-temporels et cognitifs de l'activité. Dans notre proposition, le chemin parcouru par l'apprenant est guidé par une technique de filtrage collaboratif inspirée de l'intelligence en essaim conciliant les contraintes pédagogiques avec l'autonomie des apprenants. Pour valider notre approche, nous avons réalisé un simulateur et défini un cadre logiciel opérationnel pour aider les nouveaux bacheliers à découvrir leur campus universitaire.

1 Introduction

L'apprentissage en mobilité est devenu un sujet d'intérêt car il implique de nombreux domaines de recherche concernant des contextes d'usages et de technologie complexes. De nombreux auteurs ont montré le rôle et la difficulté d'appréhender ensemble les multiples dimensions que constituent les interactions entres individus, les types de contenus pédagogiques, le temps ou l'espace d'apprentissage (Sharples, 2009). D'autre part, l'apprentissage en mobilité a été reconnu pour sa capacité à motiver les apprenants car ils peuvent construire leurs propres connaissances en collaborant avec les autres (Hung et al., 2013). Ils sont plus autonomes et transfèrent moins la responsabilité de leur apprentissage sur l'enseignant ou le créateur de contenu (Goodyear et al., 2009). L'orchestration de l'apprentissage en mobilité fait référence à la gestion des activités en temps réel selon des graphes pédagogiques qui se situent dans des plans distincts (individus, groupe, activité, localisation, ...). Cela entraîne des contraintes telles que la segmentation des horaires ou l'aménagement des territoires. Dans ce contexte spatio-temporel évolutif, l'orchestration doit aussi couvrir l'adaptation des activités d'apprentissage, les évaluations pédagogiques, etc. (Glahn et Specht, 2010). Ceci engendre des coûts importants d'organisation et de tutorat. La gestion de ces contraintes par une orchestration centralisée rigide, est peu adaptée au contexte d'une sortie pédagogique car l'apprenant doit pouvoir garder une certaine maîtrise de ses choix d'apprentissage et de son parcours.

Dans cet article, nous présentons un nouveau style de recommandation pour une orchestration décentralisée des activités d'apprentissage en mobilité. Cette technique s'appuie sur un filtrage collaboratif exploitant l'activité antérieure des apprenants mais en prenant en compte les contraintes pédagogiques et la localisation. Elle s'inspire du mode de fonctionnement de l'intelligence en essaim qui est un domaine relativement récent de l'intelligence artificielle. Les premiers travaux de recherche dans ce domaine sont partis du constat que les colonies d'insectes sociaux (fourmis, abeilles, etc.) arrivaient à un haut degré d'organisation global avec très peu d'intelligence individuelle grâce au mode particulier de communication entre les individus de la colonie. Aujourd'hui, ces algorithmes sont utilisés dans des domaines variés (traitement d'images, aide à la décision). La suite de cet article présente les premiers résultats de cette réflexion. Nous commençons par passer en revue les différentes techniques de recommandation applicables à la conception d'un scénario d'apprentissage. Puis, après avoir présenté notre approche, nous discutons des premiers résultats obtenus par notre simulateur.

2 Travaux connexes

Les systèmes de recommandation sont connus depuis de nombreuses années pour suggérer des contenus comme des livres ou des films aux usagers, essentiellement dans un contexte de commerce en ligne (Biancalana et al., 2011). Récemment, les domaines d'applications se sont rapidement élargis, en particulier dans le domaine de la formation (Candillier et al., 2009). En plus d'un large éventail de domaines, la cible de la recommandation a aussi dépassé le simple cadre des contenus pour aborder celle des contacts, des créneaux d'agenda où de la localisation. Nous présentons dans ce qui suit une revue des principaux systèmes collaboratifs susceptibles de recommander des Objets d'Apprentissage (OAs) dans le cadre d'un scénario d'apprentissage en mobilité.

D'une manière générale, les études visant la compréhension et la modélisation des usages en mobilité et le développement des terminaux mobiles ont permis d'envisager la formation sous un nouvel angle (Benayoune et Lancieri, 2005). Les premières tentatives se sont concentrées sur la manipulation du contexte (contenu, temps, espace géographique) vu comme une extension de la recommandation de contenu. Par la suite, les travaux se sont concentrés sur une prise en compte plus fine du profil des apprenants et de la dimension pédagogique. Dès 2005, des auteurs ont appliqué le principe de l'intelligence en essaim pour optimiser un environnement d'apprentissage. Dans ces propositions, les lieux et les activités sont organisés dans un graphe d'hyperliens dont la structure peut être optimisée pour faciliter le processus d'apprentissage (Valigiani et al., 2005) ou identifier des contenus ciblés pour former un parcours d'apprentissage personnalisé (Wang et al., 2008; Kurilovas, 2013). D'autres méthodes se basent sur les évaluations faites par chaque apprenant sous forme de notes attribuées aux POIs pendant la visite (De Spindler et al., 2006; Phichaya-anutarat et Mungsing, 2014; Ye et al., 2011; Sang et al., 2012; Cheng et al., 2013; Zheng et al., 2012). Ces travaux exploitent l'historique et l'activité individuelle en cours de visite. De cette manière, les apprenants peuvent collaborer par leurs commentaires, ou en consultant les notes (annotations ou évaluations des contenus) de leurs pairs. Cependant, si cette injection de connaissance permet d'optimiser la recommandation, elle est contraignante en terme d'organisation et n'atteint que partiellement l'objectif pédagogique. Un autre problème indirectement lié au rôle du tuteur concerne le démarrage à froid, sans données d'historique permettant de produire les premières recommanda-

tions. Ce problème se pose dans la plupart des travaux de la littérature. Dans notre cas, c'est la vision de l'enseignant qui va orienter les premières recommandations. En résumé, l'état de l'art fournit plusieurs travaux qui pourraient être utilisés dans un scénario de type sortie pédagogique. Cependant, notre approche est fortement liée au contexte du domaine et doit répondre aux objectifs de l'apprentissage tout en permettant une collaboration entre les apprenants.

3 Notre contribution

L'attitude des apprenants en situation de mobilité ne peut pas être prévue en détails pendant la phase de conception du scénario pédagogique mais elle peut être supervisée et ajustée pendant son déroulement. Dans cette perspective, le POI peut être considéré comme un objet d'apprentissage (OA), associé à un emplacement. Il peut contenir, une ou plusieurs ressources (textes, hyperliens...) et des activités à réaliser (remplissage d'un questionnaire, etc.) (Dennouni et al., 2014). Ainsi, la conception du scénario peut se ramener au choix et à l'ordonnancement des POIs en fonction des objectifs pédagogiques.

Dans l'intelligence en essaim, les phéromones interviennent comme des marqueurs de fréquentation qui ont besoin d'être renforcés par des passages successifs pour rester représentatifs (Dorigo et al., 1996). A partir d'un POI, plusieurs chemins sont possibles mais le chemin optimal dépend de ce niveau d'importance. Dans notre approche, l'instructeur peut modifier ce niveau de représentativité en agissant sur le dosage des phéromones. Cette prise en compte combinée entre les orientations pédagogiques de l'enseignant et la liberté des apprenants est un aspect important de notre scénarisation pédagogique adaptative. Nous verrons plus bas comment réaliser cette combinaison. A ce stade, nous pouvons proposer l'algorithme suivant.

Début

1.Initialiser la matrice phéromone par l'instructeur de la visite.

Pour Chaque nouvelle classe ***Faire***

Pour Chaque apprenant de cette classe ***Faire***

Tantque (État courant != État cible) ***Faire***

2. P= calculer les probabilités de transition à partir du POI courant.

3. Se déplacer à l'état suivant en fonction des valeurs de P et déposer la phéromone sur le lien visité.

4. État courant = État suivant.

FinTantque

5. Évaluer et marquer la solution trouvée.

Fait

6. Déterminer la meilleure solution trouvée.

7. Déposer la phéromone sur tous les arcs appartenant à cette solution et enlever sur les autres.

Fait

Fin.

Les probabilités de transitions entre POIs, utilisées dans cet algorithme utilisent la formule 1, où S, P, PH et PR sont des matrices de dimension N (Nombres de POIs).

$$p_{ij}^{k}(t) = \frac{S_{ij}(t)^{\alpha} PH_{ij}(t)^{\beta} P_{ij}(t)^{\gamma}}{\sum_{l \in D} S_{il}(t)^{\alpha} PH_{il}(t)^{\beta} P_{il}(t)^{\gamma}} \quad (1)$$

La matrice S décrit le scénario pédagogique, les lignes représentent les POIs où se trouve l'apprenant et les valeurs en colonnes indiquent les différentes transitions possibles vers les

prochains POIs à visiter. Pendant la phase de planification de la sortie pédagogique, l'instructeur commence par localiser les POIs sur le terrain et recense les parcours qui permettent d'atteindre les objectifs pédagogiques. La matrice P représente la fréquentation des étapes du parcours. Chaque élément P_{ij} s'incrémente à chaque fois qu'un apprenant transite du POI(i) vers le POI(j). La matrice PH (phéromones) est mise à jour à chaque fois qu'un apprenant emprunte un chemin proche (voir plus loin) d'un des chemins pédagogiques (i.e. matrice S). PR (la matrice relative aux probabilités de transition) est mise à jour en fonction des valeurs des 3 matrices précédentes pour calculer en temps réel la probabilité de transition d'un POI vers un autre. La meilleure probabilité fournira la recommandation du prochain POI. Le paramètre α (respectivement β et γ) permet d'intégrer l'élévation à la puissance de chaque élément de la matrices S (respectivement PH et P).

Pour évaluer la pertinence pédagogique des chemins empruntés, notre système calcule pour chaque apprenant i l'écart pédagogique EP_i (formule 2). Cette métrique correspond au minimum des Distances de Hamming (DH) entre le Chemin Emprunté par l'Apprenant (CEA) et les Chemins Désirés par l'Enseignant ($CDE_1..CDE_n$) où n représente le nombre des chemins pédagogiques identifiés. Cette mesure représente aussi un bon moyen de comparer les chemins parcourus par les différents apprenants et peut aussi être considérée comme un indicateur de performance car elle permet de comparer les différents types de recommandation entre elles.

$$EP_i = MinDH(CEA, CDE_1, DH(CEA, CDE_2), .., DH(CEA, CDE_n) \qquad (2)$$

Pour évaluer l'influence des différentes stratégies pédagogiques (en fonction des valeurs de α, β et γ) nous avons proposé trois variantes de recommandation de POIs. La recommandation de la Solution de la Majorité (RSM) correspond aux valeurs α=0, β=1 et γ=1. Cette option correspond à du filtrage collaboratif classique. La Recommandation Pédagogique (RP) se base essentiellement sur la matrice (S). Les paramètres $\alpha = 1$, $\beta = 1$ et $\gamma = 0$ éliminent l'influence des choix individuels (P) sauf s'ils vont dans le sens des options souhaitées par l'instructeur (PH). La Recommandation Collaborative (RC) prend en compte les souhaits de l'instructeur et les interactions entre apprenants (α=1, β=1 et γ=1).

Notre simulateur permet de faire varier les paramètres α, β et γ, le nombre d'apprenants et le taux d'approbation des choix recommandés. Ce dernier point permet d'intégrer la liberté des apprenants de ne pas suivre la recommandation. Nos premiers résultats apparaissent dans la figure 1. Cette dernière montre la moyenne des écarts moy(RSM) (respectivement moy(RP) et mo(RC)) entre le chemins recommandés par la RSM (respectivement RP et RC) et les chemins désirés par l'enseignant en fonction du nombre d'apprenants (le taux d'acceptation est de 75%).

On observe que la stratégie RC s'adapte à nos contraintes car elle fournit l'écart le plus faible. Cela correspond à une pédagogie conciliant les choix du tuteur et une autonomie des apprenants. D'autre part, RP est plus performante que le filtrage collaboratif classique (RSM). Ces premiers résultats sont donc encourageants même s'ils doivent être approfondis. En particulier, nous envisageons d'étudier l'influence du niveau des paramètres α, β et γ, celle du taux d'acceptation, le nombre de POI, etc.Dans le but de valider notre approche pour l'orchestration des activités mobiles d'apprentissage, nous avons mis en œuvre un projet de type découverte de campus universitaire. Ce cadre logiciel permet d'intégrer la RC pour faire le bon filtrage collaboratif des POIs dans le cadre de notre scénario de sortie pédagogique.

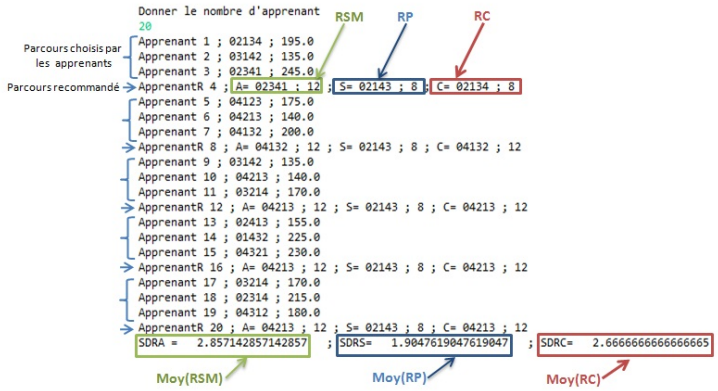


FIG. 1 – *Exemple de comparaison entre la RSM, la RP et la RC des POIs 0,1,2,3 et 4 en fonction du nombre d'apprenants à l'aide de notre simulateur.*

4 Conclusion

La logique de nos travaux repose sur une comparaison entre le modèle simulé et la réalité du comportement des apprenants. Notre simulateur permet d'évaluer l'influence des différents paramètres contextuels à l'orchestration des activités d'apprentissage en mobilité. La complexité de cette approche de la pédagogie, mêlant les contraintes de formation avec l'autonomie et la collaboration entre apprenants, peut difficilement être approchée sans outils d'analyse. Ce simulateur est destiné à éclairer des expérimentations de terrain. L'analyse des traces d'activités, les choix faits par les étudiants ainsi que les résultats du contrôle des connaissances permettront d'approfondir le modèle simulé. Dans cette perspective, nous avons développé un prototype pour aider les nouveaux bacheliers à découvrir leur campus universitaire.

Références

Benayoune, F. et L. Lancieri (2005). Toward a modelization of mobile learners behavior for the design and the evaluation of advanced training systems. *IADIS International Journal on WWW/Internet*.

Biancalana, C., F. Gasparetti, A. Micarelli, et G. Sansonetti (2011). An approach to social recommendation for context-aware mobileservices. *ACM Transactions on Intelligent Systems and Technology 4*, 325–334.

Candillier, L., K. Jack, F. Fessant, et F. Meyer (2009). State-of-the-art recommender systems. *Collaborative and Social Information Retrieval and Access*, 325–334.

Cheng, C., H. Yang, M. Lyu, et I. King (2013). Where you like to go next : Successive point-of-interest recommendation. *Proceedings of the 23 International Joint Conference on Artificial Intelligence*, 2605–2611.

De Spindler, A., R. Spindler, M. Norrie, M. Grossniklaus, et B. Signer (2006). Spatio-temporal proximity as a basis for collaborative filtering in mobile environments.

Dennouni, N., Y. Peter, L. Lancieri, et Z. Slama (2014). To a geographical orchestration of mobile learning activities. *iJIM International Journal of Interactive Mobile Technologies. ISSN : 1865-7923 8*, 35–41.

Dorigo, M., V. Maniezzo, et A. Colorni (1996). Ant system : optimization by a colony of cooperating agents. *IEEE Transactions on Systems 26*, 29–41.

Glahn, C. et M. Specht (2010). Embedding moodle intoubiquitous computing environments. *In Publications and Preprints LMedia*.

Goodyear, P., Y. Hwang, T. Lin, et I. Su (2009). Seamless connection between learning and assessment-applying progressive learning tasks in mobile ecology inquiry. *Educational Technology & Society ISSN 1436-4522 2*, 194–205.

Hung, P., Y. Hwang, T. Lin, et I. Su (2013). Seamless connection between learning and assessment-applying progressive learning tasks in mobile ecology inquiry. *Educational Technology & Society ISSN 1436-4522*, 194–205.

Kurilovas, E. (2013). Recommending suitable learning scenarios according to learners' preferences : An improved swarm based approach. *Computers in Human Behavior*.

Phichaya-anutarat, P. et S. Mungsing (2014). Hybrid recommendation technique for automated personalized poi selection. *International journal of information technology (IJIT) 1*, 01–09.

Sang, J., T. Mei, J. Tao Sun, C. Xu, et S. Li (2012). Probabilistic sequential pois recommendation via check-in data. *ACM ISBN 978-1-4503-1691-0/12/11*.

Sharples, M. (2009). Mobile learning : Small devices, big issues. *Technology-Enhanced Learning*.

Valigiani, G., Y. Jamont, et C. Bourgeois Republique (2005). Experimenting with a real-size man-hill to optimize pedagogical paths. *ACM Symposium on Applied Computing*.

Wang, T., K. Wang, et Y. Huang (2008). Using a style-based ant colony system for adaptive learning. *Expert Systems with Applications 34*, 2449–2464.

Ye, M., P. Yin, W. Lee, et D. Lee (2011). Exploiting geographical influence for collaborative point-of-interest recommendation. *ACM*, 325–334.

Zheng, W., B. Cao, Y. Zheng, et X. Xie (2012). Towards mobile intelligence : Learning from gps history data for collaborative recommendation. *ScienceDirect Artificial Intelligence*, 17–37.

Summary

This article addresses the topic of the dynamic organization (orchestration) of mobile learning activities during the conduct of field trips. In this context, adaptive learning helps to consider the spatial, temporal and cognitive parameters of the activity. In our proposal, the path travelled by the student is guided by a collaborative filtering technique inspired by Ant Colony Optimisation that conciliate the learning scope with the autonomy of learners. To validate our approach, we have developed a simulator and a dedicated software framework to help new students in the discovery of their campus.

LC3: un modèle spatial et sémantique pour découvrir la connaissance dans les jeux de données géospatiaux

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

*Laboratoire LE2I, UMR-6306 CNRS, Checksem, Université de Bourgogne
9, rue Alain Savary, 21078 Dijon, France
benjamin.harbelot@checksem.fr,
http://www.checksem.fr
**helbert.arenas@checksem.fr
***christophe.cruz@u-bourgogne.fr

Résumé. Il est nécessaire pour les gérants de territoires d'avoir un aperçu des connaissances actuelles et de l'évolution de certaines caractéristiques de la biosphère. Les outils de télédétection enregistrent une grande quantité d'informations liées à la couverture terrestre permettant l'étude de processus dynamiques. Toutefois, le volume des jeux de données nécessite de nouveaux outils pour identifier des motifs et extraire de la connaissance. Nous proposons un modèle capable de découvrir la connaissance sur des données parcellaires et permettant l'analyse des phénomènes dynamiques à l'aide de données temporelles, spatiales et thématiques. Le modèle est appelé Land Cover Change Continuum (LC3) et se base sur les technologies du Web sémantique pour permettre une représentation accrue du contexte de l'environnement géospatial et fournir des résultats d'analyses proches de ceux des experts du domaine via des opérations de raisonnement automatique. En définitive, ce modèle permet d'améliorer notre compréhension de la dynamique des territoires.

1 Introduction

L'intérêt des systèmes d'information géographique pour les organisations privés et les gouvernements est croissant. Les systèmes d'information géographique (SIG) impliquent d'énormes quantités de micro et macro-données spatio-temporelles. L'un des objectifs de développement majeur concernant les futurs SIG est de fournir une analyse de l'évolution de l'environnement géospatial. Cependant, au regard des outils actuellement disponibles, il est difficile de fournir une analyse complète permettant de comprendre pleinement les dynamiques spatio-temporelles. Dans les SIG classiques, l'analyse consiste à lier ensemble les données pour leur donner un sens. Par conséquent, l'étude de la dynamique spatio-temporelle est fréquemment réduite à une analyse statistique. Une grande majorité de ces outils s'appuie sur des bases de données relationnelles, ce qui limite fortement les possibilités de prendre en compte le contexte de l'environnement géographique. La modélisation, l'analyse et la visualisation de phénomènes

dynamiques géospatiaux a été identifié comme un enjeu majeur du développement des SIG de la prochaine génération. Un domaine d'étude pertinent dans le cadre de la dynamique spatiale concerne l'évolution de la couverture et de l'utilisation des terres plus connue sous l'acronyme "LULCC" (Land Use Land Cover Change). Au cours de milliers d'années d'existence, les humains ont modifié l'environnement. Toutefois, ce n'est que récemment que les scientifiques ont identifié une relation entre le LULCC et la modification des conditions météorologiques. Il existe plusieurs modèles actuellement utilisés pour modéliser le LULCC, cependant il y a un besoin permanent pour de nouvelles approches afin de réévaluer les modèles actuels et les améliorer.

Dans cet article, nous présenterons le domaine de la modélisation spatio-temporelle, puis nous établirons un comparatif entre le web sémantique et la base de données relationnelle comme support d'information. La section 3 présentera notre modèle défini à l'aide des outils du web sémantique. Enfin, nous détaillerons un exemple d'utilisation de notre modèle appliqué à l'étude de la dynamique du territoire Girondin.

2 Modélisation spatio-temporelle

Afin d'introduire le domaine d'étude, nous présenterons dans cette section quelques notions spatiales et temporelles.

2.1 Représenter les entités à travers le temps

Une entité spatio-temporelle est une représentation des entités du monde réel composées d'une identité, de propriétés descriptives et de propriétés spatiales. Bien que l'identité décrit une composante fixe de l'entité, les propriétes alphanumériques et spatiales peuvent varier au fil du temps et représentent la partie dynamique de l'entité. Lorsque l'identité d'une entité varie, un type particulier d'évolution intervient qui transforme l'entité spatio-temporelle en une nouvelle entité. Dans la littérature, il existe deux principaux types d'entités spatio-temporelles : 1) les objets en mouvement, comme par exemple un taxi se déplaçant dans les rues d'une ville, et 2) les objets changeant, par exemple, une région dont les frontières administratives évoluent dans le temps.

2.2 L'identité

Un concept important en ce qui concerne l'évolution des entités est l'identité. Elle caractérise l'unicité d'un objet indépendamment de ses attributs ou valeurs. Elle est la caractéristique qui distingue un objet de tous les autres. L'identité est essentielle dans la conception et la modélisation d'un phénomène. Son importance lors de la modélisation de systèmes dynamiques a été identifié par des recherches antérieures comme (Del Mondo et al., 2013), (Muller, 2002). Cependant, cette notion est très subjective car elle dépend des critères choisis par l'utilisateur pour définir l'identité d'une entité. En général, les critères pour la définition de l'identité dépendent du domaine d'étude.

2.3 Les relations de filiation

Pour pallier cette limitation, les relations de filiation définissent les liens de succession qui existent entre les différentes représentations d'un même objet à des instants différents. Dès lors, cette approche suppose une représentation détaillée de la dynamique spatio-temporelle de l'environnement. Les relations de filiation sont particulièrement adaptées à la modélisation de motifs génériques de changement comme les divisions ou les fusions d'entités. En outre, ces changements spatiaux peuvent révéler des changements dans la nature de l'entité. Pour cette raison, les relations de filiation sont intimement liée à la notion d'identité. Cette relation est essentielle pour maintenir l'identité d'une entité qui évolue et pour suivre son évolution dans le temps.

Des recherches antérieures ont identifié deux grands types de relations de filiation : la continuation et la dérivation (Harbelot et al., 2014), (Stell et al., 2011). Dans le premier cas, l'identité ne varie pas. L'entité continue d'exister, mais a subi un changement. Alors que dans le second cas, une nouvelle entité est créée à partir de l'entité "parent" lors du changement subi par l'entité. Les relations de dérivation peuvent impliquer plusieurs entités en même temps.

2.4 Étude des différentes modélisations

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 composites (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'évènements, 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 données 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.

3 Le web sémantique comme support de l'information géographique

L'objectif d'un système d'information est de stocker des données d'une manière organisée pour modéliser un domaine. Lors de la modélisation d'un domaine d'application, deux visions différentes peuvent être appliquées pour représenter la connaissance. La première est l'hypothèse du monde fermé, plus connue sous l'acronyme CWA (Closed World Assumption), et suppose que toute déclaration qui est vraie est également connue pour être vraie. Par conséquent, ce qui n'est pas actuellement connu comme étant vrai, est faux. Le contraire de l'hypothèse du monde fermé est l'hypothèse du monde ouvert ou OWA (Open World Assumption), indiquant que le manque de connaissances ne signifie pas la fausseté. Chacune de ces hypothèses est étroitement liée à des technologies spécifiques. OWA est souvent liée au Web sémantique tandis que CWA est traditionnellement associé aux bases de données relationnelles. A l'heure actuelle, beaucoup de modèles spatio-temporels utilisent des bases de données relationnelles. Dans ces travaux, nous avons souhaité étudier le Web sémantique comme support de la modélisation spatio-temporelle. Plus spécifiquement, nous baserons notre modèle sur une ontologie à l'aide du langage OWL [1].

3.1 Comparatif entre l'hypothèse du monde ouvert et du monde fermé

L'hypothèse du monde fermé invite à définir « ce qui est possible ». A l'inverse, l'hypothèse du monde ouvert permet de statuer « ce qui n'est pas possible ». Lorsqu'une ontologie OWL est vide alors tout est possible. Ce n'est que lorsque l'on contraint progressivement l'ontologie qu'elle devient plus restrictive. Dans les domaines bien définis (réservation de vols ou de livres), le modèle relationnel constitue une approche adaptée. L'hypothèse du monde fermé est performante pour faciliter la validation des données pendant les opérations de transaction. Le nombre de faits négatifs sur un domaine donné est généralement beaucoup plus grand que le nombre de faits positifs. Ainsi, dans de nombreuses applications, le nombre de faits négatifs est si grand que leur représentation explicite peut devenir pratiquement impossible. Dans de tels cas, il est plus simple et plus rapide de définir seulement tous les faits «vrais» connus plutôt que d'énumérer également les «faux». Cependant, le modèle relationnel est un paradigme où l'information doit être complète et décrite par un schéma unique. Le modèle relationnel suppose que tous les objets et les relations qui existent dans le domaine sont ceux qui sont explicitement représentés dans la base de données, et qui identifient de manière unique les noms d'objets dans ce domaine. Cela rend l'hypothèse du monde fermé et ses hypothèses connexes un mauvais choix lorsque l'on tente de combiner des informations provenant de sources multiples, pour faire face à l'incertitude ou l'incomplétude du monde.

1. http ://www.w3.org/2001/sw/wiki/OWL

3.2 Discussion

L'avantage principal de l'hypothèse du monde ouvert et du Web Sémantique (généralement associé à cette hypothèse) est de permettre aux informations d'être réutilisables. La réutilisation d'une ontologie permet d'assembler, étendre, spécialiser ou encore adapter les connaissances définies à partir d'autres ontologies. De cette manière, le Web Sémantique offre une bonne flexibilité pour permettre l'intégration de nouvelles connaissances lorsque l'application nécessite des connaissances spécifiques. Lorsqu'une ontologie est étendue, toutes les déclarations définies comme étant vraie le reste. Enfin, l'approche du Web Sémantique permet de fournir des mécanismes d'inférences afin de générer de la connaissance au sein d'une application. De son côté, l'approche relationnelle s'avère être un candidat tout indiqué pour la validation des données. Cependant, les irrégularités et l'incomplétude sont une limite à la conception du modèle relationnel. L'approche Web Sémantique dépasse ces limites en proposant une structure flexible du schéma des données. De plus, la dissociation explicite entre le schéma (TBox) et les données (ABox) offre un environnement propice pour l'interprétation des données représentées. Par conséquent, l'incomplétude du monde ouvert peut être partiellement comblée grâce à des raisonneurs capables d'auto-alimenter le système en se basant sur des contraintes. Dans beaucoup d'application du web sémantique, il est intéressant d'utiliser OWL pour définir des contraintes d'intégrités qui doivent être satisfaites par les données. Cependant, l'hypothèse du monde ouvert et le rejet de l'hypothèse du nom unique sont jusqu'alors un frein au développement des contraintes d'intégrités en OWL. En effet, les conditions définies pour lever des violations de contraintes en monde fermé, génèrent de nouvelles connaissances dans les applications de raisonnement basées sur OWL. Dès lors, il apparait nécessaire de fournir des applications hybrides combinant à la fois les raisonnements en monde ouvert avec les capacités de validation de contraintes du monde fermé.

4 Le modèle LC3

Le modèle LC3 a été proposé pour répondre au besoin de représenter les changements qui s'opèrent sur des entités spatiales. Afin de formaliser notre modèle, nous utiliserons la logique de description (Baader et Nutt (2003)).

4.1 Modélisation d'une couche spatio-temporelle

Notre modèle tente de représenter des entités dynamiques évoluant dans le temps. Ces entités sont appelées des *timeslices* dans le cadre de nos travaux. Chacun d'entre eux se définit selon quatre composantes que sont : l'identité, l'espace, le temps, et la sémantique intrinsèque de l'entité. L'identité est la composante la plus importante du modèle. Généralement, lorsque la modélisation porte sur des parcelles de terrain, une classe correspond à une couverture du sol spécifique et définit l'identité du *timeslice*, cependant d'autres critères peuvent être utilisés pour souligner l'unicité d'un *timeslice*. Les ontologies possèdent la particularité de pouvoir organiser des classes sur différents niveaux sémantiques en utilisant une taxonomie. Chaque classe décrit un concept et la taxonomie permet d'associer les *timeslices* à des concepts plus ou moins spécifiques.

$$Timeslice \sqsubseteq \top \tag{1}$$

Dans notre modèle, la classe *Timeslice* désigne le concept plus général et peut être spécialisée par une hiérarchie. Dès lors, les concepts spécifiques sont utiles pour distinguer les entités représentées tandis que les concepts généraux permettent au contraire de les regrouper sémantiquement. Un tel agencement des concepts créé un indice de profondeur au sein de la hiérarchie et permet d'évaluer les différents niveaux sémantiques qu'il est possible d'exploiter. L'équation suivante formalise la hiérarchie.

$$C_{i+2} \sqsubseteq C_{i+1} \sqsubseteq C_i \sqsubseteq ... \sqsubseteq C_0 \tag{2}$$

où C_0 correspond à la classe Timeslice et $i + 2$ représente la profondeur de la hiérarchie.

Afin de représenter le temps, nous nous basons sur l'approche suggérée par (Artale et Franconi, 1998) en considérant le domaine temporel comme une structure linéaire composée par un ensemble de points temporels ($TemporalPoint$).

$$TemporalPoint \sqsubseteq \top \tag{3}$$

Tous les éléments de type $TemporalPoint$ suivent un ordre strict, qui oblige tous les points entre deux instants temporels t_1 et t_2 à être ordonné dans le temps. En sélectionnant une paire de points temporels $[t_o, t_f]$, il est possible de définir intervalle fermé de points ordonnés définissant ainsi des intervalles de temps ($Intervalle$).

$$\begin{aligned} Interval \sqsubseteq & \top \\ Interval \equiv & \exists hasStartPoint.TemporalPoint \sqcap \\ & \exists hasEndPoint.TemporalPoint \end{aligned} \tag{4}$$

Dans certaines applications, la structure du temps peut également être définie selon des instants de temps ($TemporalPoint$). Afin de représenter à la fois des intervalles et des instants de temps, nous définissons le concept $Time$ ($\mathcal{T}$).

$$Time \equiv TemporalPoint \sqcup Interval \tag{5}$$

Dans notre modèle, nous définissons la propriété $hasTime$ qui possède comme *Domain* la classe $Time$, ainsi il est possible d'utiliser les classes $TemporalPoints$ et $Intervals$ selon les besoins de modélisation :

$$\forall hasTime.\mathcal{T} \tag{6}$$

En définitive, dans nos travaux, les différents états des entités spatio-temporelles sont représentés par la classe $TimeSlice$ ($\mathcal{TS}$). Cette classe comprend quatre composantes : 1) spatiale,

qui est la représentation géométrique de l'entité ($\mathcal{G}$) ; 2) identité, qui associe chaque état à l'entité ($\mathcal{O}$) qu'il représente ; 3) temporelle, pour décrire le temps ($\mathcal{T}$) durant lequel le *timeslice* est valide ; 4) un ensemble de propriétés alphanumériques, qui décrivent les caractéristiques de l'entité durant la période de validité du *timeslice*. L'équation 7 représente la formalisation de la classe *TimeSlice* dans le cadre de nos travaux.

$$\begin{aligned}\mathcal{TS} \equiv & \exists hasGeometry.\mathcal{G} \sqcap \\ & \exists TimeSliceOf.\mathcal{O} \sqcap \\ & \exists hasTime.\mathcal{T} \sqcap \\ & \exists hasProperties.\overline{\mathcal{TS}}\end{aligned} \tag{7}$$

4.2 Modélisation d'une transition spatio-temporelle

Sur une zone géographique possédant une couverture dynamique du territoire, une même région peut être associée à différents *timeslices* entre deux instants de temps consécutifs. Afin de représenter ces associations dans le temps, nous définissons la relation $hasFiliation$ dans notre modèle. Par conséquent, la classe $Timeslice$ est définie par un $domain$ (voir Equation 8) et un $range$ (voir Equation 9).

$$\exists hasFiliation \sqsubseteq \mathcal{TS} \tag{8}$$

$$\top \sqsubseteq \forall hasFiliation.\mathcal{TS} \tag{9}$$

Cette propriété est essentielle pour représenter le lien entre deux entités. Cependant, la connaissance portée par cette propriété est trop pauvre pour bien comprendre l'évolution. Dans notre modèle, nous proposons de spécialiser cette propriété au travers de différentes couches de connaissance. Les équations 10, 11 et 12 formalisent la hiérarchie de connaissance établie sur la base de cette propriété en utilisant le principe de subsomption.

$$hasFiliation \equiv hasContinuation \sqcup hasDerivation \tag{10}$$

Afin de spécialiser nos relations de filiation, nous intégrons les contraintes d'identité pour distinguer les objets qui ont changé leur nature de ceux dont l'identité est restée invariante. Ainsi, nous distinguons deux types de filiations que sont les continuations et les dérivations.

$$\begin{aligned}hasContinuation \equiv \\ hasEquality \sqcup hasGrowth \sqcup hasReduction \sqcup hasAnnexation \sqcup hasSeparation\end{aligned} \tag{11}$$

$$\begin{aligned}hasDerivation \equiv \\ Conversion \sqcup hasSplit \sqcup hasFusion \sqcup partOfAnnexation \sqcup partOfSeparation\end{aligned} \tag{12}$$

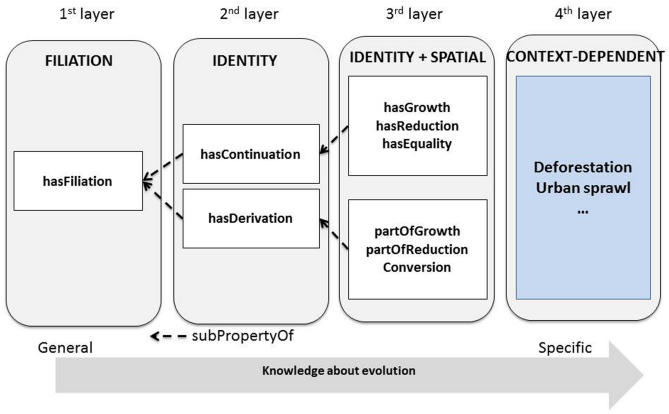


FIG. 1 – *Les différentes couches pour qualifier l'évolution*

La dernière couche générique de notre modèle ajoute des contraintes spatiales pour aboutir à des motifs détectables au sein du jeu de données. Ces motifs sont définis au travers des relations suivantes :

- **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 d'exister 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.
- **Stabilité :** aucun changement n'a eut lieu sur l'entité.
- **Conversion :** la géométrie est restée identique tandis que l'identité a varié.

5 Exemple de la gestion des risques d'inondation dans la région de la Gironde

La section précédente présente les bases de notre modèle dans l'optique de fournir un support d'analyse pour l'étude de la dynamique des territoires. Notre méthode tente de décrire et

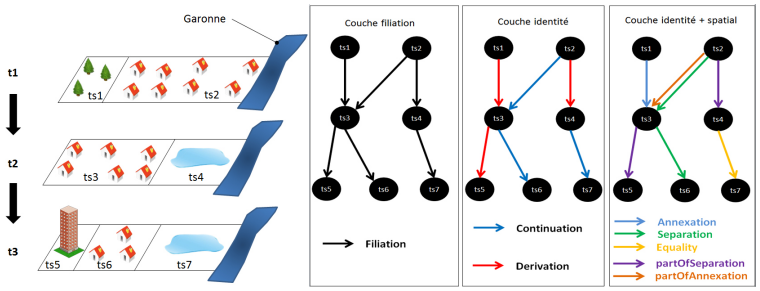


FIG. 2 – *Liste des motifs génériques détecté dans le modèle LC3*

d'offrir une meilleure compréhension de l'évolution appliquée à l'étude du territoire Girondin. Cet exemple se base sur le jeu de données Corine Land Cover qui offre une couverture de tout le territoire européen pour les années 1990, 2000 et 2006. Pour faciliter la compréhension, nous illustrerons notre approche en schématisant le territoire étudié. La Figure 2 présente notre exemple d'évolution ainsi que sa représentation sous forme de graphe au travers des différentes couches proposées par notre modèle. La dernière couche du modèle est la plus expressive. Elle permet de définir des motifs génériques d'évolution indépendants de tout domaine d'application pour servir de support à l'inférence de phénomènes. En revanche, les phénomènes peuvent être très nombreux. Par conséquent, il est impossible de les représenter au sein d'une couche générique car chacun d'entre eux est spécifique au domaine d'application. Pour contourner cette limitation, notre approche utilise les données contextuelles présentent au sein de l'ontologie que l'on appelle communément "ontologie de domaine". L'utilisation de cette ontologie de domaine couplée aux motifs génériques définis dans la dernière couche du modèle permet de fournir une interprétation de ces motifs pour aboutir à la qualification d'un phénomène. La Figure 3 présente l'ensemble des phénomènes inférés. Pour se faire, notre système se base sur l'utilisation de règles dites "d'inférences" qui permettent de générer de la connaissance automatiquement à partir de celle déjà présentes dans le système d'information. Il existe plusieurs outils dans la littérature permettant de fournir de telles règles. Dans ces travaux, nous utilisons le langage SPARQL 1.1 [2] qui offre la possibilité d'interroger, créer, supprimer ou mettre à jour les informations présente dans notre ontologie. L'algorithme 1 montre un exemple de règle permettant d'inférer un phénomène d'inondation en se basant sur un motif de séparation.

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 l'exemple, la partie de l'entité qui varie est passée de l'état de *zone urbaine* à l'état d'*étendue d'eau* et induit une réduction de la zone urbaine. En parallèle, la partie de l'entité qui n'a pas variée est également impliquée dans un motif d'*annexation* ce qui indique une *expansion urbaine* (il est à noter au passage que cette *expansion urbaine* se produit en empiétant sur la forêt connexe à la *zone urbaine* ce qui révèle un phénomène de *déforestation*). Enfin le dernier phénomène observable et qui peut être inféré sur le même principe de règle est un phénomène d'*intensification urbaine*. Ce phénomène peut être inféré lorsqu'un motif de *séparation* est observé sur une *zone urbaine* avec une partie se transformant en zone *industrielle et commerciale*.

2. http ://www.w3.org/TR/sparql11-query/

```
insert
{
        ?p checksem:Inondation ?c
}
where
{
        ?p checksem:PartOfSeparation ?c .
        ?p a checksem:ArtificialSurfaces .
        ?c a checksem:Waterbodies .
}
```

Algorithme 1 : Règle SPARQL pour inférer un phénomène d'inondation

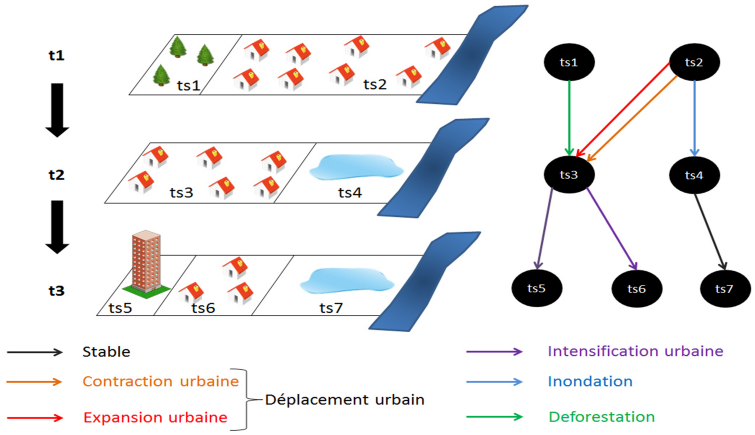


FIG. 3 – *Liste des phénomènes inférés dans le modèle LC3*

D'après les conclusions tirées de la précédente analyse, il est intéressant de noter qu'entre les instants T1 et T2, la zone urbaine est à la fois impliquée dans un phénomène d'*expansion* et de *contraction urbaine* ce qui peut sembler contradictoire de prime abord. En réalité la combinaison de ces deux phénomènes révèle un phénomène plus complexe qui est un *déplacement de la zone urbaine*. L'algorithme 2 présente la requête permettant de détecter ce type de phénomène complexe.

```
insert
{
        ?p checksem:UrbanMove ?c
}
where
{
        ?p checksem:hasAnnexation ?c .
        ?p checksem:hasSeparation ?c .
        ?p a checksem:ArtificialSurfaces .
        ?c a checksem:ArtificialSurfaces .
}
```

Algorithme 2 : Règle SPARQL pour inférer un phénomène de déplacement urbain

L'analyse fournie par notre modèle permet jusqu'ici de répondre à des questions telles que :

où, quoi, quand et comment a lieu le changement d'une entité spatio-temporelle ? Cependant, la question principale à laquelle un expert tente généralement de répondre est : *pourquoi ce changement a t-il eu lieu ?*. Les causes à l'origine d'un phénomène sont souvent nombreuses et complexes, toutefois notre modèle offre des capacités accrues d'analyses en permettant de mettre en interaction différents phénomènes connexes. Par exemple, au travers du motif de *séparation* détecté entre T1 et T2, notre modèle établie une corrélation entre la *réduction du territoire urbain* et le phénomène d'*inondation*. De manière similaire, le motif d'*annexation* détecté entre T1 et T2 met en relation le phénomène d'*expansion urbaine* et le phénomène de *déforestation*. La *contraction* et l'*expansion urbaine* pouvant être assimilées à un *déplacement urbain*, il est alors aisé de conclure que la zone inondée a engendré un déplacement de la zone urbaine. Par conséquent, la forêt située en bordure de cette zone urbaine a subit un phénomène de *déforestation*.

6 Conclusion et travaux futurs

Fort du constat que les systèmes d'information géographiques actuels possèdent des capacités limitées pour représenter le contexte de l'environnement géospatial, nous avons proposé dans cet article un modèle basé sur les technologies du web sémantique capable de fournir des outils de raisonnement afin de faire émerger de nouvelles connaissances à partir de celles déjà présentes dans le système d'information. La structure sous forme de graphe permet d'autre part de travailler aisément sur la notion de relation. La relation de filiation permet de suivre les entités dans le temps ainsi que d'accroitre la connaissance liée à l'évolution. Plusieurs couches d'expressivité ont été définie sur la base de cette relation. A terme, cette relation permet de révéler l'existence de phénomènes qui peuvent d'être mis en corrélation afin d'améliorer notre compréhension de l'évolution du territoire.

Notre modèle a été testé sur des jeux de données parcellaires uniquement. Nos futurs travaux s'orientent vers la prise en compte d'autres structures de données telles que les lignes ou les points. En reprenant l'exemple fournit à la fin de cet article, cela permettrait de mettre en corrélation les inondations avec la Garonne qui coule le long de la zone inondée. La cohabitation de ces différents types de données impose de définir des relations plus complexes ainsi que de nouveaux motifs génériques.

7 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

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

Artale, A. et E. Franconi (1998). A Temporal Description Logic for Reasoning About Actions and Plans. *Journal of Artificial Intelligence Research 9*(1), 463–506.

Baader, F. et W. Nutt (2003). Basic description logics. *Description logic handbook*, 43–95.

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

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

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

Harbelot, B., H. Arenas, et C. Cruz (2014). A semantic model to query spatial-temporal data. *Information Fusion and Geographic Information Systems (IF AND GIS 2013)*, 75–89.

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, 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.

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. *Spatial Information Theory*, 148–167.

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

Summary

It is necessary for land managers to be provided both an overview of existing knowledge and changes about the evolution of certain characteristics of the biosphere. Remote sensing tools monitore a large amount of land cover informations for studying dynamic processes. However, the volume of data sets requires new tools to identify patterns and extract knowledge. We propose a model enabling the knowledge discovery on landparcel data by analyzing dynamic phenomena using temporal, spatial and thematic data. The model is called Land Cover Change Continuum (LC3) and is based on Semantic Web technologies to enable a better representation about the context of the geospatial environment and provide analysis results close to those of experts ones via reasoning systems. Ultimately, this model can improve our understanding of spatio-temporal dynamics in land territories.

Exploitation de données spatiales provenant d'articles scientifiques pour le suivi des maladies tropicales

Iana Atanassova*, Marc Bertin**, Tomi Kauppinen***

*Centre Tesnière, Université de Franche-Comté, Besançon, France
iana.atanassova@univ-fcomte.fr
**CIRST, Université de Québec à Montréal, Montréal, Canada
bertin.marc@gmail.com
***Department of Computer Science, Aalto University School of Science, Finlande
tomi.kauppinen@aalto.fi

Résumé. Nous nous intéressons à la représentation des informations géographiques extraites d'articles scientifiques. En utilisant des outils de Traitement Automatique des Langues et de géo-codage, nous avons traité la revue PLOS Neglected Tropical Diseases afin de produire des données spatiales liées aux articles sous forme de Linked Data. Les résultats montrent une exploitation spatiale et temporelle des données représentant les zones du monde concernées par les maladies tropicales à différentes échelles.

1 Introduction

Ces dernières années, de nombreuses recherches sont menées sur l'exploitation de sources textuelles en complément des systèmes de surveillance épidémiologique (Arsevska et al., 2014). L'extraction d'informations spatiales d'un document reste une tâche non triviale comme le souligne Tahrat et al. (2013). Nous proposons dans cet article une approche pour étudier les maladies tropicales en fonction des informations spatiales extraites d'articles scientifiques. En effet, certaines revues thématiques permettent d'envisager des traitements géographiques à partir d'une étude de corpus des métadonnées liées à l'article. Dans la majorité des cas, les mots-clés ne donnent pas d'informations sur les lieux ou les périodes étudiés. L'enrichissement des méta-données classiques avec des données spatiales et temporelles provenant des textes permettra d'analyser les études scientifiques du point de vue spatio-temporel (Kauppinen et al., 2013). Une telle approche permettrait de répondre à des questions comme *« Quelles régions ont été étudiées en relation avec la Dengue ? »* ou *« Quels lieux ont des co-occurrences dans des études scientifiques dans tel ou tel domaine ? »*. Une problématique plus générale consiste à rendre compte de la nature d'un objet via son association avec des attributs géographiques, tels que des noms de villes, pays, régions et géo-coordonnées. Par exemple, la présence de noms de lieux explicites et de géo-coordonnées associées à des objets, tels que des pages web (Wang et al., 2005b; Borges et al., 2007; Inoue et al., 2002), rendent possible des analyses de ces objets en tant que phénomènes spatiaux. De nombreux travaux proposent des techniques d'enrichissement de ressources par l'établissement de relations avec des informations spatiales

dans des contextes différents (Jones et al. (2002); Wang et al. (2005a); Bucher et al. (2005); Purves et al. (2007); Markowetz et al. (2005)).

Dans cet article nous proposons une méthodologie permettant d'identifier et d'extraire des localisations à partir des publications scientifiques. Notre hypothèse de travail est que l'extraction, la catégorisation et l'affectation de ces données spatiales permettront, à travers un enrichissement des méta-données, d'exprimer les propriétés spatiales des études scientifiques, en établissant des liens avec des localisations géographiques. Notre approche utilise des outils issus du Traitement Automatique de Langues afin de traiter les informations spatiales dans des documents en plein texte.

Notre principale contribution dans cet article est de proposer une lecture des localisations géographiques à travers le filtre des articles scientifiques. Nous illustrons différentes méthodes de visualisations faisant partie des applications possibles. Ces résultats sont destinés à faciliter des décisions pour cibler de nouvelles études, à la veille scientifique et au développement de systèmes de recherche d'information orientés autour des données spatiales.

2 Méthodologie

L'extraction des termes de localisation à partir de corpus d'articles scientifiques permet de représenter la dimension spatiale des études afin de répondre à la question *« Quelles sont les régions / les localisations géographiques / ... qui sont liées à cette étude ? »*. Pour cela, nous cherchons à lier les articles, considérés comme des objets, à des informations spatiales. Dans cette approche, nous excluons les méta-données, telles que les affiliations et les adresses des auteurs. Après l'identification des termes de localisation dans des textes, notre méthode s'appuie sur une désambiguïsation des termes extraits et l'identification des géo-coordonnées afin de produire des visualisations.

2.1 Corpus

Comme corpus d'expérimentation, nous avons utilisé les articles en plein texte du journal *PLoS Neglected Tropical Diseases (PLOS NTDs)*, qui est publié par la *Public Library of Science* [1] et disponible en libre accès. PLOS NTDs publie des articles de recherche examinés par des pairs, dans le domaine des maladies tropicales peu étudiées et traitent de leurs aspects scientifiques, médicaux et sanitaires. Le choix de ce journal a été motivé par sa portée — *"a group of poverty-promoting chronic infectious diseases, which primarily occur in rural areas and poor urban areas of low-income and middle-income countries"* — qui suggère que le fait de rendre explicites les localisations présentes dans les textes en tant que métadonnée apportera des informations précieuses sur les maladies en question.

L'observation montre que les articles de recherche publiés dans PLOS NTDs contiennent un grand nombre d'informations géospatiales. Tous les articles sont disponibles en accès libre en format XML en utilisant le schéma *Journal Article Tag Suite (JATS)* [2]. Nous avons traité l'ensemble complet de 1 872 articles de recherche qui ont été publiés sur une période de 6 ans dans PLOS NTDs, d'août 2007 à août 2013.

1. `http://www.plosntds.org/`
2. Ce standard est une application des normes NISO Z39.96-2012. JATS est une extension de *NLM Archiving and Interchange DTD* (`http://jats.nlm.nih.gov`)

2.2 Extraction des termes de localisation géographique

Afin d'implémenter l'extraction des termes de localisation géographique des textes, nous considérons les étapes suivantes (voir figure 1) :

1. identification des termes de localisation dans des textes ;
2. filtrage des noms de pays ;
3. géocodage des termes identifiés ;
4. analyse des données spatiales à travers des visualisations et analyse des corrélations.

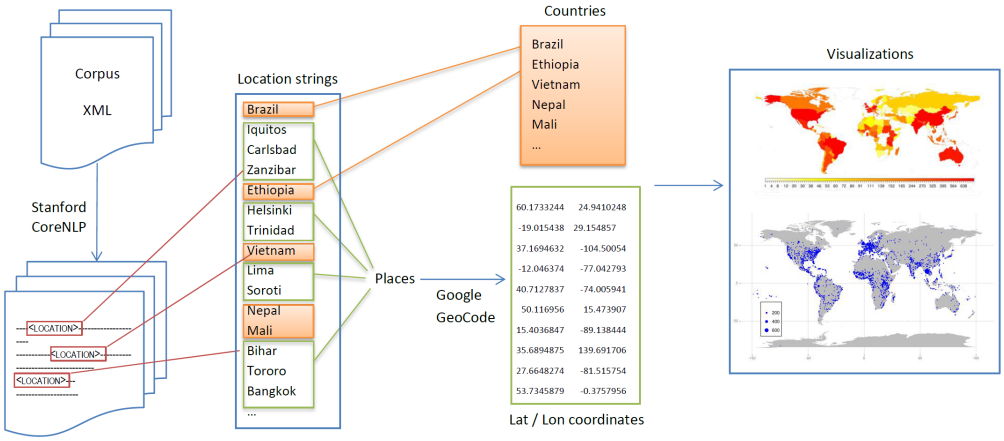


FIG. 1 – *Etapes de l'extraction d'informations spatiales des textes*

Afin d'identifier les localisations, nous nous appuyons sur une Reconnaissance d'Entités Nommées (REN) [3], qui est une technique bien connue en Traitement Automatique des Langues et qui utilise des méthodes d'apprentissage automatique. Le but des systèmes de REN est d'identifier les entités nommées dans des textes et les annoter avec des catégories différentes, telles que *location*, *organization*, *proper name*, *date*. Nous avons réalisé l'extraction des termes de localisation en utilisant l'outil de REN de Stanford CoreNLP [4] (Manning et al., 2014).

Dans notre étude, nous nous intéressons particulièrement aux différents types de localisations : pays, villes ou régions. Par ailleurs, les informations géographiques présentes dans des textes sont souvent incomplètes et ambiguës, ce qui implique des traitements supplémentaires après l'extraction afin d'assurer la qualité des données générées. Par exemple, nos premières expérimentations ont révélé qu'une partie des termes extraits en tant que localisations par le système de REN sont des noms de virus et ne correspondent pas aux localisations géographiques. Des traitements ont été mis en place à des fins de nettoyage :

1. vérification si le terme extrait est le nom d'un pays. Si oui, nous pouvons établir un lien entre ce pays et l'article traité.

3. voir par exemple Cucerzan et Yarowsky (2002); Zhou et Su (2002)
4. Stanford CoreNLP Named Entity Recognizer

2. si le terme extrait n'est pas le nom d'un pays, nous pouvons émettre l'hypothèse qu'il est le nom d'une ville ou une autre localisation plus petite qui pourra être représentée par un point. Dans ce cas, nous utilisons Google GeoCode API [5] afin de retrouver la latitude et longitude.

La distinction entre les pays et les autres localisations (villes, régions) est nécessaire notamment pour des fins de visualisation. En effet, nous supposons que les localisations qui ne sont pas des pays correspondent à des territoires relativement petites et peuvent être représentées par des points sur une carte géographique à l'échelle mondiale. Ainsi, nous avons deux principaux types de données spatiales pour chaque article : pays mentionnés dans l'article et autres termes de localisation dans le texte. Une méthodologie peut être considérée afin de prendre en compte la taille des territoires des localisations identifiées de façon précise pour pouvoir obtenir des visualisations plus fines.

Dans la première étape nous avons utilisé la liste de pays traitée par la norme ISO 3166 [6], qui fournit une correspondance entre les noms des pays en anglais et des codes des pays reconnus internationalement.

La deuxième étape de ce traitement a pour fonction, en plus de l'identification des géo-coordonnées, de filtrer les expressions pour lesquelles Google GeoCode API ne retourne pas de résultats. La table 1 présente le nombre de pays, localisations et géo-coordonnées identifiés dans le corpus. Sur un total de 24 660 occurrences de termes de localisations, autour de 85% (20 990) ont été convertis en géo-coordonnées avec succès. Cette étape permet d'éliminer une partie des entités nommées qui ont été identifiées par CoreNLP et qui ne correspondent pas à des noms de lieux. Cependant, cette méthode ne nous permet pas la distinction entre les différents sens des noms de localisations ambiguës.

TAB. 1 – *Occurrences de pays et termes de localisations dans le corpus*

	Pays	Localisations (CoreNLP)	Localisations avec géo-coordonnées (Google GeoCode API)
Total	24 197	24 660	20 900
Nb moyen par article	12,93	13,17	11,16

Remarquons que certains localisations sont présentes avec une très grande fréquence dans le corpus. De plus, les occurrences de seulement 15 pays et 200 termes de localisations représentent la moitié de toutes les occurrences. La table 2 présente le nombre total de pays et termes de localisation distincts dans le corpus.

TAB. 2 – *Pays et termes de localisation distincts*

Pays	Termes de localisation	Géo-coordonnées
150	4 168	3 249

5. http://maps.googleapis.com/maps/api/geocode/
6. http://www.iso.org/iso/country_codes.htm

2.3 Génération de Linked Data

Pour chaque article, nous avons obtenu une liste de pays et une liste de localisations avec leurs coordonnées latitudinales et longitudinales. Cela nous permet de générer des données sous forme de Linked Data, qui représentent les informations géographiques issues de l'article. Ce format a été choisi afin de pouvoir fournir des données réutilisables pour des applications externes. Elles peuvent être exploitées dans des visualisations, mais également en tant que ressources dans des systèmes de recherche d'informations ou d'extraction de connaissances. Les données sont accessibles à l'adresse suivante : `http://linkedscience.org/data/spatialaboutness/`.

La génération de ces données en tant que ressources indépendantes établit un lien fort entre la géographie et les études de terrain. La publication de ces résultats sous forme de Linked Data permet de mettre en place une base de connaissances incrémentale afin de gérer et partager les données spatiales pour l'utilisation par la communauté via des services SPARQL.

Nous avons utilisé le vocabulaire *Linked Science Core Vocabulary (LSC)* [7], qui a été spécifiquement créé pour représenter des propriétés des recherches scientifiques, y compris les données temporelles et spatiales liées aux études scientifiques. En se basant sur les termes de LSC, nous avons converti les informations géographiques extraites des articles en des triplets RDF.

```
@prefix ns:<http://linkedscience.org/lsc/ns#> .

journaldoi:journal.pntd.0000321
    lsc:isAboutRegion
        aboutloc:Akonolinga, aboutloc:Ayos, aboutloc:Bu,
        aboutloc:Nyong, aboutloc:Thailand;
    a lsc:Research .

journaldoi:journal.pntd.0000355
    lsc:isAboutRegion
     aboutloc:Mahottari, aboutloc:Muzaffarpur, aboutloc:Pune,
     aboutloc:Rajshahi, aboutloc:Vaishali, aboutloc:Varanasi;
    a lsc:Research .
```

FIG. 2 – *Exemples de triplets RDF qui décrivent la dimension spatiale des études scientifiques*

La figure 2 montre deux exemples de triplets RDF représentés dans la syntaxe Turtle [8]. Le champ *journaldoi* permet d'identifier chaque article de façon unique par son *Digital Object Identifier (DOI)*. Le premier exemple exprime le fait que l'article identifié par *journal.pntd.0000321* est une étude scientifique qui concerne les régions d'Akonolinga, Ayos, Bu, Nyong et Thailand.

7. `http://linkedscience.org/lsc/ns/`
8. `http://www.w3.org/TeamSubmission/turtle/`

3 Résultats

Les données spatiales que nous avons obtenues à partir des articles peuvent être exploitées dans plusieurs types d'analyses. Nous proposons différentes visualisations géographiques, ayant pour objectif principal de mettre en évidence les zones géographiques qui concentrent le plus grand nombre de recherches ainsi que les zones peu étudiées. Les visualisations ont été générées en utilisant *R-studio* et la librairie *rworldmap*.

Nous considérons deux représentations principales : une au niveau des pays et une sur une échelle plus fine. De plus, comme le corpus que nous avons traité est homogène et couvre six ans, nous pouvons examiner et visualiser les tendances, en prenant en compte la date de publication des articles.

3.1 Analyse au niveau des pays

La visualisation sur la figure 3 montre une carte thermique des pays mentionnés dans les articles du corpus. Comme le montre cette représentation, certaines régions ont fait l'objet d'un très grand nombre d'études, par exemple l'Afrique du Sud, alors que d'autres régions comme l'Afrique Centrale sont peu étudiées. Cette carte donne ainsi un premier aperçu des régions constituant des centres d'intérêt pour le journal.

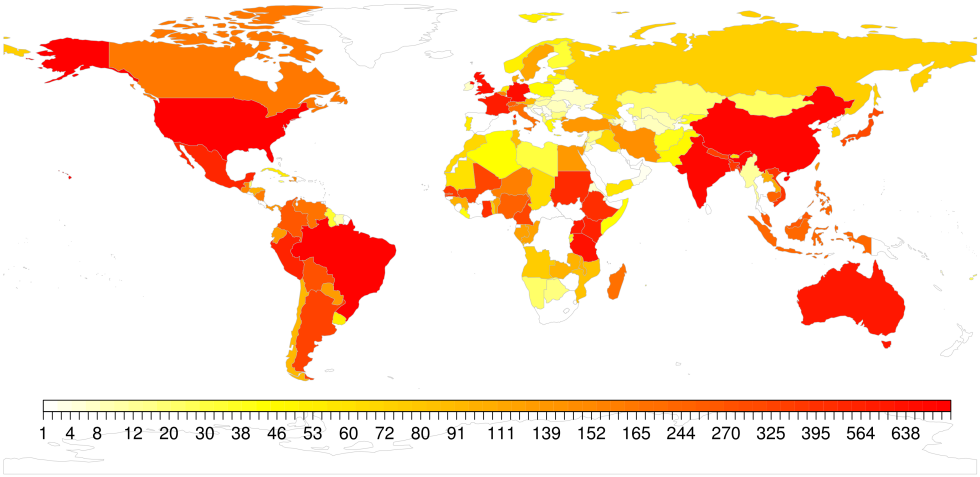


FIG. 3 – *Pays mentionés dans le corpus par nombre d'occurrences.*

Notons que plusieurs pays non-tropicaux sont également présents avec des fréquences élevées, dont les États-Unis, certains pays européens, le Japon, etc. Ces occurrences proviennent bien du corps des textes et non pas des métadonnées des articles. En effet, ces pays sont souvent mentionnés dans le corpus en lien avec des laboratoires, vaccins, établissements hospitaliers, etc.

Si en plus nous prenons en compte les années des publications, les données permettent de détecter et observer les tendances dans la recherche. Par exemple, nous pouvons représenter

sur un graphe le nombre d'occurrences des noms de pays par année, comme le montre la figure 4. En prenant comme exemple le continent d'Afrique, nous avons considéré les cinq pays ayant le plus d'occurrences. Le graphe à gauche présente le pourcentage du nombre d'occurrences de chaque pays par rapport à toutes les occurrences pour une année donnée. Le graphe à droite présente le pourcentage du nombre d'articles qui mentionnent chaque pays par rapport à tous les articles publiés dans la même année. La différence entre ces deux graphes provient du fait que le même article peut contenir de multiples occurrences d'un pays, qui seront alors comptées plusieurs fois pour le graphe à gauche et une seule fois pour le graphe à droite. Comme le corpus couvre la période d'août 2007 à août 2013, les données pour 2007 sont sur seulement 4 mois, et les données pour 2013 sont sur 8 mois.

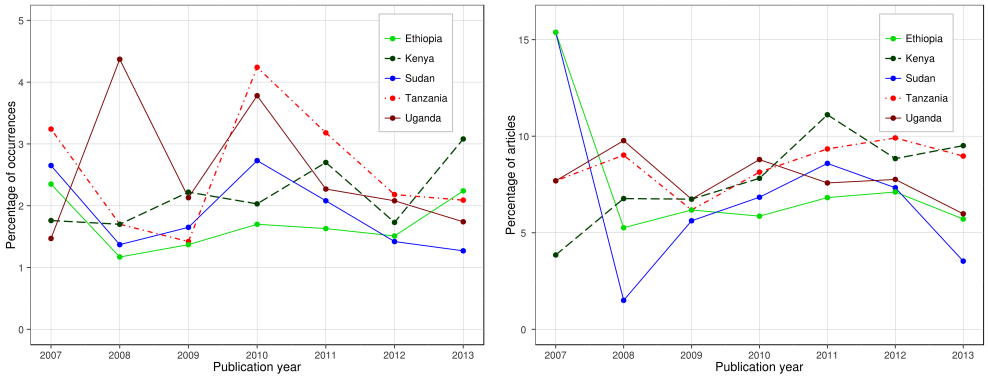


FIG. 4 – *Pays par nombre d'occurrences par année : les 5 pays en Afrique les plus cités*

Ces résultats montrent que les 5 pays le plus fréquemment mentionnés en Afrique se trouvent en forte proximité géographique : Kenya, Tanzanie, Ouganda, Soudan et Éthiopie. Les graphes indiquent un nombre relativement élevé d'occurrences de Tanzanie et Ouganda pour 2009 et 2011, ainsi qu'un intérêt émergeant pour le Kenya jusqu'à 2011. En 2012 et 2013, Kenya et Tanzanie sont cité dans presque 10% des articles, alors que le nombre d'occurrences des autres pays diminue. Globalement, nous pouvons observer une hausse significative du pourcentage d'articles liés à cette région entre 2009 et 2011 : en effet, en 2011 les articles mentionnant ces cinq pays constituent plus de 35% de tous les articles.

Nous avons également examiné les co-occurrences entre les pays les plus cités. La figure 5 montre une partie des corrélations obtenues. Nous pouvons observer les corrélations les plus importantes entre Ouganda et Kenya (0,28) et entre Soudan et Éthiopie (0,26). Cependant, ces valeurs ne sont pas élevées, ce qui signifie que peu de pays apparaissent ensemble dans les articles de façon systématique.

3.2 Analyse utilisant les géo-coordonnées

Pour toutes les localisations qui ont été extraites des textes, nous avons obtenu le nombre d'occurrences et les géo-coordonnées. La carte sur la figure 6 montre toutes les localisations géographiques. Les tailles des points sont relatives aux nombres d'occurrences. La grande

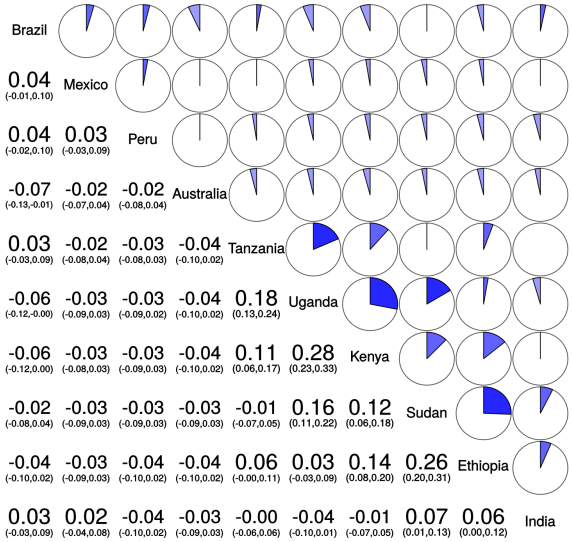


FIG. 5 – *Corrélations entre occurrences de pays*

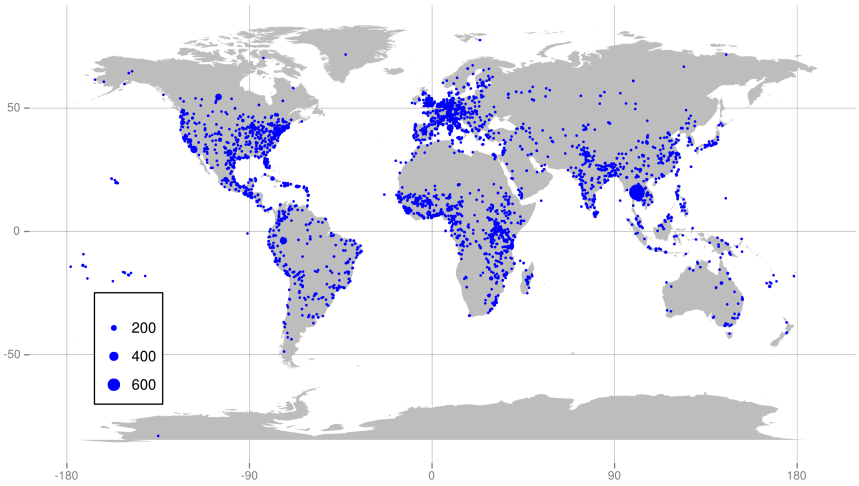


FIG. 6 – *Visualisation de localisations géographiques dans le corpus. Les tailles de points correspondent aux nombres d'occurrences.*

concentration de points dans certaines régions indique leur importance du point de vue de l'étude des maladies tropicales.

Ces données peuvent être exploitées également en prenant en compte les années de publication afin de visualiser l'évolution du nombre de recherches liées à chaque région. La figure 7 montre les localisations géographiques pour des différentes années de publication [9]. L'émergence de nouvelles régions d'intérêt peut être observée. Par exemple, un nombre croissant d'études entre 2008 et 2013 concernent la région autour de Bangkok, Thaïlande.

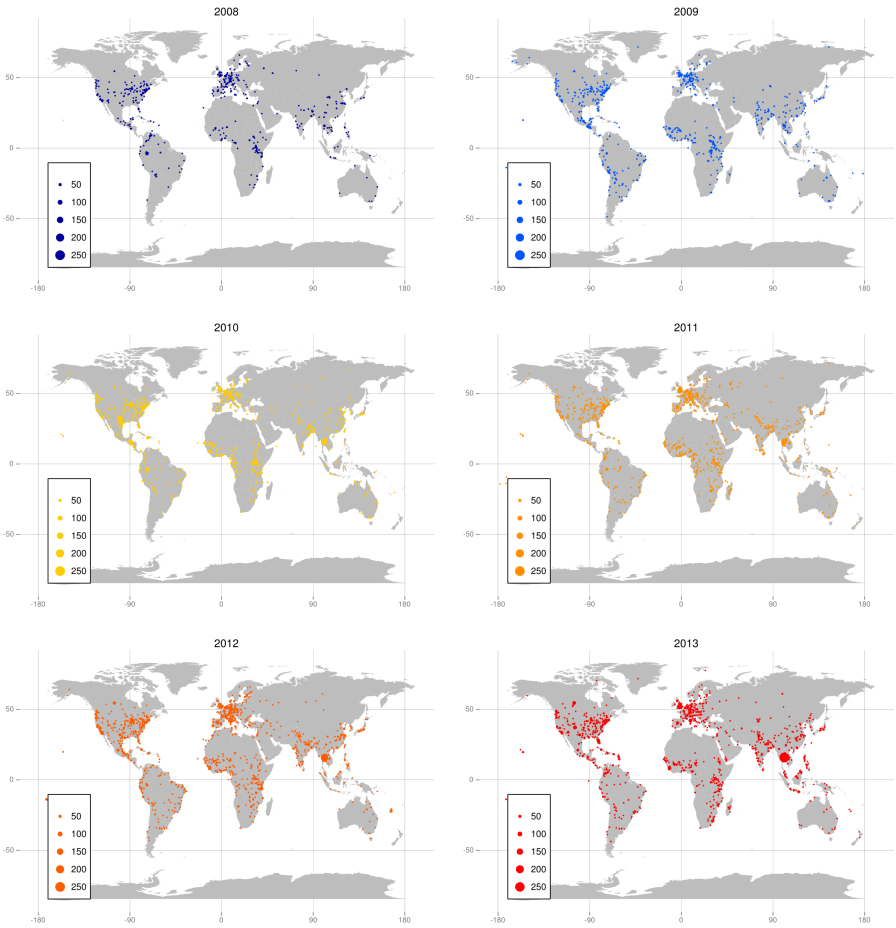


FIG. 7 – *Visualisation de localisations géographiques par année*

9. Une interface en ligne permettant l'exploitation interactive de ces données se trouve à l'adresse suivante : `http://linkedscience.org/demos/spatialaboutness/`

4 Discussion

Cette étude a pour but de montrer la pertinence de l'exploitation de données géo-spatiales dans le cadre de corpus scientifique spécifique à un domaine comme PLOS PNTDs. Les raisons de la présence de localisations géographiques dans des études scientifiques sont variées. Comme le montrent les exemples dans la table 3, les localisations sont liées à des épidémies, des vaccins, provenances de virus ou d'échantillons, adresses de compagnies médicales, etc.

TAB. 3 – *Exemples de phrases contenant des localisations*

Phrase	Type de localisation
RVFV has led to outbreaks in *Egypt* and the *Arabian Peninsula* with the potential to spread to the *United States* and *Europe*.	Epidémies
Briefly, One Step RT-PCR Kit (Qiagen ; *Valencia, CA*) was used for the RT-PCR reactions.	Adresse de compagnie médicale
A separate, internal control reaction for the detection of RNAse P was performed on the clinical samples from *Nicaragua* and *Sri Lanka*.	Provenance des échantillons
In a 2-site ID vaccine trial in *Thailand*, antibody levels varied 2.2 fold between different hospitals.	Campagne d'immunisation

Les limites de cette étude sont de nature linguistique. D'une part, les noms des localisations géographiques peuvent être polysémiques. La désambiguïsation des entités nommés peut être envisagée en faisant appel à des connaissances encyclopédiques (voir Bunescu et Pasca (2006)). Cependant, dans cette étude nous n'avons pas effectué de désambiguïsation et nous nous sommes appuyés sur les premiers résultats de Google GeoCode API. Il est possible que cette limitation ait introduit des erreurs dans les visualisations. D'autre part, l'extraction des localisations par l'outil CoreNLP ne permet pas de prendre en compte certaines expressions linguistiques qui contribuent à préciser un lieu. Or, il pourrait être utile d'exploiter la distinction entre, par exemple "*southern Tanzania*" et "*rural districts in northwest Tanzania*". La prise en compte de telles variations permettra d'obtenir une plus grande précision dans les représentations.

Il en résulte que la catégorisation des localisations pourra fournir des données pour des analyses plus fines avec des applications en recherche d'information et en veille. Par exemple, cela permettra de distinguer entre les localisations qui sont des adresses de laboratoires ou de compagnies et celles qui sont des foyers de contagion.

5 Conclusion

Notre objectif était de rendre compte de la dimension spatiale des études et de fournir les données nécessaires à une agrégation visuelle des résultats. En s'appuyant sur l'outil de Reconnaissance d'Entitées Nommées de Stanford, nous proposons une approche afin d'extraire les localisations géographiques liées aux études de la revue PLOS Negleted Tropical Diseases. Les résultats montrent l'évolution dans l'espace et le temps des zones porteuses de maladies tropicales comme le montrent par exemple les visualisations obtenues pour la Thaïlande.

Notre futur travail s'articulera autour de la catégorisation des localisations extraites à partir des textes. Un autre objectif est la construction d'un service d'agrégation et de partage de données géographiques liées à des études scientifiques, accessibles via SPARQL [10] ou GeoSPARQL [11] sous forme de Linked Data. Cela permettra la ré-utilisation de ces données par des applications et services externes.

Le fait d'établir des liens entre des études scientifiques et des informations géo-spatiales permet de fournir de nouveaux descripteurs des publications pour l'enrichissement de métadonnées. Cette approche permet une nouvelle lecture des articles scientifiques à travers les données géographiques. Il s'agit d'une première étude montrant des applications autour des représentations spatiales liées aux études scientifiques, notamment par des visualisations géographiques. Les résultats soulignent la nécessité d'une étude linguistique des contextes d'entités nommées dans les textes afin de catégoriser les informations géographiques. Cela nous permettra, à terme, de proposer une ontologie montrant les relations entre laboratoires et foyers de contagion.

Remerciement

Nous remercions Benoit Macaluso de l'Observatoire des Sciences et des Technologies (OST) [12], Montréal, Canada, pour le moissonnage du corpus PLOS.

Références

Arsevska, E., M. Roche, R. Lancelot, P. Hendrikx, et B. Dufour (2014). Exploiting textual source information for epidemiosurveillance. *Metadata and Semantics Research*, 359.

Borges, K. A., A. H. Laender, C. B. Medeiros, et C. A. Davis Jr (2007). Discovering geographic locations in web pages using urban addresses. In *Proceedings of the 4th ACM workshop on Geographical information retrieval*, pp. 31–36. ACM.

Bucher, B., P. Clough, H. Joho, R. Purves, et A. K. Syed (2005). Geographic ir systems : requirements and evaluation. In *Proceedings of the 22nd International Cartographic Conference*, Volume 201, pp. 11–16.

Bunescu, R. C. et M. Pasca (2006). Using encyclopedic knowledge for named entity disambiguation. In *EACL*, Volume 6, pp. 9–16.

Cucerzan, S. et D. Yarowsky (2002). Language independent ner using a unified model of internal and contextual evidence. In *proceedings of the 6th conference on Natural language learning-Volume 20*, pp. 1–4. Association for Computational Linguistics.

Inoue, Y., R. Lee, H. Takakura, et Y. Kambayashi (2002). Web locality based ranking utilizing location names and link structure. In *Web Information Systems Engineering Workshops, International Conference on*, pp. 56–56. IEEE Computer Society.

Jones, C. B., R. Purves, A. Ruas, M. Sanderson, M. Sester, M. Van Kreveld, et R. Weibel (2002). Spatial information retrieval and geographical ontologies an overview of the spirit

10. http://www.w3.org/TR/rdf-sparql-query/
11. http://www.opengeospatial.org/standards/geosparql
12. http://www.ost.uqam.ca/

project. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 387–388. ACM.

Kauppinen, T., A. Baglatzi, et C. Keßler (2013). Linked Science : Interconnecting Scientific Assets. In T. Critchlow et K. Kleese-Van Dam (Eds.), *Data Intensive Science*. USA : CRC Press.

Manning, C. D., M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, et D. McClosky (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, pp. 55–60.

Markowetz, A., Y.-Y. Chen, T. Suel, X. Long, et B. Seeger (2005). Design and implementation of a geographic search engine. In *WebDB*, Volume 2005, pp. 19–24.

Purves, R. S., P. Clough, C. B. Jones, A. Arampatzis, B. Bucher, D. Finch, G. Fu, H. Joho, A. K. Syed, S. Vaid, et al. (2007). The design and implementation of spirit : a spatially aware search engine for information retrieval on the internet. *International Journal of Geographical Information Science 21*(7), 717–745.

Tahrat, S., E. Kergosien, S. Bringay, M. Roche, et M. Teisseire (2013). Text2geo : from textual data to geospatial information. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, pp. 23. ACM.

Wang, C., X. Xie, L. Wang, Y. Lu, et W.-Y. Ma (2005b). Detecting geographic locations from web resources. In *Proceedings of the 2005 workshop on Geographic information retrieval*, pp. 17–24. ACM.

Wang, L., C. Wang, X. Xie, J. Forman, Y. Lu, W.-Y. Ma, et Y. Li (2005a). Detecting dominant locations from search queries. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 424–431. ACM.

Zhou, G. et J. Su (2002). Named entity recognition using an hmm-based chunk tagger. In *proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 473–480. Association for Computational Linguistics.

Summary

In this paper we present an approach for the extraction of geographic information from scientific articles in the biomedical domain. The idea is make sense of geographic dimension of articles via geo-spatial visualizations and enhanced information retrieval. We evaluate the approach by using the journal PLOS Neglected Tropical Diseases as a data source. We perform full-text analysis of the articles and produce Linked Data to describe the geographic aboutness of scientific studies. We make use of visualizations to present the results and discuss their implications for the future work.

Sequential pattern mining for customer relationship analysis

Kiril Gashteovski[**,***],Thomas Guyet[*]
René Quiniou[**], Alzennyr Gomes Da Silva[***], Véronique Masson[****]

[*]AGROCAMUS-OUEST/IRISA – UMR6074, thomas.guyet@irisa.fr
[**]INRIA – Centre de Rennes
[***]EDF R&D, ICAME, alzennyr.gomes-da-silva@edf.fr
[****]Université Rennes 1/IRISA – UMR6074

Abstract. Customer Relationship Management (CRM) is a set of tools for managing a company's interactions with its customers. Extracting the frequent interaction behaviors from the CRM database is useful to better understand the customers habits and to improve the company services. In this article, we propose a preliminary work on the study of an interactive tool. Our tool is based on the TGSP algorithm of Yen and Lee (2013) to extract frequent sequential patterns with time-gap information. The extracted patterns are visualized in a tree view and the analyst can interact with this view to support its analysis of the patterns.

1 Introduction

Customer Relationship Management (CRM) comprises a set of tools for managing the interactions between a company and its customers. The main objective of CRM is to develop long-term relationships with regular customers and to acquire new ones. Thus, CRM aims to understand the wishes and requirements of customers in order to let the company improve the quality of services provided to customers and propose new ones. A final objective for the company is to increase the amount of sales. With the development of loyalty cards, personal spaces on company websites, records from call centres, mailing, etc., companies are getting big databases of customer records describing the interactions between a customer and the company. By analysing customers' records, the data analysts from commercial departments expect to characterize interesting consumption behaviours and to discover clusters of similar customers.

The application of data mining tools in CRM has become an emerging trend in the global economy. According to Ngai et al. (2009), appropriate data mining tools, which are good at extracting and identifying useful information and knowledge from huge customer databases, are among the best supporting tools for making relevant CRM decisions. Amongst the data mining techniques, pattern mining techniques are suited to provide meaningful and comprehensive descriptions of frequent behaviours from the customer data records. These frequent behaviours can be used to predict the future needs of the customers. For instance, commercial services can advertise and recommend products to retain customers and maintain customer value. Such frequent behaviors can also be used to refill the stocks in advance once a promotion is scheduled. Customer data records contain timestamped data and so it is desirable to

take the temporal or at least the sequential dimension into account. Eichinger et al. (2006) uses sequence pattern mining combined with decision tree analysis to make behaviour predictions in telecommunications.

The main objective of the data analysts is to propose the right service to a customer at the right moment. If rules or sequential patterns can predict the interaction event that can follow a sequence of actions or events, they cannot predict at what precise time such an action have the highest probability to occur. Consequently, the commercial services can not plan their commercial actions precisely. Thus, sequential patterns should provide timing information on the relative occurrence of events or actions in the patterns. This is the objective of temporal data mining. In this context, our goal is to devise a method for extracting temporal patterns from large CRM databases that will inform the analyst about the time-delay between elements of the interactions between a company and its customers.

In this article, we present a tool to support the analysis of sequential data from a real CRM database. This tool relies on an adaptation of the TGSP algorithm (Yen and Lee, 2013) for extracting frequent patterns. The tool provides also pre and post-mining functionalities and, an interface to visualize and to browse the extracted patterns.

2 Pattern mining with temporal information

Itemset and association rule mining has been an important domain of data mining. A lot of research has been done on this problem, but a many opened questions still remain. Agrawal and Srikant (1994) were the first to propose a simple formal setting of the problem and an efficient algorithm, which they call *Apriori*, for mining all the frequent itemsets in a transaction database.

Sequential pattern mining extends itemset mining by taking into account the sequentiality of items. Mining operates on a sequence database $\mathcal{D} = \{S_1, S_2, \ldots, S_N\}$, where each element is a **sequence** $S = \langle s_1, s_2, \ldots, s_m \rangle$ consisting of itemsets $s_i = \{\tau_1, \tau_2, \ldots, \tau_m\}$. The order of elements in a sequence relates to time (e.g. purchase date). A sequence $\langle a_1, a_2, \ldots, a_m \rangle$ **is contained** in another sequence $\langle b_1, b_2, \ldots, b_k \rangle$ if there exist integers $i_1 < i_2 < \ldots < i_m$ such that $a_1 \subseteq b_{i_1}, a_2 \subseteq b_{i_2}, \ldots a_m \subseteq b_{i_m}$. Typically, a sequence is associated with a user. We say that a user **supports** some sequence S_i if S_i is contained in the sequence of that user. The **support of a sequence** S_i is the ratio of the number of sequences that support S_i over the total number of sequences n in the database (or number of users). The **minimum support** is a user-defined threshold $\sigma \in [0, 1]$. For some sequence S_i, if $support(S_i) \geqslant \sigma$, we say that S_i is a **frequent sequence** or a **sequential pattern**. Mining sequential patterns consists in extracting all the frequent sequences in a database $\mathcal{D}$. Several methods have been proposed for sequential pattern mining. The most well-known algorithms are GSP (Srikant and Agrawal, 1996), PrefixSpan (Pei et al., 2001), SPADE (Zaki, 2001) and BIDE (Wang and Han, 2004).

Sequential pattern mining takes the time dimension into account as denoted by the order of itemsets in sequences. However, representing time this way is very restrictive. No qualitative or quantitative information tells more precisely how much time separates events in a sequence. A lot of research has been done lately on temporal data mining and has produced various algorithms answering different types of questions for the temporal dimension.

FACE (Dousson and Duong, 1999) is an algorithm for discovering alarm correlation in a complex dynamic system. An alarm is a pair (A, t), where A is an alarm type and t is

the time of occurrence of this alarm. $(A_1, 2)$ means that an alarm labelled A_1 has occurred at time 2. An alarm log, $\mathcal{L}$, is an ordered list of alarms (ordered in time). For example, $\mathcal{L} = (A, 1), (B, 5), (C, 5), (A, 8), (D, 10), (B, 10), (F, 14)$ describes an alarm log, where the alarm A occurred at time 1, next, at time 5, alarms B and C have occurred, and so on. FACE finds specific temporal patterns called chronicle models in alarms log. FACE is based on the GSP algorithm (Srikant and Agrawal, 1996). A chronicle model is a constraint graph, *i.e.* a directed weighted graph where the nodes are alarm types and the arcs contain two kinds of information: the first is the arc direction (*e.g.* $A \rightarrow B$ means *"A occurred, followed by B"*; the second is the arc weight represented by a time interval telling what is the expected time between A and B. For example, $A_1 \xrightarrow{[t_1,t_2]} A_2$ means *an alarm of type A occurred and after at least t_1 and at most t_2 time-units, it was followed by an alarm of type B.*

I-Apriori and I-PrefixSpan (Chen et al., 2003) are two algorithms for time-interval sequential pattern mining in sequence databases. Here, event pairs contain an event label and a timestamp. The mining algorithm finds patterns in the format $\langle e_1, i_1, e_2, i_2, ..., i_{n-1}, e_n \rangle$, where e_{ind} is an event and i_{ind} is an interval label which denotes the time delay between the occurrence of e_{ind} and the occurrence of e_{ind+1}. However, there is one tedious detail with these algorithms: the analyst has to provide the set of intervals I as an input.

TGSP (*Time-Gap Sequential Patterns*) (Yen and Lee, 2013) is an algorithm for mining non-redundant time-gap sequential patterns where time intervals are found automatically. This algorithm makes use of additional parameters, ε and δ, defined in the CLIQUE algorithm (see (Yen and Lee, 2013; Agrawal et al., 2005)).

All the temporal sequential pattern mining algorithms above consider only instantaneous or point-based events (having no duration). Event timestamps specify when they starts (and finishes). Some methods take into account interval-based event. Guyet and Quiniou (2011) define a temporal item as $\mathcal{A} = (A, [l, u])$ where A is an event ID, and l, u are timestamps, which represent respectively the beginning and the end of the event. A temporal sequence consists of sets of temporal items. The authors have proposed a PrefixSpan strategy (depth first search) to mine temporal patterns containing quantitative temporal constraints between events.

Concerning our project, TGSP was the one that answers our questions the most, and we will present it in more details in the section 3.2. Next, we propose an alternative implementation of TGSP.

3 Integrated tool for mining a CRM database

First, this section gives details about the process of mining CRM data. Next, we present time-gap sequential patterns and the TGSP algorithm. Finally, our implementation of TGSP is briefly described.

The CRM data mining process described in figure 1 is a loop where the main steps are:

1. select customers sharing some contextual properties from the CRM database. This step generates the set of customer pathways, *i.e.* sequences of timestamped events. The customers are selected on criteria concerning their contract and their personal information.

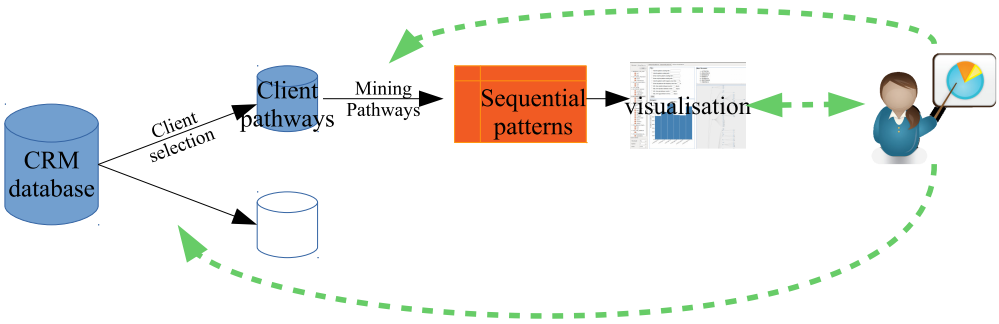


FIG. 1 – *CRM database analysis overview.*

2. extract the frequent behaviors from the set of pathways. We have implemented an algorithm based on TGSP to extract sequential patterns with quantitative time delays (gaps).
3. visualize the frequent patterns. The analyst can visualize and browse the set of frequent patterns to identify interesting behaviors, including quantitative time-gaps.
4. modify the mining settings (dashed-green arrows) to undertake a new mining process. This includes modifying the customer pathways database or selecting new mining parameters.

In this section, we present the two first steps of the exploration process. The next section will be dedicated to the interactive analysis of extracted pattern.

3.1 CRM database and customer pathway

In the CRM context, an event type refers to a certain kind of interaction between a customer and the company. An event refers to the occurrence of an event type in a customer pathway. Let I be the set of all possible event types. The number of possible event types is $n = |I|$.

Definition 1 (Event)**.** *The pair $E = (e, t)$ denotes an event E where $e \in I$ is an event type and $t \in \mathcal{T} \in \mathbb{N}$ is a timestamp denoting the time of occurrence.*

Definition 2 (Time-sequence)**.** *A time-sequence is an ordered set of events $s = \langle E_1, E_2, \ldots, E_k \rangle$ such that $\forall E_i, E_i = (e_i, t_i), e_i \in I, i = \{1, 2, \ldots k\}$, $t_i \in \mathbb{N}$, $t_i \leq t_{i+1}$. A time-sequence containing k events is called a **k-time-sequence**.*

Without limiting the generality of the definition above, we assume that the timestamp 0 is associated with the first item of every sequence of the database.

*The sequence $seq = \langle e_1, e_2, \ldots, e_k \rangle$, excluding event timestamps but maintaining the order of events, is called the **sequence** of the time-sequence s.*

A CRM database is a set of tuples $\mathcal{D} = \langle C, D, S \rangle$, where C is a customer identifier, $D = \langle d_1, \ldots, d_p \rangle$ is a set of values in $\mathbb{N}$ or $\mathbb{R}$ for p attributes, and S is a time-sequence. D gives the values of contextual attributes, *e.g.* personal customer information or contract details. The time-sequence S of customer C describes the pathway of customer C.

Definition 3 (Sequence database)**.** *Let $\mathcal{C}$ be a CRM database. A customer query $\mathcal{Q}$ describes the acceptable attribute values that fix the context of a mining process. Executing such a query yields the set of customers from $\mathcal{D}$ that match the* ***customer query****. The set of sequences associated with the selected customers is called the sequence database. We denote by N the number of customers satisfying the query.*

A customer query selects the customers from whom the analyst would like to identify interesting behaviours. For instance, the analyst may be only interested by `CSP+` customers. As a consequence, the sequence database to be mined will only contain sequences of such customers. The sequence database construction is a simple preprocessing step in our tool.

In the following, we assume that $\mathcal{D}$ is the sequence database that the analyst wants to mine in order to extract frequent temporal patterns.

3.2 TGSP Algorithm

The TGSP algorithm extracts sequential patterns comprising information about the time gap between any two adjacent items (see Yen and Lee (2013)). First, we present time-gap sequential patterns and next we detail TGSP.

3.2.1 Time-gap sequential patterns

Definition 4 (Time-gap sequence)**.** *A sequence $S = \langle e_1, (\tau_1), e_2, (\tau_2), \ldots, (\tau_{k-1}), e_k\rangle$, where $\forall e_i, e_i \in I$ and $\tau_i = [\tau_i^l, \tau_i^u]$, $\tau_i^l \leq \tau_i^u$, is called a* ***time-gap sequence*** *and τ_i is called a* ***time-gap interval****. τ_i indicates that there is at least τ_i^l and at most τ_i^u time-units between e_i and e_{i+1}.*

Let $s = \langle E_1, E_2, \ldots E_k\rangle$ be a time-sequence where $\forall i, E_i = \langle e_i, t_i\rangle$. The time-gap sequence ***associated*** *with s is $S = \langle e_1, [t_2 - t_1, t_2 - t_1], e_2, [t_3 - t_2, t_3 - t_2], \ldots, [t_k - t_{k-1}, t_k - t_{k-1}], e_k\rangle$. For all $i < k$, we have $\tau_i = [t_{i+1} - t_i, t_{i+1} - t_i]$.*

Any time-sequence can be translated into a time-gap sequence. Thus, the time sequence database $\mathcal{D}$ is also a time-gap sequence database.

Definition 5 (Sequence inclusion – **contains** relation)**.** *Let $p = \langle a_1, (\tau_1), a_2(\tau_2), \ldots, (\tau_{n-1}), a_n\rangle$ and $q = \langle b_1, (\omega_1), b_2, (\omega_2), \ldots, (\omega_{m-1}), b_m\rangle$ be two time-gap sequences. q* ***contains*** *p iff there exists $1 \leq k_1 < k_2 < \ldots < k_n \leq m$, such that $a_1 = b_{k_1}, a_2 = b_{k_2}, \ldots a_n = b_{k_n}$ and $\tau_1 \subseteq [\sum_{j=k_1}^{k_2} \omega_j^l, \sum_{j=k_1}^{k_2} \omega_j^u], \tau_2 \subseteq [\sum_{j=k_2}^{k_3} \omega_j^l, \sum_{j=k_2}^{k_3} \omega_j^u], \ldots, \tau_{n-1} \subseteq [\sum_{j=k_{m-1}}^{k_m} \omega_j^l, \sum_{j=k_{m-1}}^{k_m} \omega_j^u]$ (interval inclusions).*

A time-sequence p ***contains*** *a time-sequence q iff the time-gap-sequence associated with p* ***contains*** *the one associated with q.*

Example 1 (Sequence inclusion)**.** *In this definition, $[\sum_{j=k_{m-1}}^{k_m} \omega_j^l, \sum_{j=k_{m-1}}^{k_m} \omega_j^u]$ is an interval describing the time-gap between the two events $b_{k_{m-1}}$ and b_{k_m}. In fact, ω_i^l (resp. ω_i^u) gives the minimum (resp. maximum) delay between b_i and b_{i+1}, so the minimum (resp. maximum) delay between b_k and b_l is $\sum_{j=k}^{l} \omega_j^l$ (resp. $\sum_{j=k}^{u} \omega_j^u$).*

For instance, let $q = \langle a[1, 2]b[3, 4]c\rangle$, there is a delay of 1 to 2 time units between a and b, but there is a delay of 4 (= 1 + 3) to 6 (= 2 + 4) time units between a and c. Then, q contains the sequence $p = \langle a[4, 5]c\rangle$.

The time-gap sequence c supports the time-gap sequence s iff the time-sequence C related to c contains a time-subsequence S such that the related time-gap sequence of S is equal to s. The **support** of a time-gap sequence s, noted $supp(s)$, is the number of time-gap sequences that support s in the database.

Definition 6 (Time-gap sequential pattern)**.** *Let σ be a user-specified **minimum support**. A sequential pattern p is frequent in the time-gap sequence $\mathcal{D}$ iff $supp(p) \geq \sigma$.*

3.2.2 TGSP algorithm

The TGSP algorithm adopts a breadth first strategy similar to the Apriori algorithm. The algorithm intertwines the evaluation of the support of the $k-1$ candidate sequences with the generation of new candidates sequences of size k. For time-gap sequential pattern, an additional step generates and evaluates the time-gaps inserted in the time-gap patterns. TGSP makes use of the CLIQUE clustering algorithm to generate time-gap sequential pattern candidates from k-time-gap-table. Before explaining the algorithm, we define the k-time-gap-table.

Definition 7 (Minimal sub-time-sequence)**.** *Let $p = \langle E_1, E_2, \ldots E_k \rangle$ be a time-sequence, and $\langle x_i, x_j \rangle, 1 \leq i \leq j \leq k$ be a 2-sequence. If there exists $E_i = (x_i, t_i)$ and $E_j = (x_j, t_j), 1 \leq i \leq j \leq k$ then $q = \langle E_i, E_j \rangle$ is a sub-time-sequence of p ($q \subseteq p$). If $\forall m \in [i, j], \nexists x_m$ such that $x_i = x_m$ or $x_j = x_m$, then q is a minimal sub-time-sequence of p.*

Definition 8 (k-time-table)**.** *A k-time-table is associated with each frequent k-sequence. It contains a set of pairs $(user_{id}, \langle t_1, t_2, \ldots, t_k \rangle)$, where $user_{id}$ is the ID of a user which supports the relevant k-sequence, and $\langle t_1, t_2, \ldots, t_k \rangle$ is a list of time-points (a time-point per time-sequence) related to the minimal sub-time sequences of the k-sequence.*

Definition 9 (k-time-gap-table)**.** *A k-time-gap-table associated with some k-time-gap-sequence contains a set of pairs $(user_{id}, \langle \tau_1, \tau_2, \ldots, \tau_{k-1} \rangle)$, $\forall \tau_i, \tau_i = t_{i+1} - t_i$ where $user_{id}$ is the user ID which supports the k-sequence, and the second element is a list of corresponding time-gaps between every adjacent items in the sequence. We will also refer to a record of a $k-$time-gap-table as to a **data point** in the $(k-1)-$dimensional space.*

Firstly, the algorithm scans the database in order to find the frequent 1-sequences, *i.e.* the frequent items.

Assuming that the frequent $(k-1)$-time-gap sequential patterns have been computed (as well as their $(k-1)$-sequences, without time-gaps), the next step is the k-sequence candidate generation. Let $p = \langle a_1, a_2, \ldots, a_{k-1} \rangle$ and $q = \langle b_1, b_2, \ldots, b_{k-1} \rangle$, where $a_2 = b_1, a_3 = b_2, \ldots, a_{k-1} = b_{k-2}$, be two frequent $k-1$-sequences. Then, the sequence $c = \langle a_1, a_2, \ldots, a_{k-1}, b_{k-1} \rangle$ is a k-candidate sequence. If the intersection of the sets associated with $user_{id}$ for p and q in the corresponding time-tables has a size smaller than the minimum support count then candidate c is deleted. Otherwise, the two $(k-1)$-time tables are joined into a k-time-table for candidate c.

For each frequent k-sequence, the algorithm generates a k-time-gap-table. Each record of a k-time-gap-table can be considered as a data point in a $(k-1)$-dimensional space (a k sequence contains $k-1$ intervals). For each frequent sequence separately, the CLIQUE clustering algorithm divides the space in "rectangular units" according to the user-defined parameter $\varepsilon \in \mathbb{N}$ (the **length parameter**). Each unit is represented as $U = \langle u_1, u_2, \ldots u_{k-1} \rangle$ where each

value $u_i = [l_i, u_i]$, $1 \leq i \leq k-1$, is an interval in the i-th dimension. Suppose that $k = 2$ and $\varepsilon = 2$. Then the units in the space will be: $u_{1_0} = [0, 0], u_{1_1} = [1, 2], u_{1_2} = [3, 4], u_{1_3} = [5, 6], \ldots, u_{1_n} = [l_{1_n}, h_{1_n}]$, where $h_{1_n} = \varepsilon \times n$ and $l_{1_n} = h_{1_n} - \varepsilon + 1$. For each k-time-gap table, a data point can be represented as $V = \langle v_1, v_2, \ldots, v_{k-1} \rangle$. For a data-point V and a unit $U = \langle u_1, u_2, \ldots u_{k-1} \rangle$, if $\forall i,\ l_i \leq v_i \leq h_i$ where $u_i = [l_i, u_i]$, then the unit U **contains** the data point V. We will denote this simply as $V \in U$. The total number of data points per k-time-gap table is Δ. The **density of a unit** U is defined by:

$$density(U) = \frac{|\{V | V \in U\}|}{\Delta}$$

A unit U is called a **dense unit** if $density(U) \geq \delta$, where $\delta \in [0, 1]$ is a user-defined parameter. For each k-time-gap table, the CLIQUE clustering algorithm finds the dense units, and will eventually return the minimal description of the time-intervals between the adjacent items in the frequent sequence. The CLIQUE clustering generates the k-time-gap sequential patterns.

After generating the k-time-gap sequential patterns, the algorithm performs the pruning step. This step prunes the records in the k-time tables which are not contained in any k-time-gap sequential pattern, because no $(k+1)$-time-gap sequential patterns can be generated from them (the anti-monotonicity property). After this, the algorithm checks if the number of records in each k-time table is lower than σ. If this is the case, then this k-time table is deleted.

3.3 TGSP modification

Running the CLIQUE algorithm at every candidate generation step can be very costly, especially for small values of ε and δ. Our objective was to evaluate the adequacy of the TGSP algorithm to CRM mining and, so, we have adopted a simplification of TGSP. The efficiency of the algorithm will be improved in future work.

Our implementation of the TGSP algorithm avoids intertwining k-sequence construction and k-time-gap sequence generation. It first extracts all the frequent sequential patterns (without time-gaps) using a classical PrefixSpan algorithm and then, for each extracted sequential pattern, it applies the CLIQUE algorithm to extract the dense units that generate the time-gap sequential patterns.

It should be noted that our algorithm can generate frequent sequences from which no frequent time-gap-sequential pattern can be extracted. It appends while a dense unit yields only unfrequent time-gap sequences. In such cases, the CLIQUE algorithm would be executed more times than required.

4 Browsing sequential patterns

Once the TGSP algorithm have been run, the analyst will get lots of patterns. To mine accurate behaviours, a low frequency threshold may be required. As a consequence, the analyst will have to identify the most interesting patterns amongst the large amount of extracted patterns. Our tool has two functionalities to support this task: firstly, we propose a tree view of the pattern set with which the user can interact to explore the individual patterns (especially,

the time-gaps distributions); secondly, we propose filtering options in order to post-prune the patterns interactively.

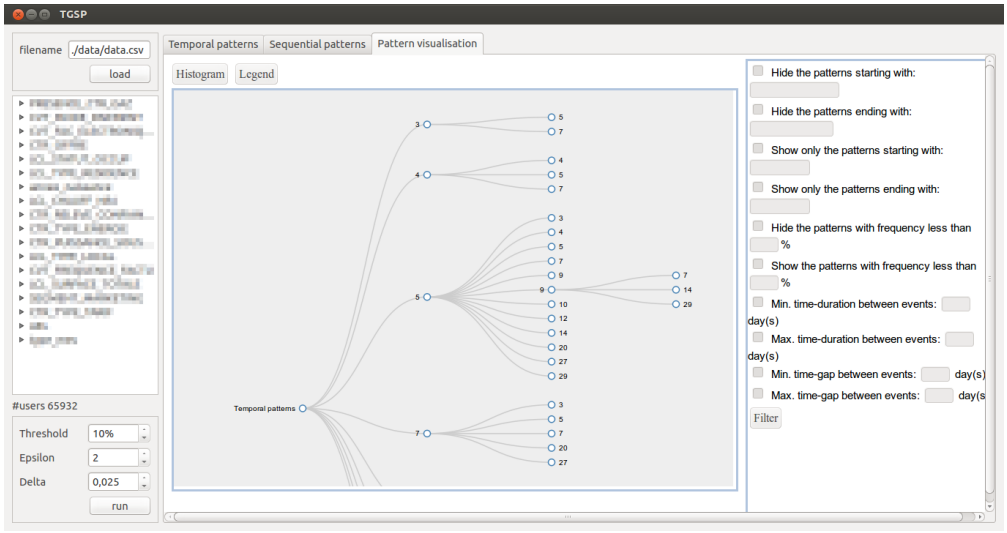


FIG. 2 – *Customer data mining interface*

Figure 2 illustrates the interface of the tool for discovery and visualization of time-gap sequential patterns. On the left, there is a panel with several check-box options. These represent the customer properties, *i.e.* the data of a customer provided when he signed a contract with the company. Relying on these options, the analyst will be able to specify a query for pre-processing the CRM database (see section 3.1). For instance, the analyst may want to perform the analysis only on the *customers that use gas and born after 1980*. Once the selection is made on the left panel, the mining process is performed on those customers only, not on all the customers. Each time the query is modified, the sequence database is updated accordingly and a new mining process is executed.

Right below the pre-processing options, three dialog boxes can be used to set the frequency threshold σ, and the CLIQUE parameters ε and δ. By default they are set to 10%, 2 and 0.075 respectively.

The central panel is dedicated to the visualization of patterns. The two other tabs of the central pattern contain textual outputs. The right panel contains some filtering options (see section 4.2). Finally, "popping windows" can display the event dictionary and dynamic time-gap histograms.

4.1 Pattern visualization

Wong et al. (2000) give several ways to visualize of sequential patterns. The most intuitive and most expressive one appears to be the visualization of sequential patterns as trees. Thus, we propose an interactive visualization of the patterns set as a simple tree view. The analyst can interact with the tree by expanding or by collapsing the nodes. On the other hand, he can

get some contextual information by dragging the mouse over the nodes (see figure 3, on the right).

This view does not highlight the time-gaps that characterize the patterns. We have chosen to visualize the time-gap distribution in a separate window (see figure 3, on the left). By clicking on a some leaf of the tree, the analyst can display an histogram representing the distribution of every time-gap for the related pattern.

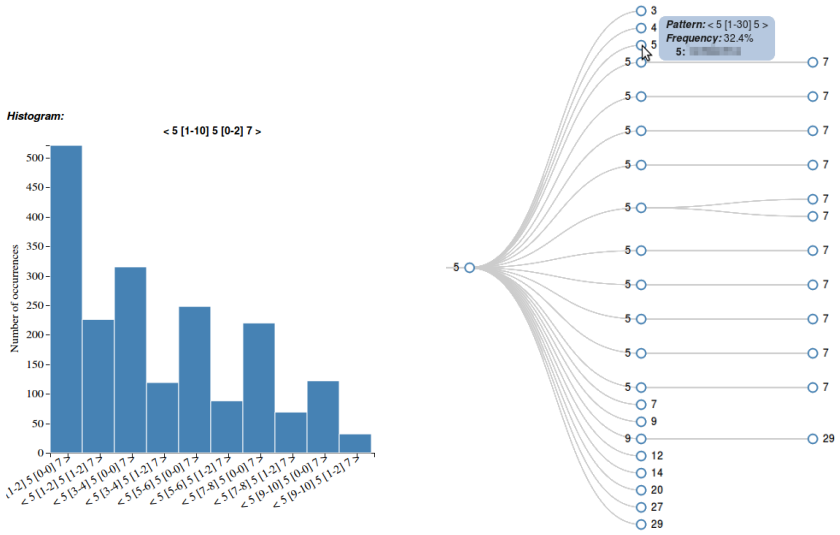


FIG. 3 – *Pattern interactive visualization: on the left, the histogram of the time-gaps a pattern, on the right: the tree view of the pattern set with interactive feedbacks.*

4.2 Filtering options

To reduce the size of the tree and to make easier the exploration of patterns, we propose some filtering options. With tree filters, the tool can dynamically hide the nodes that do not match the filters. The filtering options are in Conjunctive Normal Form (CNF) where each clause of the CNF is a different condition regarding the time-gap sequential patterns.

The tree illustrated in the figure 2 shows a part of the TGSP patterns extracted from all the customers, with minimum support of 1%, $\varepsilon = 2$ and $\delta = 0.075$ (270 patterns). We can see how useful the filtering options can be. By narrowing down the patterns that the analyst is looking for, the number of displayed patterns can be greatly reduced. This makes the discovery of potentially interesting patterns much more precise, more elegant and faster. There might even be cases where thousands of patterns would be displayed, which would make searching interesting patterns very difficult (almost impossible) without the filtering options.

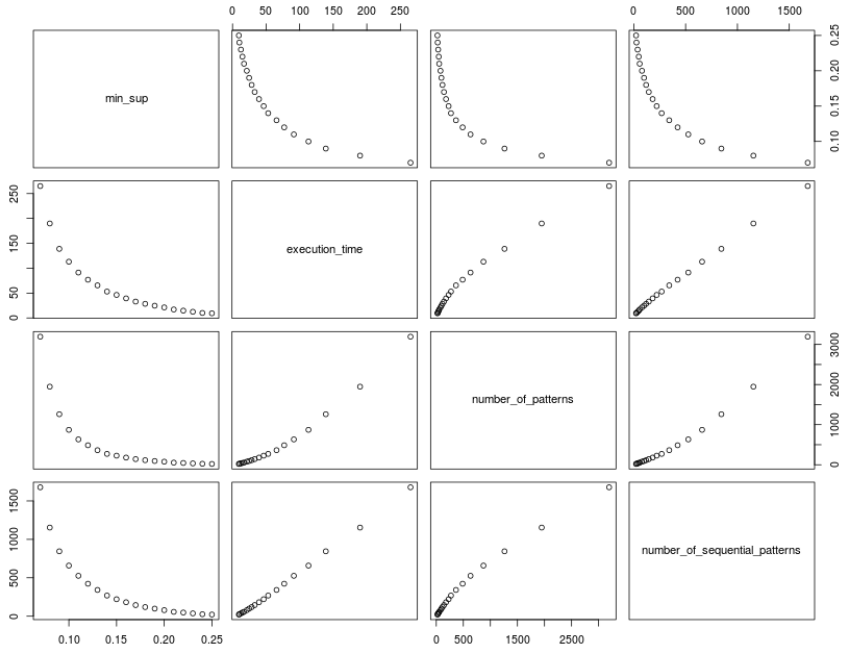


FIG. 4 – *Experiment on synthetically generated data by the IBM quest data generator software*

5 Implementation and experiments

The TGSP algorithm have been implemented with C++/Python. The interactive tool have been implemented with Python/Qt4 and javascript technologies (*D3JS* library).

5.1 Algorithm evaluation on synthetic datasets

We conducted several experiments on synthetic data generated by the IBM quest data generator software. Figure 4 illustrates the results of a synthetic dataset generated with respect to simple properties: the number of users is set to 10K, the number of different items is 100, and the other properties are left to be the default values proposed by the data generator.

The number of sequential patterns has a strong linear correlation with the number of temporal patterns (98.8% according to the Pearson product-moment correlation coefficient). The decrease of the minimum support leads to an exponential growth on the execution time because the slowest part of the TGSP algorithm is the extraction of the k-time-gap tables. If m sequential patterns are found by PrefixSpan, then the algorithm goes m times through the whole database.

5.2 Interactive tool experiments

Our tool has been developed to analyse a CRM dataset provided by our partner company. The prototype has been evaluated on a small real dataset containing around 65000 customers with 180 different events. In this dataset the event distribution is rather odd (some events occur frequently in the customer pathway, including repetitions). As a consequence, a low threshold leads to large number of patterns. The use of TGSP could reduce this amount of patterns.

On all the customers' data, with $\sigma = 1\%$, $\epsilon = 2$ and $\delta = 0.075$, the number of patterns is 270 and the execution time is around 4.5 minutes.

The final interactive tool have been tested by the commercial service of the company. Their feedback was positive. The computation time has been estimated reasonable, but this dataset is only a sample of the real CRM database and the algorithms will have to be improved in the future.

6 Conclusion

We proposed to use the TGSP algorithm to mine Customer Relationship Management (CRM) Database. This algorithm extracts frequent sequences with quantitative time-gaps that inform the analyst about the typical duration between the occurrences of two events of the sequence.

We also tackle the visualization of these time-gap sequential patterns. Besides the pattern mining algorithms, the human eye can be very efficient pattern mining tool, so a good intuitive visual representation is useful. We complete the visualization with interactions to support the analyst in its exploration of the frequent patterns.

From the conducted experiments, we can see that the hardest tasks in the algorithm is computing the k-time-gap-tables. This part of the algorithm scans the whole database m times (m being the number of sequential patterns found by PrefixSpan). A more efficient approach would scan the database less times, maybe by pruning some sequences in the time-gap-tables counting process.

References

Agrawal, R., J. Gehrke, D. Gunopulos, and P. Raghavan (2005). Automatic subspace clustering of high dimensional data. *Data Mining and Knowledge Discovery 11*(1), 5–33.

Agrawal, R. and R. Srikant (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB)*, pp. 487–499.

Chen, Y.-L., M.-C. Chiang, and M.-T. Ko (2003). Discovering time-interval sequential patterns in sequence databases. *Expert Systems with Applications 25*, 343–354.

Dousson, C. and T. V. Duong (1999). Discovering chronicles with numerical time constraints from alarm logs for monitoring dynamic systems. In *Proceedings of the 16th International Joint Conference on Artifical Intelligence (IJCAI)*, pp. 620–626.

Eichinger, F., D. D. Nauck, and F. Klawonn (2006). Sequence mining for customer behaviour predictions in telecommunications. In *Workshop on Practical Data Mining: Applications, Experiences and Challenges*, pp. 3–10.

Guyet, T. and R. Quiniou (2011). Extracting temporal patterns from interval-based sequences. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1306–1311.

Ngai, E., L. Xiu, and D. Chau (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications 36*(2, Part 2), 2592 – 2602.

Pei, J., J. Han, M.-A. B., and P. H. (2001). PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *Proceedings of the 17th International Conference on Data Engineering (ICDE)*, pp. 215–224.

Srikant, R. and R. Agrawal (1996). Mining sequential patterns: Generalizations and performance improvements. In *Proceedings of the 5th International Conference on Extending Database Technology: Advances in Database Technology (EDBT)*, pp. 3–17.

Wang, J. and J. Han (2004). BIDE: Efficient mining of frequent closed sequences. In *Proceedings of the 20th International Conference on Data Engineering (ICDE)*, pp. 79–90.

Wong, P. C., W. Cowley, H. Foote, E. Jurrus, and J. Thomas (2000). Visualizing sequential patterns for text mining. In *Proceedings of the Symposium on Information Vizualization (INFOVIS)*, pp. 105–111.

Yen, S.-J. and Y.-S. Lee (2013). Mining non-redundant time-gap sequential patterns. *Applied Intelligence 39*(4), 727–738.

Zaki, M. J. (2001). SPADE: An efficient algorithm for mining frequent sequences. *Machine Learning 42*(1-2), 31–60.

Résumé

La gestion de la relation client (CRM) est un ensemble d'outils pour la gestion des interactions d'une entreprise avec ses clients. Extraire les séquences fréquentes d'interaction à partir d'une base de données CRM est utile pour mieux comprendre les habitudes des clients et pour améliorer le service qui leur est rendu. Dans cet article, nous présentons un travail préliminaire sur le développement d'un outil interactif d'extraction de séquences fréquentes avec durées inter-événements. Nous proposons un outil qui inclut l'adaptation de l'algorithme TGSP de Yen and Lee (2013) pour extraire des motifs séquentiels fréquents avec des durées inter-événements. Les motifs extraits sont visualisés par un arbre représentant l'ensemble des séquences fréquentes. L'analyste peut interagir avec cette vue des motifs pour en faciliter l'analyse.

Exploration de données temporelles avec des treillis relationnels

Cristina Nica*, Xavier Dolques*, Agnès Braud*
Marianne Huchard**, Florence Le Ber*

*ICube, Université de Strasbourg, CNRS, ENGEES
prenom.nom@engees.unistra.fr, agnes.braud@unistra.fr
http://icube-bfo.unistra.fr
**LIRMM, Université de Montpellier 2, CNRS
huchard@lirmm.fr
https://www.lirmm.fr

Résumé. Cet article présente une méthode d'exploration de données temporelles à l'aide de treillis relationnels. Elle s'applique à un jeu de données composé de séquences de valeurs concernant des paramètres physico-chimiques et biologiques mesurés dans des cours d'eau. L'objectif est d'extraire des sous-séquences fréquentes reliant les deux types de paramètres. Nous montrons sur un petit exemple que l'analyse relationnelle de concepts (ARC) permet de mettre en évidence l'influence dans le temps des paramètres physico-chimiques sur les paramètres biologiques.

1 Introduction

La fouille de données séquentielles (Agrawal et Srikant, 1995) est un domaine de recherche actif qui comporte de nombreuses applications. Par exemple, en étudiant les données acquises auprès de clients d'un magasin, on cherchera à prévoir et favoriser leurs prochains achats ; en étudiant les suivis temporels d'un processus industriel, on cherchera à caractériser les séquences d'évènements qui mènent à un incident. Dans le cas qui nous préoccupe ici, il s'agit d'anticiper et de prévenir l'évolution de l'état des cours d'eau, en examinant les séquences de prélèvements effectués sur les rivières. Plus particulièrement nous essayons de mettre en évidence l'effet dans le temps d'un état physico-chimique de l'eau sur son état biologique (populations animales et végétales) en recherchant des répétitions fréquentes de sous-séquences ayant valeur de règles. Ce travail est mené dans le cadre du projet Fresqueau [1] en interaction avec des hydrobiologistes. Une première approche, s'appuyant sur la recherche de motifs temporels, a été mise en œuvre avec succès par Fabrègue et al. (2014).

Dans cet article, nous explorons une deuxième approche, en considérant ce problème temporel comme une variante d'un problème relationnel. Nous exploitons une technique d'analyse de données relationnelles fondée sur l'analyse de concepts formels (ACF) (Ganter et Wille,

1. http://engees-fresqueau.unistra.fr/

1997), l'analyse relationnelle de concepts (ARC), qui est développée depuis une dizaine d'années (Hacene et al., 2013). Cette méthode a été utilisée sur de nombreuses applications mettant en jeux des données relationnelles, en particulier en génie logiciel pour l'analyse d'éléments UML (Arévalo et al., 2006; Dolques et al., 2012) ou pour la détection de défauts (Moha et al., 2008). Cette approche est également utile pour factoriser des classes redondantes, en exploitant les attributs et les relations entre classes (Miralles et al., 2015).

Appliquer l'ARC à l'analyse de données séquentielles requiert une modélisation spécifique que nous allons expliciter dans ce papier à partir d'un exemple simplifié. En revanche nous ne présenterons pas de résultats chiffrés car l'exploitation complète des données pose des difficultés non encore résolues. Dans la suite de l'article nous présenterons ces données, puis les fondements théoriques de la modélisation utilisée ; nous montrerons ensuite sa mise en œuvre et le type de résultats qu'elle permet d'obtenir avant de conclure.

2 Contexte et données

De nombreuses questions de recherche touchant à l'environnement impliquent la prise en compte de données temporelles ou spatiales. Dans le cadre du projet Fresqueau, nous nous intéressons à l'état des cours d'eau, et nous développons des méthodes de fouille de données pour exploiter les données disponibles permettant d'évaluer cet état. Ces données sont de natures et d'origines différentes : elles concernent la qualité de l'eau, l'hydrologie, les stations de mesures, etc. mais également l'occupation du sol au voisinage des cours d'eau (Braud et al., 2014). Les données de qualité de l'eau en particulier sont produites par les agences de l'eau et l'ONEMA (Office National de l'Eau et des Milieux Aquatiques), et se déclinent en trois sous-ensembles :

1. données concernant l'état physico-chimique de l'eau et des sédiments ; macropolluants (nitrates, matières organiques, ...) et micropolluants (pesticides, ...) ;
2. données concernant l'état des peuplements biologiques floristiques et faunistiques : cet état est synthétisé dans des indices biologiques, parmi lesquels l'indice biologique global normalisé (IBGN) (AFNOR, 2004) est le plus fréquemment utilisé ;
3. données concernant l'état physique : il s'agit de l'hydromorphologie du cours d'eau (état des berges, du lit mineur, du lit majeur, ...) et des conditions hydrologiques (débits) et hydrauliques (vitesse, géométrie du cours d'eau).

Ces données sont issues de résultats d'analyse de prélèvements effectués régulièrement sur les réseaux de mesures nationaux. Chaque station d'un réseau de mesures est ainsi caractérisée théoriquement par une note annuelle d'un ou plusieurs indices biologiques, et par des valeurs bimensuelles de paramètres physico-chimiques. Le tableau 1 présente un extrait du jeu de données. À chaque station (colonne 1) sont associées plusieurs dates (sous la forme mois/année, colonne 2) auxquelles sont effectués des prélèvements physico-chimiques (par exemple, NH_4^+, NKJ) ou biologiques (synthétisés sous la forme d'un indice, l'IBGN ici). Ici une seule station est présentée pour des questions de place. Le nombre d'attributs est également limité. On remarque sur ce tableau que les données sont éparses : en particulier, alors que les paramètres physico-chimiques sont mesurés tous les deux à trois mois, l'IBGN est réalisé au mieux une fois l'an, en période d'étiage, donc généralement en été.

Numéro Station	Mois / année	NH_4^+	NKJ	NO_2^-	PO_4^{3-}	Phosphore total	IBGN
2	01/04	0,043	0,146	0,421	-	-	-
	04/04	-	-	-	0,325	0,093	-
	07/04	2,331	7,993	0,252	0,132	0,066	-
	08/04	-	1,414	-	-	-	-
	09/04	-	-	-	-	-	8
	11/04	0,117	0,0844	-	0,188	-	-
	12/04	-	-	-	0,067	0,078	-
	03/05	-	0,182	0,0310	0,137	-	-
	06/05	0,004	-	0,012	0,035	0,034	-
	08/05	-	-	-	-	-	10

TAB. 1 – *Extrait du jeu de données avec différents paramètres physico-chimiques (ammonium, nitrate de Kjeldahl, nitrite, orthophosphate, phosphore total) et un indice biologique (IBGN)*

Pour mettre en œuvre la méthode choisie, différents prétraitements sont nécessaires. En particulier, il faut transformer l'information numérique en information qualitative et pour cela nous nous appuyons sur des références du domaine permettant d'agréger les données collectées sur les stations de mesures. De plus, pour tenir compte des connaissances du domaine dans le processus d'analyse, nous ne considérerons par la suite qu'une partie des données. Concrètement, nous éliminons les dates de prélèvements physico-chimiques trop loin dans le temps (au-delà de 4 mois) et avant une date de prélèvement biologique, car nous cherchons à étudier l'effet, limité dans le temps, de l'état physico-chimique du cours d'eau sur la biologie.

Les données ont donc d'abord été discrétisées en utilisant la norme SEQ-Eau [2], modifiée à la marge selon l'avis des hydro-écologues travaillant dans le projet Fresqueau. Cette discrétisation transforme les données initiales en cinq classes de qualité, "Très bon", "Bon", "Moyen", "Mauvais" et "Très mauvais" representées par cinq couleurs *Bleu*, *Vert*, *Jaune*, *Orange* et *Rouge*. La norme SEQ-Eau permet également de regrouper les paramètres initiaux en 15 macro-paramètres : ainsi les paramètres PO_4^{3-} et phosphore total sont rassemblés en un seul paramètre qualitatif PHOS (matières phosphorées) qui prend la valeur la plus basse des deux. Les paramètres NH_4^+, NJK et NO_2^- sont eux rassemblés dans le macro-paramètre AZOT (matières azotées hors nitrate).

Le résultat de la discrétisation appliquée au tableau 1 est présenté dans le tableau 2. Une station (*Site 1*) a été ajoutée. Concernant la station *Site 2*, on remarque que le nombre de dates considérées a été réduit de moitié.

3 Analyse de concepts formels

L'analyse de concepts formels (ACF) est une méthode de classification qui s'applique à des jeux de données constitués d'objets décrits par des attributs (Ganter et Wille, 1997). D'un point de vue mathématique, l'ACF permet d'extraire des données un ensemble de concepts munis

2. http://sierm.eaurmc.fr/eaux-superficielles/fichiers-telechargeables/grilles-seq-eau-v2.pdf

Site	Date	AZOT	PHOS	IBGN
$Site\ 1$	06/07	Bleu	-	-
	07/07	Vert	Bleu	-
	09/07	-	-	Bleu
	02/08	Bleu	Vert	-
	04/08	Vert	-	-
	05/08	-	-	Jaune
$Site\ 2$	07/04	Orange	Jaune	-
	08/04	Vert	-	-
	09/04	-	-	Orange
	06/05	Bleu	Vert	-
	08/05	-	-	Jaune

TAB. 2 – *Jeu de données discrètes obtenu à partir du tableau 1*

d'une structure hiérarchique. Pour l'illustrer, nous utilisons l'exemple du tableau 2 contenant les données biologiques et physico-chimiques discrétisées.

On considère en entrée du processus ACF un contexte formel qui est un triplet $\mathcal{K} = (O, A, I)$, où O est un ensemble d'objets, A un ensemble d'attributs et I une relation binaire entre O et A, $I \subseteq O \times A$. Le tableau 2 nécessite donc un pré-traitement, qui consiste ici à transformer une relation multi-valuée en une relation binaire. Le contexte $\mathcal{K}_{\texttt{sites}}$ résultant est présenté au tableau 3, avec :

- $O = \{S1\_06/07, S1\_07/07, S1\_09/07, S1\_02/08, S1\_04/08, S1\_05/08, S2\_07/04, S2\_08/04, S2\_09/04, S2\_06/05, S2\_08/05\}$ (site et date sont accolés),
- $A = \{\text{AZOT}^O, \text{AZOT}^J, \text{AZOT}^V, \text{AZOT}^B, \text{PHOS}^J, \text{PHOS}^V, \text{PHOS}^B, \text{IBGN}^O, \text{IBGN}^J, \text{IBGN}^B\}$ (^{O} pour *Orange*, ^{J} pour *Jaune*, ^{V} pour *Vert*, ^{B} pour *Bleu*)
- Les couples de la relation I sont désignés par une croix à la jonction d'une ligne et d'une colonne.

Site-Date	AZOT^O	AZOT^J	AZOT^V	AZOT^B	PHOS^O	PHOS^J	PHOS^V	PHOS^B	IBGN^O	IBGN^J	IBGN^B
$S1\_06/07$				×							
$S1\_07/07$			×					×			
$S1\_09/07$											×
$S1\_02/08$				×			×				
$S1\_04/08$			×								
$S1\_05/08$										×	
$S2\_07/04$	×					×					
$S2\_08/04$			×								
$S2\_09/04$									×		
$S2\_06/05$				×			×				
$S2\_08/05$										×	

TAB. 3 – *Jeu de données binaires obtenu à partir du tableau 2*

On considère maintenant un sous-ensemble d'objets $X \subseteq O$ et un sous-ensemble d'attributs $Y \subseteq A$. On peut alors définir deux opérateurs de Galois, notés $'$, s'appliquant aux sous-ensembles X et Y.

$$X' = \{y \in A | \forall x \in X : (x, y) \in I\} \texttt{ et } Y' = \{x \in O | \forall y \in Y : (x, y) \in I\}$$

Un concept formel est une paire (X, Y) où $X = Y'$ et $Y = X'$. X s'appelle l'extension et Y l'intension du concept. X est l'ensemble maximal d'objets décrits par tous les attributs de Y et Y est l'ensemble maximal d'attributs partagés par tous les objets de X. Par exemple, $\{S1\_06/07\}' = \{\text{AZOT}^B\}$ et $\{\text{AZOT}^B\}' = \{S2\_06/05, S1\_06/07, S1\_02/08\}$: le couple $(\{S2\_06/05, S1\_06/07, S1\_02/08\}, \{\text{AZOT}^B\})$ est un concept formel.

L'ensemble de concepts $\mathcal{C}_K$ construit sur le contexte $\mathcal{K}$ est muni d'un ordre, défini de la façon suivante. Soit $C_1 = (X_1, Y_1)$ et $C_2 = (X_2, Y_2)$ deux concepts formels, $C_1 \leq C_2 \Leftrightarrow X_1 \subseteq X_2 \Leftrightarrow Y_2 \subseteq Y_1$. C_1 est un sous-concept de C_2 et C_2 un sur-concept de C_1. Par exemple le concept $(\{S2\_06/05, S1\_06/07, S1\_02/08\}, \{\text{AZOT}^B\})$ est un sur-concept du concept $(\{S2\_06/05, S1\_02/08\}, \{\text{PHOS}^V, \text{AZOT}^B\})$. L'ensemble $\mathcal{C}_K$ muni de la relation $\leq$ est un treillis de Galois ou treillis de concepts, $\mathcal{L}_K = (\mathcal{C}_K, \leq)$. Le treillis obtenu à partir du contexte $\mathcal{K}_{\texttt{sites}}$ est présenté sur la figure 1.

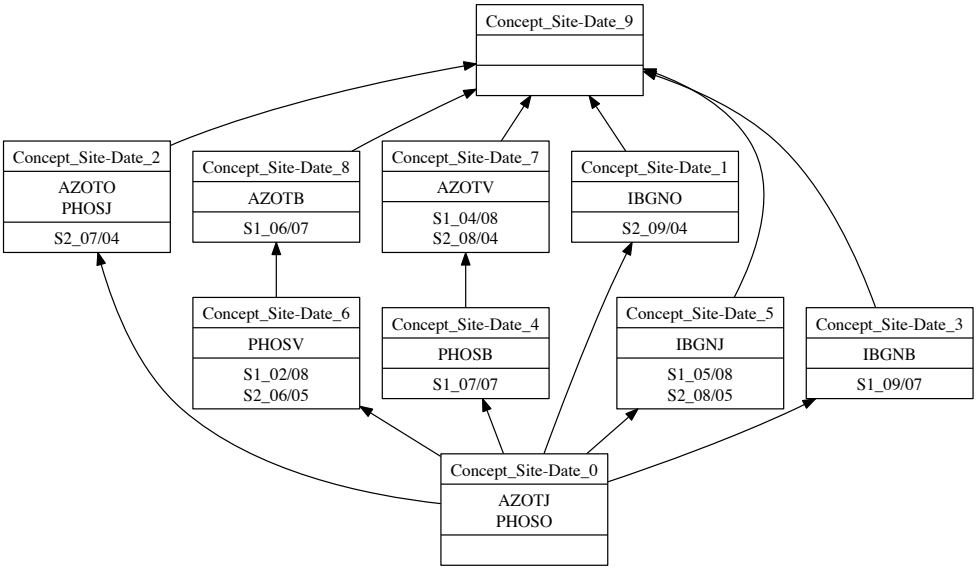


FIG. 1 – *Treillis de concepts obtenu à partir du contexte* $\mathcal{K}_{\texttt{sites}}$

Dans la figuration utilisée, chaque concept est représenté par son extension et son intension simplifiées : seuls les attributs et objets introduits par le concept y sont représentés. Chaque concept hérite des extensions simplifiées de ses sous-concepts et des intensions simplifiées de ses sur-concepts. Par exemple le concept dénommé Concept_Site-Date_6 correspond au concept $(\{S2\_06/05, S1\_02/08\}, \{\text{PHOS}^V, \text{AZOT}^B\})$ décrit ci-dessus ; l'attribut AZOT^B

n'apparaît pas dans l'intension simplifiée, il est hérité du sur-concept Concept_Site-Date_8. Le concept le plus général (en haut dans le treills) a une extension maximale, couvrant tous les objets, tandis que le concept le plus spécifique (en bas dans le treillis) a une intension maximale, couvrant toutes les propriétés. Dans ce treillis particulier, le concept le plus général a une intension vide (aucune propriété n'est partagée par tous les objets) et le concept le plus spécifique a une extension vide (aucun objet ne réunit toutes les propriétés).

4 Analyse relationnelle de données temporelles

Comme montré ci-dessus, l'analyse de concepts formels prend en entrée une unique relation binaire objet-attribut et ne permet donc pas d'exploiter la relation temporelle de précédence inscrite dans le jeu de données traité. C'est pourquoi nous utilisons l'analyse relationnelle de concepts qui permet de prendre en compte des relations inter-objets par application itérative de l'ACF sur un ensemble de contextes formels, qu'on appelle une famille relationnelle de contextes (Hacene et al., 2013). Plus formellement, une famille relationnelle de contextes est composée de n contextes formels objet-attribut $\mathcal{K}_i = (O_i, A_i, I_i)$, $i \in [1..n]$, et de m contextes relationnels objet-objet $\mathcal{R}_j = (O_k, O_l, r_j)$, $j \in [1..m]$ où O_k et O_l sont respectivement des ensembles d'objets de $\mathcal{K}_k$ et $\mathcal{K}_l$ et $r_j \subseteq O_k \times O_l$.

Lors de la première itération du processus ARC, un treillis est généré pour chaque contexte formel $\mathcal{K}_l$. Aux itérations suivantes, les concepts créés à l'étape précédente sont intégrés sous forme d'attributs dans les contextes formels $\mathcal{K}_k$ pour enrichir la description des objets. En effet un concept C_l du treillis $\mathcal{L}_{K_l}$ contient dans son extension un sous-ensemble d'objets de O_l et grâce à une opération d'échelonnage il est possible d'utiliser la relation inter-objets $r_j \subseteq O_k \times O_l$ entre les objets de $\mathcal{K}_k$ et les objets de C_l pour créer une relation objet-concept. Dans la suite nous utiliserons l'opérateur *existentiel* qui crée une relation $\exists r_j$ entre un objet $o \in O_k$ et un concept C_l dès que $r_j(o)$ a une intersection non vide avec l'extension de C_l. L'attribut $\exists r_j(C_l)$ est alors ajouté au contexte $\mathcal{K}_k$.

Au cours du processus, chaque contexte formel ainsi augmenté permet de générer un nouveau treillis où les attributs ajoutés peuvent faire émerger de nouveaux concepts. Lorsqu'aucun nouveau concept n'apparaît lors d'une nouvelle itération, le processus de l'ARC a atteint un point fixe et s'arrête.

4.1 Modélisation des données

Comme nous l'avons dit ci-dessus, le processus ARC prend en entrée une famille relationnelle de contextes qui contient un ensemble de contextes formels et un ensemble de contextes relationnels. La création d'une telle famille relationnelle de contextes implique d'évaluer la connaissance à obtenir et le type de la donnée initiale (par exemple temporelle, spatiale). Dans l'exemple traité, l'objectif est d'analyser l'influence des paramètres physico-chimiques sur les paramètres biologiques. La phase d'analyse s'appuie sur les aspects temporels, c'est-à-dire sur l'évolution de la qualité des paramètres prélevés par station durant une période de temps spécifique. Deux types de relations sont considérés, les relations temporelles (un prélèvement précède un autre prélèvement) et les relations de valeur (un prélèvement a pour valeur la classe *Jaune* ou *Verte*, etc.). Quatre contextes formels doivent donc être créés : le

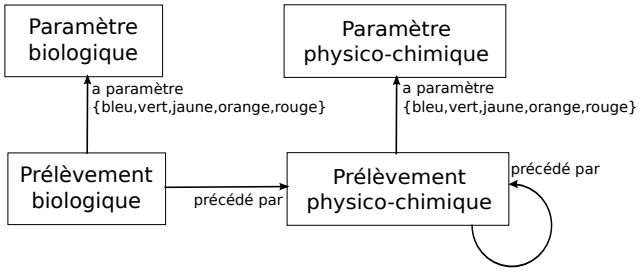


FIG. 2 – *Schéma relationnel illustrant la modélisation de la famille relationnelle de contextes*

contexte des paramètres biologiques (réduit ici à l'IBGN), le contexte des paramètres physico-chimiques (AZOT, PHOS), le contexte des prélèvements biologiques et le contexte des prélèvements physico-chimiques. La figure 2 illustre la mise en relation de ces différents contextes par les relations temporelles (entre contextes des prélèvements) et les relations de valeur (entre contextes des prélèvements et des paramètres).

Reconsidérons le tableau 2. Les contextes objet-attribut sont construits directement à partir des en-têtes ligne et colonne de ce tableau : le tableau 4 montre le contexte des paramètres biologiques, $\mathcal{K}_{paramBio}$ (à gauche), et le contexte des prélèvements biologiques, $\mathcal{K}_{prelBio}$ (au milieu). Le contexte des paramètres a pour particularité d'avoir un ensemble d'objets et un ensemble d'attributs identiques ($\{IBGN\}$ dans l'exemple). Le contexte des prélèvements a un ensemble d'attributs vide.

La construction des contextes relationnels exploite quant à elle les relations sous-jacentes au tableau 2. Celui-ci contient des données multi-valuées, c'est-à-dire qu'un paramètre y est décrit par différentes valeurs de qualité. À chaque valeur correspond un contexte relationnel entre les prélèvements biologiques et les paramètres biologiques ; de même pour les prélèvements physico-chimiques et les paramètres physico-chimiques. De manière concise, il y a sept relations binaires : $R_{aParamYX} = \mathcal{K}_{prelX} \times \mathcal{K}_{paramX}$ où $Y \in \{Vert, Bleu, Orange, Jaune\}$ si $X \in \{bio\}$ ou $Y \in \{Bleu, Jaune, Orange\}$ si $X \in \{phc\}$. Le tableau 4 (à droite) illustre un des contextes relationnels mentionnés précédemment. Enfin, les dates associées aux prélè-

TAB. 4 – *De gauche à droite : contexte des paramètres biologiques, contexte des prélèvements biologiques et contexte relationnel des prélèvements biologiques ayant la valeur Jaune*

$\mathcal{K}_{paramBio}$	IBGN
IBGN	×

$\mathcal{K}_{prelBio}$
S1_09/07
S1_05/08
S2_09/04
S2_08/05

$\mathcal{R}_{aParamJBio}$	IBGN
S1_09/07	
S1_05/08	×
S2_09/04	
S2_08/05	×

vements donnent une information sur les relations temporelles entre les contextes formels de prélèvements. La relation $\mathcal{R}_{bioPrecedeParPhC} = \mathcal{K}_{prelBio} \times \mathcal{K}_{prelPhC}$, exprime le fait qu'un paramètre physico-chimique a été prélevé avant un paramètre biologique sur une même station. La relation $\mathcal{R}_{phcPrecedeParPhC} = \mathcal{K}_{prelPhC} \times \mathcal{K}_{prelPhC}$ est un contexte relationnel cy-

clique. Elle indique qu'un paramètre physico-chimique a été prélevé avant un autre paramètre physico-chimique sur une même station.

Le processus ARC est orienté de la façon suivante. L'ACF est appliquée une seule fois sur les contextes de paramètres $\mathcal{K}_{paramBio}$ et $\mathcal{K}_{paramPhC}$ dont les ensembles objets et attributs sont identiques. Les contextes des prélèvements $\mathcal{K}_{prelBio}$ et $\mathcal{K}_{prelPhC}$, qui ne contiennent que des objets au départ, évoluent au cours du processus par ajout d'attributs relationnels, ce qui permet de mettre en évidence des répétitions, sur différentes stations, d'enchaînements de prélèvements et de leurs valeurs.

4.2 Mise en œuvre de l'ARC

Dans cette partie nous exploitons la famille de contextes relationnels décrite ci-dessus pour contruire un ensemble de treillis permettant de relier les contextes.

Tout d'abord, la procédure itérative de l'ARC commence par l'application d'un algorithme standard pour créer des hiérarchies de concepts à partir des quatre contextes formels $\mathcal{K}_{paramBio}$, $\mathcal{K}_{paramPhC}$, $\mathcal{K}_{prelBio}$ et $\mathcal{K}_{prelPhC}$. La figure 3 présente les résultats de cette première étape, résultats qui servent d'entrée à l'étape suivante.

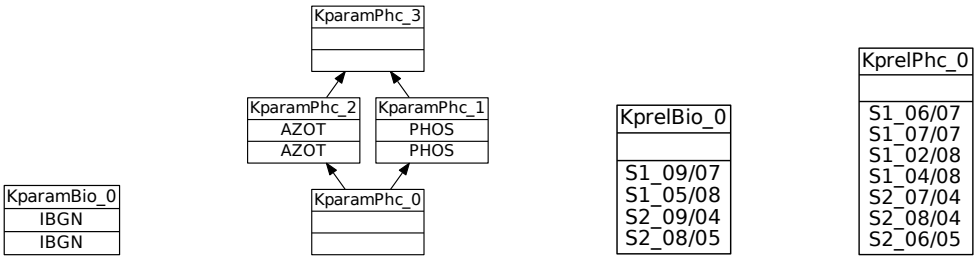


FIG. 3 – *De gauche à droite : treillis issus des contextes formels $\mathcal{K}_{paramBio}$, $\mathcal{K}_{paramPhC}$, $\mathcal{K}_{prelBio}$ et $\mathcal{K}_{prelPhC}$ lors de la première étape du processus ARC*

Dans un deuxième temps, les treillis de la figure 3 sont utilisés pour générer une nouvelle famille de contextes en utilisant l'opérateur d'échelonnage existentiel sur les relations $\mathcal{R}_{bioPrecedeParPhC}$, $\mathcal{R}_{phcPrecedeParPhC}$ et les contextes relationnels qui donnent la qualité des paramètres physico-chimiques et biologiques pour chaque échantillonnage. Le tableau 5 décrit le contexte augmenté $\mathcal{K}_{prelBio}+$ obtenu à cette deuxième étape. Brièvement, le contexte $\mathcal{K}_{prelBio}$ est étendu avec deux types d'attributs relationnels : les attributs de la forme générale $\exists\mathcal{R}_{aParamXBio(KparamBio\_0)}$, où $X \in \{Orange, Jaune, Bleu\}$ représentent le fait qu'un prélèvement biologique appartient à la classe de qualité X de l'indice IBGN ; l'attribut $\exists\mathcal{R}_{bioPrecedeParPhC(KprelPhC\_0)}$ dénote quant à lui qu'il y a au moins un prélèvement physico-chimique du concept $KprelPhC\_0$ précédant le prélèvement biologique considéré.

Le processus se poursuit de la même façon jusqu'à un point fixe obtenu ici après deux autres étapes. La figure 4 présente les treillis obtenus à l'issue du processus complet, durant lequel les deux treillis des paramètres biologiques et physico-chimiques ne changent pas.

Dans le treillis $\mathcal{L}_{\mathcal{K}prelBio}$, le concept $KprelBio\_3$ révèle que la classe de qualité *Bleu* de l'indice IBGN est influencée, durant 4 mois, par la classe de qualité *Bleu* du macro-paramètre

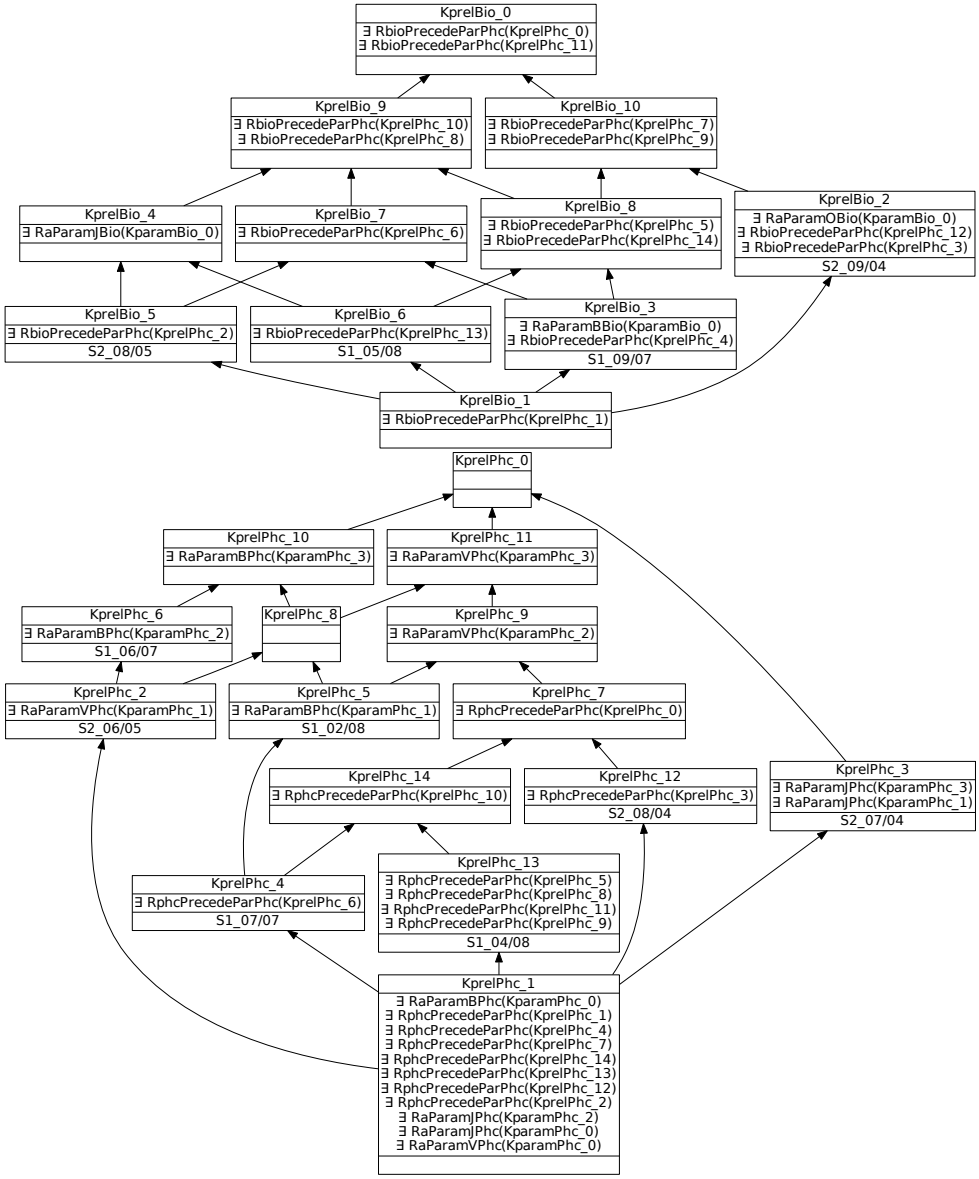


FIG. 4 – *Les deux treillis $\mathcal{L}_{\mathcal{K}prelBio}$ (en haut) et $\mathcal{L}_{\mathcal{K}prelPhC}$ (en bas) obtenus à la dernière étape du processus ARC (comparer avec leur forme initiale en figure 3)*

TAB. 5 – *Contexte augmenté $\mathcal{K}_{prelBio}+$ obtenu à la deuxième étape du processus ARC*

KprelBio	∃ RaParamOBio(KparamBio_0)	∃ RaParamBBio(KparamBio_0)	∃ RbioPrecedeParPhc(KprelPhc_0)	∃ RaParamJBio(KparamBio_0)
S1_09/07		×	×	
S1_05/08			×	×
S2_09/04	×		×	
S2_08/05			×	×

PHOS et par la classe de qualité *Vert* du macro-paramètre AZOT. De plus ces paramètres ont été affectés par la classe de qualité *Bleu* du macro-paramètre AZOT. La figure 5 montre l'enchaînement de concepts qui permet d'expliquer ce concept $KprelBio\_3$.

Parallèlement, le treillis $\mathcal{L}_{\mathcal{K}_{prelPhC}}$ contient le concept $KprelPhC\_12$ qui regroupe les prélèvements physico-chimiques décrits par la classe de qualité *Vert* du macro-paramètre AZOT. Cette valeur est influencée par la classe de qualité *Jaune* du même paramètre et du macro-paramètre PHOS.

Pour résumer, on voit dans ce petit exemple que l'ARC peut être utilisée pour révéler des relations temporelles entre paramètres physico-chimiques, qui ont une influence à court terme sur les valeurs des indices biologiques. Les résultats obtenus ici n'ont aucune valeur statistique compte tenu de la petitesse du jeu de données exemple utilisé.

5 Discussion et conclusion

De nombreux travaux ont porté sur l'exploitation de données hydroécologiques, que ce soit avec des approches statistiques classiques ou plus récemment, des méthodes d'apprentissage automatique. Les treillis ont été utilisés par (Bertaux et al., 2009), avec pour objectif d'aider l'expert à la constitution de groupes de taxons (ici des plantes aquatiques) partageant des caractéristiques communes. Les travaux exploitant les techniques d'apprentissage ont généralement pour objectif de mettre en relation des caractéristiques physiques ou physico-chimiques des rivières et les populations de taxons (faune ou flore) qui les habitent. Ainsi, Dakou et al. (2007) utilisent des arbres de décision pour prédire l'adéquation des habitats à certains macro-invertébrés, tandis que Kocev et al. (2010) utilisent les arbres de régression multiple pour étudier l'impact des conditions physico-chimiques du milieu sur des communautés d'algues

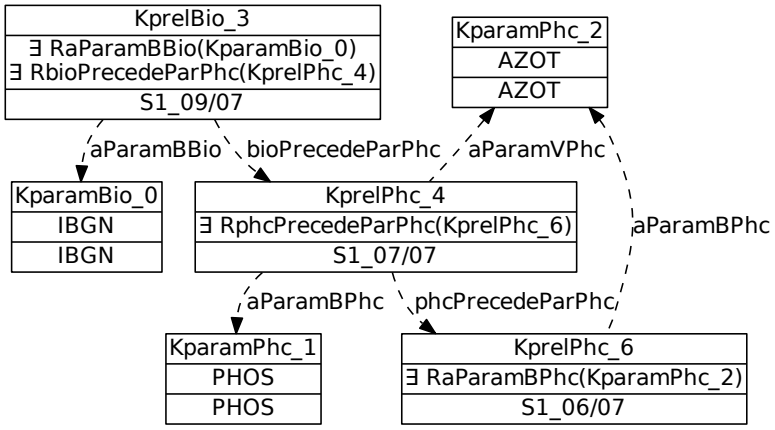


FIG. 5 – *Analyse du concept $KprelBio\_3$: liens entre concepts à travers les différents treillis*

microscopiques. Avec les mêmes techniques, Džeroski et al. (2000) ont cherché à prédire des valeurs de paramètres physico-chimiques à partir de paramètres biologiques (abondance des taxons). En revanche, à notre connaissance, les aspects temporels ne sont pas pris en compte comme nous l'avons fait ici et comme l'ont fait auparavant Fabrègue et al. (2014).

Par rapport à ce dernier travail, qui s'appuie sur la recherche de motifs partiellement ordonnés dans les séquences de données hydroécologiques, nous avons ici mis en œuvre une approche originale pour l'exploration de données temporelles fondée sur l'analyse relationnelle de concepts. Dans un délai proche, nous explorerons le même jeu de données que celui utilisé par Fabrègue et al. (2014). Sa taille importante conduira à des problèmes d'échelle, que nous pourrons résoudre en le segmentant selon les points d'intérêt des utilisateurs (par exemple selon les classes de valeurs des indices biologiques) ou en filtrant les concepts construits selon un seuil appliqué sur la taille de leur extension, ce d'autant que nous sommes à la recherche de sous-séquences fréquentes. Nous pourrons alors mener une comparaison complète avec les résultats obtenus par recherche de motifs.

Remerciements

Cette recherche est financée par l'Agence Nationale de la Recherche dans le cadre du projet ANR 11 MONU 14 Fresqueau.

Références

AFNOR (2004). Qualité de l'eau : détermination de l'Indice Biologique Global Normalisé (IBGN). XP T90-350.

Agrawal, R. et R. Srikant (1995). Mining sequential patterns. In *International Conference on Data Engineering*, ICDE, pp. 3–14.

Arévalo, G., J.-R. Falleri, M. Huchard, et C. Nebut (2006). Building Abstractions in Class Models : Formal Concept Analysis in a Model-Driven Approach. In *MoDELS 2006*, pp. 513–527.

Bertaux, A., F. Le Ber, A. Braud, et M. Trémolières (2009). Identifying Ecological Traits : A Concrete FCA-Based Approach. In *Formal Concept Analysis*, LNAI 5548, pp. 224–236.

Braud, A., S. Bringay, F. Cernesson, X. Dolques, M. Fabrègue, C. Grac, N. Lalande, F. Le Ber, et M. Teisseire (2014). Une expérience de constitution d'un système d'information multi-sources pour l'étude de la qualité de l'eau. In *Atelier "Systèmes d'Information pour l'environnement", Inforsid 2014, Lyon*.

Dakou, E., T. D'Heygere, A. P. Dedecker, P. L. Goethals, M. Lazaridou-Dimitriadou, et N. Pauw (2007). Decision Tree Models for Prediction of Macroinvertebrate Taxa in the River Axios (Northern Greece). *Aquatic Ecology 41*, 399–411.

Dolques, X., M. Huchard, C. Nebut, et P. Reitz (2012). Fixing Generalization Defects in UML Use Case Diagrams. *Fundam. Inform. 115*(4), 327–356.

Džeroski, S., D. Demšar, et J. Grbović (2000). Predicting chemical parameters of river water quality from bioindicator data. *Applied Intelligence 13*(1), 7–17.

Fabrègue, M., A. Braud, S. Bringay, C. Grac, F. Le Ber, D. Levet, et M. Teisseire (2014). Discriminant temporal patterns for linking physico-chemistry and biology in hydro-ecosystem assessment. *Ecological Informatics 24*, 210–221.

Ganter, B. et R. Wille (1997). *Formal Concept Analysis : Mathematical Foundations* (1st ed.). Secaucus, NJ, USA : Springer-Verlag New York, Inc.

Hacene, M. R., M. Huchard, A. Napoli, et P. Valtchev (2013). Relational concept analysis : mining concept lattices from multi-relational data. *Ann. Math. Artif. Intell. 67*(1), 81–108.

Kocev, D., A. Naumoski, K. Mitreski, S. Krstić, et S. Džeroski (2010). Learning habitat models for the diatom community in Lake Prespa. *Ecological Modelling 221*(2), 330–337.

Miralles, A., M. Huchard, X. Dolques, F. Le Ber, T. Libourel, C. Nebut, et A. Osman-Guédi (2015). Méthode de factorisation progressive pour accroître l'abstraction d'un modèle de classes. *Revue d'Ingéniérie des Systèmes d'Information*. À paraître.

Moha, N., A. R. Hacene, P. Valtchev, et Y.-G. Guéhéneuc (2008). Refactorings of Design Defects Using Relational Concept Analysis. In *ICFCA 2008*, pp. 289–304.

Summary

This article describes a temporal data mining method based on relational lattices. This method is applied on a sequence dataset, dealing with physico-chemical and biological parameters sampled in watercourses. Our aim is to reveal frequent sub-sequences linking the two parameter types. We use a small example to show that relational concept analysis (RCA) is able to highlight temporal impacts of physico-chemical parameters on biological parameters.

Vers un système d'Aide à la Décision Multicritères et Spatiotemporel pour la Surveillance Epidémiologique

Farah Amina Zemri*, Djamila Hamdadou*, Karine Zeitouni**

*BP 1524 El M'Naouer 31000 Oran Algérie
http://www.univ-oran.dz
zemriamina@gmail.com
dzhamdadoud@yahoo.fr
**45, avenue des Etats-Unis 78035 Versailles, France
Karine.Zeitouni@prism.uvsq.fr
http://www.prism.uvsq.fr

Résumé. Dans le contexte d'une approche guidée données visant à détecter les facteurs réels et responsables de propagation des épidémies et à expliquer son émergence ou réémergence, nous sommes amenés à étudier les systèmes SOLAP (Spatial On Line Analysis Processing), puis à proposer une nouvelle formulation de ces systèmes intégrant la FDS (Fouille de Données Spatiales) et les MADMC (Méthodes d'Aide à la Décision Multicritères). Cet article relate d'une part, la mise en oeuvre d'un système SOLAP pour la surveillance épidémiologique « EPISOLAP » permettant la détection et la localisation des foyers d'épidémies et d'autre part son couplage avec la FDS et les MADMC permettant l'évaluation prédictive du risque sanitaire en présence d'aléa et connaissant la vulnérabilité de la population exposée. L'architecture proposée constitue une seule plate-forme décisionnelle intégrée. En fin, nous pensons que l'exploitation des capacités de la FDS qu'offrent les méthodes et techniques d'extraction de connaissances à partir de données nous permettra d'avoir dans une perspective à court terme une meilleure précision de la prédiction et ainsi des résultats plus convaincants.

1 Introduction

La prévention des épidémies est une préoccupation de la santé publique et un réel challenge. L'identification des zones d'habitation (urbaines et non urbaines) exposées aux épidémies permettrait d'aider à la circonscription de ces phénomènes de santé publique par une stratégie de prévention et une gestion réfléchie. Notamment, la prise en charge médicale de ces maladies serait plus efficace. Dans ce contexte, le rôle du système d'information est essentiel dans l'accès à l'information et l'aide à la décision envers les organismes institutionnels de santé et les collectivités territoriales en charge de cette prévention.

Afin de pouvoir identifier une bonne stratégie de prévention contre les épidémies et afin d'assurer une gestion réfléchie du phénomène de propagation de celles-ci, il est important de

concevoir un système de surveillance épidémiologique de qualité, permettant la supervision de la maladie et l'identification des zones qui présentent des foyers d'épidémies.

Traditionnellement, les systèmes d'information décisionnels s'appuient largement sur des entrepôts de données offrant des outils d'exploration multidimensionnelle des données et des indicateurs par l'analyse en ligne ou OLAP (Online Analytical Processing). S'agissant des phénomènes comme l'épidémie, les dimensions spatiales (et temporelles) sont éminemment importantes dans l'analyse et la qualification de la propagation pas du phénomène au voisinage, ainsi que son émergence ou réémergence. Les dernières années ont vu le développement de travaux sur le Spatial OLAP ou SOLAP (Bédard et al, 2005), (Bimonte et al, 2010) permettant d'intégrer les données spatiales dans l'OLAP et de lier l'exploration et la visualisation cartographique. La démarche naturelle est donc d'appliquer ces approches et les systèmes qui ont en découlé (Zemri et al, 2013) à l'analyse épidémiologique, ce que nous avons fait. Par la suite, des limites ont été détectées. Ce système ne donne pas une interprétation des résultats de l'analyse multidimensionnelle.

L'article décrit en section 2 la problématique. En section 3, notre contribution. En section 4, les principaux travaux menés dans le domaine d'aide à la décision spatiale utilisant les MADMC couplées avec les outils SIG sont présentés. La section 5 donne une description du système EPISOLAP. La section 6 décrit, quant à elle, l'approche proposée ainsi que le système EPISOLAP-MINING suggéré. Une formulation du problème multicritères est illustrée, en détail, en section 7 qui englobe une étude de cas réelle qui constitue en soit une première validation de notre approche. Enfin, nous concluons notre propos, en section 8, en donnant quelques perspectives.

2 Problématique

Le terrain d'étude dans le cadre de ce travail est la région d'Oran en Algérie. L'établissement d'une géo-localisation des zones d'habitation de population, dites « pauvres » plus exposés, permettrait d'observer les lieux où une épidémie pourrait s'étendre rapidement et le plus largement. De l'avis d'un épidémiologique de l'université d'Oran, la tuberculose représente encore un danger persistant menaçant la population de la région d'Oran (Bouziani 2000).

Le plan de relance de la lutte contre la tuberculose (2006-2015) fait partie des Objectifs du Millénaire pour le Développement (OMD), et de la nouvelle stratégie « Halte à la tuberculose » recommandée par l'OMS depuis 2006 :stopper l'augmentation de l'incidence de la tuberculose et commencer à la réduire sur tout le territoire national » (Agadir et al 2011). Or, avec l'évolution démographique et la situation socioéconomique de l'Algérie, réaliser un tel plan reste un défi et un objectif difficile à atteindre sans outil d'aide et de pilotage efficace.

Une des principales mesures adoptées par ce programme national est l'amélioration de la déclaration des cas de la tuberculose et de leur suivi par la généralisation du système électronique de surveillance. Notre projet de recherche s'intègre dans cette optique pour aider à la réalisation de tels objectifs en élaborant un système d'aide à la décision spatiale (SDSS) pour la surveillance épidémiologique. Ce système vise à bien cerner le problème de propagation de la tuberculose.

A cet égard, peu d'études étés menées sur l'impact transformations socioéconomiques que connait la zone d'étude et sur les scénarios environnementaux ayant à la propagation et à la transmission de la maladie.

3 Contribution

L'informatique décisionnelle apporte des solutions nouvelles pour la modélisation, l'interrogation et la visualisation de données dans un objectif d'aide à la décision. Les modèles multidimensionnels ou modèles d'hyper-cube sont des modèles qui permettent de structurer les données pour l'analyse décisionnelle en explicitant les notions de fait et des dimensions.

L'intégration des données spatiales dans les systèmes OLAP est un besoin réel dans de nombreuses applications. En effet, l'information géographique est très fréquemment présente implicitement ou explicitement dans les données, mais généralement sous-employée dans le processus décisionnel. Le couplage de systèmes OLAP et de Systèmes d'Information Géographique (SIG) au sein de systèmes OLAP appelé Spatial OLAP (SOLAP), est une voie prometteuse. Notre première contribution a été de mettre en ?uvre une solution SOLAP pour l'épidémiologie. Mais pour aller plus loin dans l'analyse à la fois explicative et prédictive, le SOLAP seul n'est pas suffisant. Notre seconde contribution dans cet article est la proposition d'une combinaison de la technologie SOLAP avec les Méthodes d'Analyse Multi Critères (MAMC) et les techniques de Fouille de Données Spatiales (FDS) offrant une analyse des données plus riche.

4 Travaux connexes

Dans le cadre de l'aide à la décision utilisant les SIG, plusieurs systèmes d'aide à la décision ont retenu notre attention. Dans (Joerin, 1997) le système MEDUSAT a été proposé pour optimiser la localisation spatiale de site d'une usine de traitement des déchets en Tunisie. MEDUSAT combine un outil SIG permettant la création de zones homogènes déterminées à partir de données spatiales (constituant un indice de similarité). Ces zones constituent l'ensemble des actions qui sont ensuite évaluées grâce à des Méthodes d'Aide à la Décision Multicritères. Un processus de décision pour la gestion de l'eau en milieu urbain dans (Mottier, 1999). Dans (C.Boulemia et al, 2000), les auteurs ont présenté des outils pour l'aide à la décision dans les communautés locales afin de résoudre également des problèmes de gestion de l'eau. Dans (Bensaid, 2007), l'analyse multicritères a été utilisée comme un outil pour la prise de décision pour la localisation spatiale des zones sous forte pression humaine. Une étude de cas dans le département de Naama en Algérie a été présentée dans le même ouvrage. Différents systèmes d'aide à la décision riches en outils spatiaux et en méthodes d'analyse multicritères ont été développés pour la gestion et la prise de décision dans les problèmes territoriaux (eau, air, zones naturelles, transports, énergie, déchets, planification, santé, gestion des risques,...) (Bensaid et al, 2000).

Tous ces systèmes intègrent les outils d'analyse multicritères couplés avec un SIG, mais considèrent les critères comme indépendants et sont incapables de modéliser leur interactions (interchangeabilité, corrélation, dépendance, préférentiel...). Dans (Hamdadou, 2007), nous avons discuté de manière significative l'inclusion de critères de corrélation, dans les MADMC, notamment dans la méthode "Electre Tri" (Roy, 1985), en introduisant l'intégrale de Choquet (au lieu de la somme arithmétique) comme opérateur d'agrégation. Dans (Hamdadou, 2009), l'objectif principal a été développé un système d'aide à la décision, pour la modification d'itinéraire dans le cas du transport de matières dangereuses. Enfin, un système d'aide à la décision multicritères pour le diagnostic industriel a été développé dans (Bouamrane et al, 2012).

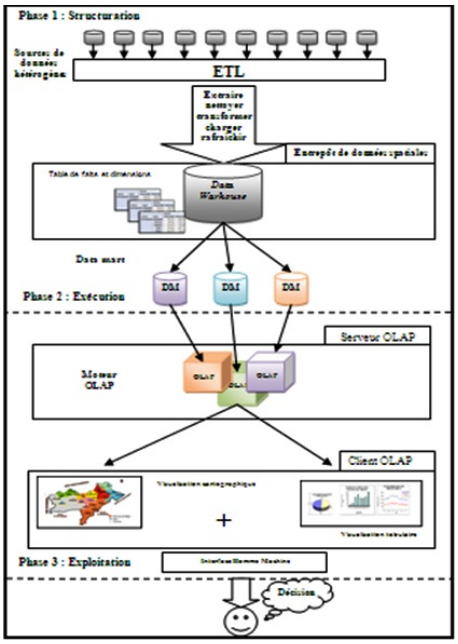


FIG. 1 – *Système d'Aide à la Décision pour la Surveillance Épidémiologique "EPISOLAP"*

5 Description du Système EPISOLAP

L'étude qui a été menée, par notre travail antérieur dans le projet "EPISOLAP" visait à identifier et à prédire le risque sanitaire selon les données de la surveillance relevées des différentes structures de santé spécialisées.

L'analyse est faite par l'outil EPISOLAP, mais elle se limite aux données alphanumériques et n'exploite guerre la localisation géographique et le lien de voisinage. Nos travaux actuels visent à faire émerger, partant de l'ensemble de ces données, des structures pertinentes du risque sanitaire qui puissent soutenir l'effort de surveillance, orienter l'action d'éradication et renforcer le système de la prévention. Dans la figure 2 nous présentons quelques résultats obtenus par l'analyse des données de surveillance par le système « EPISOLAP ».

Plus particulièrement, il s'agit d'outrepasser les limites de « EPISOLAP » et d'intégrer le caractère spatial des données (ici les foyers d'épidémie) et l'interaction avec l'environnement géographique via le concept de relations de voisinage, ce qui est très important s'agissant d'une malade à propagation rapide.

Ceci permettrait, dans cet exemple d'application, d'expliquer et de prédire le risque sanitaire menaçant la population en tenant compte de leur contexte géographique.

- A : Nombre de cas par commune
- B : Les taux d'incidence par âge et par commune
- C : Le taux d'incidence par commune
- D : Taux d'incidence par type de la maladie et par commune

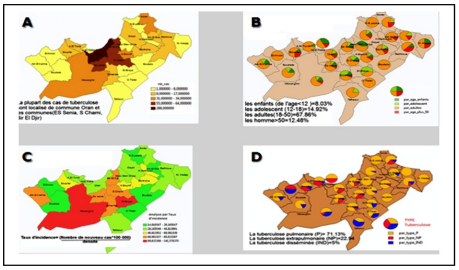


FIG. 2 – *Géo localisation et affichage cartographique- Représentation par commune : du nombre de cas (A) ; du taux d'incidence par âge (B) ; du taux d'incidence (C) ; du taux d'incidence par type de maladie (D).*

6 Approche proposée

Notre approche comprend deux types d'extension des systèmes SOLAP. L'un est d'ordre exploratoire et intégré l'identification des foyers en SOLAP avec la priorisation des actions à l'aide des MADMC. Le second est d'ordre préventif et intégré l'entrepôt de données et la navigation SOLAP avec des techniques de fouille de données spatiales prédictives permettant de mettre en évidence les facteurs de risques. Nous détaillons ci-dessous ces deux extensions.

6.1 Intégration SOLAP-MADMC

Le système « EPISOLAP » avait comme objectif la détection et la localisation des foyers d'épidémie. Or, dans le cas ou ces foyers sont nombreux, il faut pouvoir établir des priorités afin de traiter les foyers les plus à risque. Nous avons donc proposé d'effectuer un classement de ces foyers en utilisant les différents critères qui rentrent dans l'identification du risque sanitaire. Pour ce faire, nous proposons l'intégration des méthodes d'analyse multicritères qui sont des méthodes formelles ayant prouvé leur efficacité dans le domaine spatial et ont démontré leur capacité à cerner des problèmes spatiaux. Ces méthodes ont été appliquées dans les différentes études menées dans notre équipe depuis une décennie et dans des domaines différents (transport, aménagement du territoire.) la conception d'un système d'aide à la décision spatial basé sur une telle intégration permettrait de cerner le phénomène de la propagation épidémiologique, d'améliorer la surveillance épidémiologique et enfin rendre ce problème mieux contrôler les actions des décideurs. En somme :

- Le rôle du SOLAP serait la localisation des foyers d'épidémies,
- Le rôle des MADMC serait le classement des foyers d'épidémie pour faciliter l'intervention par ordre de priorité.

6.2 Couplage SOLAP-FDS

Le but de l'intégration de la technologie SOLAP et les techniques de FDS est de décrire et d'expliquer le phénomène d'épidémie observé avec « EPISOLAP » dans le but de mieux les appréhender, voire de les éviter. Plus exactement, la FDS aidera dans la découverte des relations de corrélations entre les phénomènes spéciaux et donnera une description précise

des scénarios épidémiologiques ce qui devrait améliorer la compréhension des facteurs de risque d'épidémie et orienter les actions pour les prévenir. Vis-à-vis de notre application de surveillance épidémiologique, ce couplage remplira les fonctions suivantes :

- L'outil SOLAP permettra de représenter l'aspect spatial et cartographique de notre étude grâce aux outils qu'ils combinent pour la détection et la localisation des foyers d'épidémies.
- La FDS permettra de mettre en évidence certains aspects socio environnementaux et transformations socio économiques qui influencent la dynamique spatiale de la tuberculose et contribuent dans les différents scénarios de la propagation et la transmission de la maladie dans la région d'Oran, ceux qui n'avaient pas été jusqu'à présent formalisés, ou du moins modélisés.

L'avantage du SOLAP est de fournir une analyse en ligne, une visualisation simple et rapide de l'information, une vision multidimensionnelle des données et une analyse spatiotemporelle sur une carte géographique. La fouille de données spatiales (FDS) permet quant à elle, d'extraire des connaissances à partir des données et offre une grande variété de méthodes avec des objectifs d'analyse différents.

On peut regretter d'une part que la FDD ne se fasse généralement pas en ligne et qu'elle ne traite que des données représentées sous forme tabulaire (tableau individus-variables) et d'autre part que les opérateurs SOLAP classiques permettent seulement d'agréger, de visualiser et d'explorer les données. Ceci rend le couplage entre ces deux outils plus que nécessaire pour réaliser cette complémentarité, il reste cependant plusieurs défis à lever qu'on explicite par les interrogations suivantes :

- Comment SOLAP et les entrepôts peuvent intégrer des algorithmes de FDS ?
- Comment stocker dans un entrepôt les connaissances extraites par une méthode de FDS ?
- Comment exécuter en ligne, sur des cubes parfois de grande taille, des algorithmes de fouille d'une certaine complexité et consommateurs de temps ?
- Comment modéliser de façon multidimensionnelle de telles données complexes (données géographiques) ?
- Comment faire une analyse en ligne sur ces données spatiales ?

Il est évident que l'implémentation d'un tel couplage ne sera pas sans difficultés et devra être effectuée sous le respect de certaines contraintes. En effet, contrairement à la fouille sur des données provenant de sources autres que des cubes de données, la fouille basée sur OLAP devrait tenir compte du contexte. À un instant t donné, l'utilisateur d'un cube visualise des faits avec différents niveau de granularité. On peut donc se servir de cette « photographie » instantanée des données pour effectuer la fouille de données.

La FDS et SOLAP étant issues de deux communautés scientifiques différentes, ceci explique que peu de travaux de recherche traitent du couplage entre les deux domaines. Toutefois, la vision d'intégrer l'OLAP et de la fouille de données n'est pas nouvelle. En effet, Jiawei Han avait introduit l'OLAM (On Line Analytical Mining) dans (Han, 1998), (Guran et al, 2009) La figure 3 présente le modèle d'aide à la décision intégrée « EPISOLAP MINING » proposé.

7 Formulation du problème

Le premier défit serait de modéliser le problème de la surveillance épidémiologique en un problème d'analyse multicritères en tenant compte des différents facteurs influant la propaga-

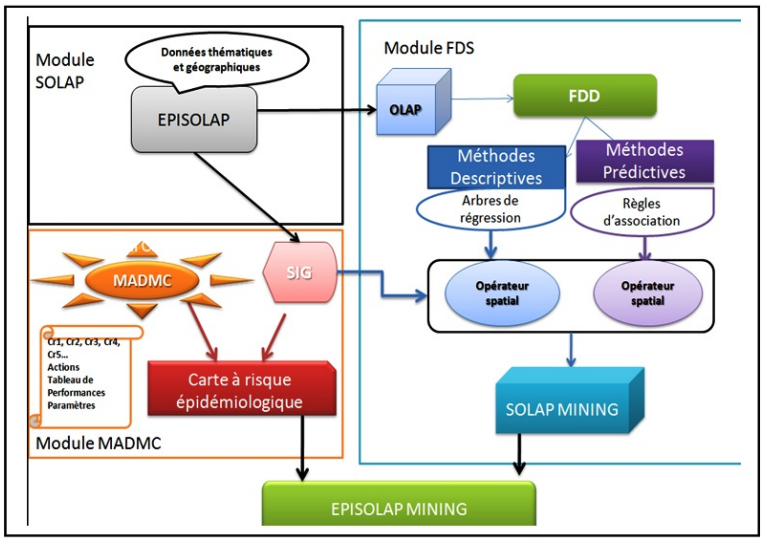


FIG. 3 – *Modèle d'Aide à la Décision "EPISOLAP MINING"*

tion de la maladie.

En suite, il faudra modéliser et intégrer des scénarios de l'impact environnemental et socio-économique dans un entrepôt de données spatio-temporel et appliquer les techniques de fouille de données spatiales afin d'expliquer les scénarios prospectifs et donner des interprétations précises aux phénomènes spatiaux.

Voici une liste non exhaustive des facteurs considérés par notre étude dans la surveillance épidémiologique :

- Facteurs environnementaux et climatiques : Pollution, humidité, température, ensoleillement,
- Facteurs spatiaux : localisations des foyers et leurs proximité par rapport à la mer, lac et des zone humides) et la considération du lien de voisinage par rapport aux autres foyers,
- Facteurs temporels : évaluation des périodes et des saisons de l'année où on peut enregistrer des pandémies dans ces régions,
- Facteurs socio économiques des individus : les conditions de vie personnelles et professionnelles des sujets malades ou susceptibles de devenir malades (taux de chômage, total des personnes recevant une allocation, indice sur l'échèle mondial de la pauvreté (ex : « index of deprivation »), total d'immeubles recensés par unité spatiale, condition d'habitat « typologie » (quantitatifs : dense, diffus, dispersés ; qualitatif : spontané, résiduel, mixte), insalubrité d'habitat (zones impropres à l'urbanisation, informel, non équipées en VRD, voiries et réseaux divers, etc.),
- Facteurs démographiques : densité de la population (surpeuplement), taux d'immigrants,
- Etc.

Tous ces facteurs et d'autres seront analysés et conjugués à l'aide des méthodes multicritères pour en tirer un classement des régions qui sont des foyers d ?épidémies ou susceptibles de devenir des foyers d'épidémie (de la zone la plus risquée à la moins risquée). Ainsi les spé-

cialistes et épidémiologistes pourront intervenir d'une manière proactive sur ces régions par ordre de priorité pour minimiser le taux de mortalité et de morbidité. Les décideurs en santé peuvent savoir quant et où intervenir, tout en ayant une idée claire et précise sur le potentiel du risque sanitaire encourue et sur la gravité de la situation épidémique rencontrée.

7.1 Intégration SOLAP-MADMC

Une des solutions proposées dans cette étude pour augmenter la détection des cas est d'identifier les populations à risque de tuberculose. L'objectif de ce volet de l'approche est justement de comprendre la dynamique de la tuberculose à Oran en fonction des facteurs cités ci-dessus afin de créer un index de vulnérabilité des populations vivant dans la région, et de produire, au final, une cartographie du risque de tuberculose. Le risque est évalué en fonction de la fréquence d'apparition d'un ou plusieurs paramètres autour des cas. On définit le risque par la présence simultané de l'aléa et de la vulnérabilité de la population Ainsi, il est possible de créer un indice de risque en fonction des paramètres qui identifient la présence de l'aléa (ex : la présence des cas contagieux) et les paramètres identifiant la vulnérabilité de la population (ex : le surpeuplement, la localisation, le type d'habitat ou encore, le taux de chômage).

7.1.1 Conception multicritères

Le modèle décisionnel proposé basé sur l'aide à la décision multicritères est largement inspirée de celui proposée par (Hamdadou, 2008) et repris dans (Zemri et al, 2012). Ce modèle se base sur les trois phases suivantes :

1. La phase de structuration

   Cette première phase vise à identifier les paramètres à prendre en compte telles que la situation géographique de la zone d'étude à l'aide d'un SIG et les différents critères et actions. Cette phase vise également à formaliser deux éléments fondamentaux de la situation décisionnelle :

   (a) Identification des actions : l'identification de toutes les actions potentielles est une étape très importante dans toute approche d'aide à la décision, en particulier lorsque les méthodes d'analyse multicritères procèdent par agrégation partielle. Il est très important que l'ensemble de toutes les actions soit complet car sa modification au cours de l'analyse peut entraîner une récurrence de l'analyse multicritères.

   (b) Identification des critères : la liste des critères obtenue en rassemblant les facteurs correspondants (sous-critères) devrait être aussi complète que possible. Ces critères doivent être liés à des contraintes et des objectifs utilisés dans les activités de production. La famille de critères les plus pertinents doit vérifier les conditions de l'exhaustivité, la cohérence et l'indépendance.

2. La phase d'exécution

   Cette deuxième phase est le processus d'analyse de l'étude. Ses deux principaux objectifs sont l'évaluation des critères, puis l'agrégation de ces informations par une analyse multicritères en exploitant les méthodes de la classification multicritères à savoir (les méthodes de la famille PROMETHEE)

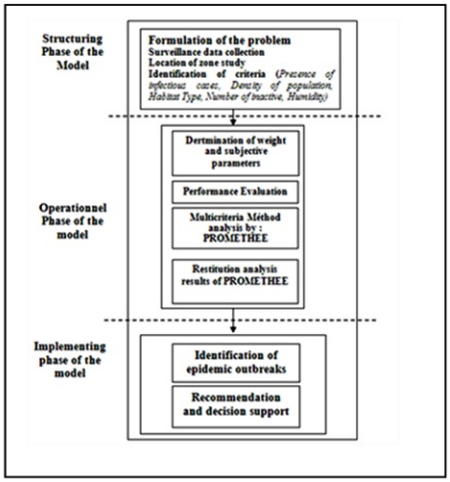


FIG. 4 – *Modèle Décisionnel Multicritéres proposé.*

3. La phase d'exploitation

   Cette troisième phase est principalement le résultat de l'acceptation sociale. Toutefois, elle comprend également la mise en oeuvre de la décision et le contrôle de cette exécution.

Les principales phases et les étapes du modèle proposé sont illustrées sur à la figure 4.

L'architecture fonctionnelle du système décisionnel proposé basé sur l'aide à la décision multicritères est illustrée sur le diagramme de la figure 5.

7.2 La méthode multicritères PROMETHEE

Le choix de PROMETHEE (Préférence Ranking Organisation Method for Enrichment Evaluations) (Hamdadou et al, 2012) s'explique par les avantages suivants :

- La simplicité et l'aspect intuitif de la méthode
- La puissance de sa fonction de préférence
- La simplicité de la phase d'exécution de la méthode

La méthode PROMETHEE I fournit à l'utilisateur un classement des différentes actions (foyers). Le problème est que cette méthode ne classe pas toutes les actions. Certaines actions peuvent rester incomparables. Méthode PROMETHEE II permet d'enlever cette in comparabilité. Le principe de cette méthode est de créer un procédé numérique de comparaison de chaque action par rapport à toutes les autres. Ainsi, il est possible de calculer plus (mérite) ou moins (démérite) de chaque action par rapport à tous les autres. Le résultat de cette comparaison permet la classification ordonnée des actions. La mise en ?uvre du procédé revient à exécuter les trois étapes suivantes :

1. Choix des critères généraux : Chacun des critères C1, C2 ... Cm est associé à un critère généralisé basé sur une fonction de préférence et les effets d'échelle sont éliminés.

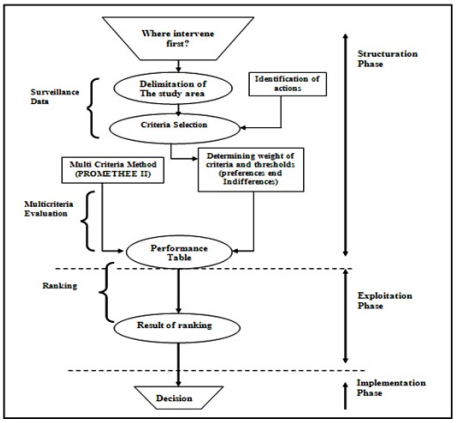


FIG. 5 – *Diagramme d'utilisation du modèle décisionnel proposé*

2. Détermination d'une relation de surclassement : Dans une deuxième phase, il est nécessaire de déterminer une relation de surclassement à travers un indice de préférence qui quantifie les préférences du décideur. L'intensité de la préférence est calculée comme suit : $p(D) = 0$ si $d \square q_j$, $p(D) = (d - q_j)/(p_j, q_j)$ si $q_j < d \square p_j$ et $p(D) = 1$ si pas a et b sont deux actions potentielles, d est la différence entre la performance de a et la performance de b $(GJ(a) - g_j(b))$. q_j est le seuil d'indifférence, et p_j est le seuil de préférence.
3. Évaluation des préférences : L'évaluation de la préférence du décideur est assurée par l'inclusion de flux entrants et sortants.
   (a) Calcul de l'indicateur de préférence

   $$(a, b) = \Sigma W_j * P_j(a, b)/\Sigma W_j \text{ ; avec } W_j \text{ est le poids du critère } j$$

   (b) Calcul des flux entrants

   $$\phi_+(a) = \Sigma(a, x)$$

   (c) Calcul des flux sortants,

   $$\phi_-(a) = \Sigma(a, x)$$

   (d) Calcul des flux mondiaux,

   $$\phi(a) = +\phi - \phi - (a)$$

Le principe général de fonctionnement par PROMETHEE (I et II) est donné par l'organigramme illustré par la figure 6.

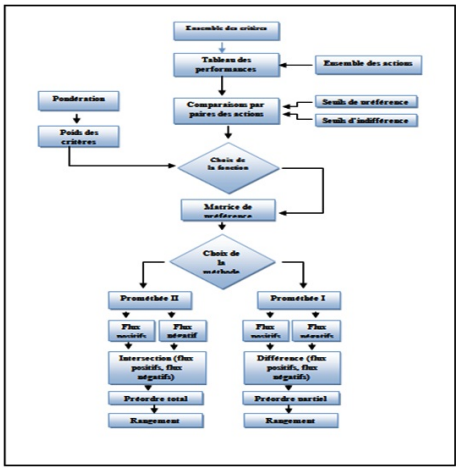


FIG. 6 – *Processus de la méthode PROMETHEE*

7.3 Expérimentation

L'étude de cas est la surveillance de la maladie de la tuberculose dans la région d'Oran en Algérie.

1. Identification de la zone d'étude

   Le terrain d'étude est la région d'Oran en Algérie qui est une des régions de l'ouest de l'Algérie et qui, souffre encore de la maladie de la tuberculose. Elle est classée en deuxième position dans le classement national du pays (Agadir et al 2011).

2. Identification des actions :

   Les actions considérées dans notre étude multicritères sont les foyers d'épidémies représentées par les 26 communes de la wilaya d'Oran.

3. Identification des critères

   Notre choix dans cette conception multicritères était porté sur les cinq critères suivants :

   (a) Aléas : Présence de cas contagieux (Critère Médicale)

   L'aléa de la tuberculose est définit en trois classes en fonction de la présence des cas contagieux (les cas prouvés) et de la présence de cas de tuberculose non contagieux (non prouvés) mais potentiellement contagieux. L'aléa est fort si, dans le foyer considéré, il y a au moins un cas contagieux ; l'aléa est faible si il n'y a aucun cas de tuberculose détecté comme contagieux mais qu'il y a au moins un cas de tuberculose recensé dans le foyer ; l'aléa est nul si aucun cas de tuberculose n'a été recensé pendant toute l'année (probabilité de l'évènement "présence d'un cas de tuberculose" = 0).

   (b) La densité (Critère démographique)

   Il existe dans la littérature peux d ?informations concernant la « vulnérabilité sociale » et les données socio-économiques qui peuvent y être associées. Parmi les

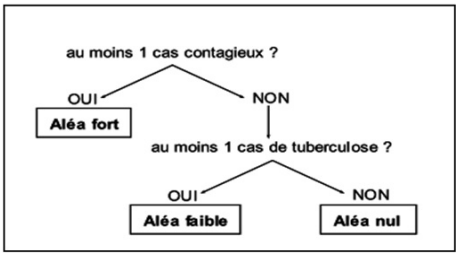


FIG. 7 – *Différentes classes du risque de la tuberculose*

facteurs socio-économiques liés à la présence de la tuberculose, et selon la disponibilité de ces paramètres vulnérables, Notre choix était porté sur le paramètre de « la densité de la population » qui nous paraisse le plus significatif. En effet, une surpopulation dans une région donnée favorise la contamination et la propagation des épidémies y compris la tuberculose qui suffit un contact direct avec le sujet atteint de la maladie pour transporter le virus par voix aérienne. Une population plus dense est donc plus vulnérable d'être contaminé en présence bien sur de l'aléa qui est ici la présence du cas contagieux.

(c) Typologie d'habitat (Critère démographique)

Ce critère est calculé par le nombre de constructions précaires occupant généralement les bidonvilles.

(d) Nombre des inactifs (Critère démographique)

Le nombre des inactifs est calculé dans notre étude par le nombre des personnes sans emploi dans les régions.

(e) Humidité (Critère climatique)

L'humidité du climat est un paramètre très important qui favorise la propagation des bactéries (par exemple, le bacille de coq responsable de la maladie de la tuberculose). Etant dans l'impossibilité d'avoir l'humidité annuelle moyenne de toutes les régions d'Oran (présence de deux stations de mesure seulement : Sénia et Arzew), nous avons identifié, à l'aide d'un spécialiste dans la météo au bureau national de la météo à Oran, une échelle de quatre points (1-2-3-4) qui classe les régions de la plus humide (mesure = 1) à la plus sèche (mesure = 4).

4. Génération du tableau de performances

Les couches d'informations intervenant dans notre modèle du risque tuberculeux, c'est-à-dire l'aléa et la vulnérabilité de la population : la densité, type d'habitation, le taux de l'humidité et le niveau de la pauvreté, croisés dans le système SOLAP, faisant apparaître les zones au niveau desquelles les interactions entre une population vulnérable face à la tuberculose et des patients atteints de tuberculose qui sont susceptibles de transmettre la maladie (Prouvé/Non prouvé) sont les plus intenses.

L'objectif de cette modélisation multicritères du problème de la propagation de la maladie de la tuberculose va aboutir à une classification des foyers d'épidémie de la plus favorable à la moins favorable. Ces zones sont supposées être les plus risquées ou en-

core les plus « dangereuses épidémiologiquement » pour la tuberculose humaine ce qui permettra d'élaborer une carte de risque tuberculeux.

Les actions dans notre étude d'analyse multicritères sont les foyers d'épidémie détectés préalablement par le système « EPISOLAP » (26 foyers envisagés).

	Nbr de cas contagieux	La densité de la population	Nbr de const précaires	Nbr des inactifs	Humi dité
Oran	196	767558	5125	9928	1
Gdyel	16	36314	101	760	2
Bir El Djir	49	88398	1305	2683	2
Hassi Bounif	25	54046	116	1356	3
Es Senia	38	78433	570	1345	2
Arzew	22	80761	120	788	1
Bethioua	6	17840	53	342	1
Marsa El Hadjaj	4	12449	44	304	1
Ain Turk	6	31776	197	671	1
El AnÃ§ar	0	9597	82	112	1
Oued Tlilet	2	16086	78	225	4
Tafraoui	2	12090	5	289	4
Sidi Chami	41	71243	1282	2324	2
Boufatis	3	11991	11	189	2
Mars El kebir	3	17149	73	357	1
Bousfer	4	13480	13480	401	1
El Karma	11	16507	255	422	2
El Braya	4	4696	15	206	3
Hassi Ben Yabka	5	11421	51	297	3
Ben Freha	11	17631	75	446	3
Hassi Mefssoukh	6	9267	113	346	3
Sidi Ben Yabka	2	7133	44	116	3
Messerguine	17	21896	286	629	3
Boutlelis	5	21303	65	564	3
Ain El Karma	5	8449	0	122	1
Ain El Biya	7	31778	5	395	2

TAB. 1 – *Tableau de performances.*

5. Poids des critères

Dans le cas fréquent où l'analyse des conséquences des actions potentielles conduit à la construction de plusieurs critères multiples est l'analyse des critères d'apporter des réponses au problème.

Pour une action donnée, et pour chaque critère un seuil de préférence p, d'indifférence q et le seuil de veto v sont estimés, en sachant que les méthodes PROMETHEE n'exploitent pas le seuil de veto qui est un paramètre subjectif exprimé par le décideur. Chaque critère est affecté d'un seuil de préférence p, d'un seuil d'indifférence q et d'un

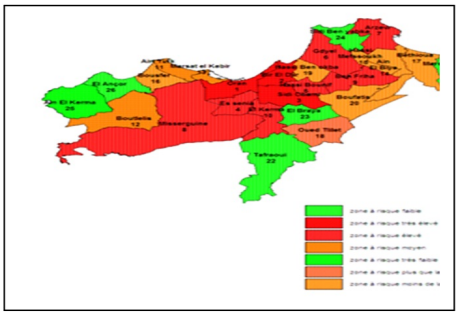


FIG. 8 – *Carte de risque tuberculeux avec Map Info Professional 11.0*

poids k reflétant sa contribution à la décision finale. Le résultat de l'analyse des conséquences est présenté dans un tableau de performances.
Pour des raisons de simplification, nous avons choisi que : p (PREF) = q (INDI) = 1

	Cr1	Cr2	Cr3	Cr4	Cr5
POIDS	3	2	1,5	2	1, 5
PREF	1	1	1	1	1
IND	1	1	1	1	1

TAB. 2 – *Tableaux de paramètres subjectifs.*

6. Résultat du rangement
   La méthode multicritères qui a été mise en oeuvre pour le classement des foyers d'épidémies sur la carte de la région d'Oran est la méthode PROMETHEE II. Celle là construit une relation de surclassement, basée sur la comparaison des actions en paires ; son but est de stocker les actions de la meilleure à la pire.
   La figure 8 montre une carte du risque qui donne une classification des foyers d'épidémies de la région la plus à risque épidémiologique à la région la moins à risque épidémiologique en utilisant une légende. Le résultat de classement a été visualisé sur une carte à l'aide du SIG Map Info Professional 11.0.

8 Conclusion et perspectives

Nous avons pu, grâce à cette étude de cas, voir l'efficacité de notre modèle proposé dans la surveillance épidémiologique, en termes de présentation des informations pertinentes aux décideurs en santé publique leur permettant de voir les relations entre les phénomènes, tout en les incitant à découvrir les connaissances et produire la bonne décision pour agir efficacement dans le temps et dans l'espace. Ce processus de prédiction généré par l'intégration de la technologie SOLAP et des méthodes d'aide à la décision multicritères d'une part, et les techniques de fouille de données spatiales d'autre part, fait toute l'originalité de notre contribution.

Même si des lacunes, présentées en plusieurs interrogations dans cet article, demeurent pour atteindre nos objectifs de départ ; nous avons initié dans le présent document un modèle prédictif utilisant l'analyse multicritères pour établir une carte à risque épidémiologique aidant les décideurs en santé publique à prendre les dispositifs nécessaires pour éviter un risque pour la santé.

Dans nos travaux futurs, nous essaierons d'intégrer les techniques d'exploration de données du système EPISOLAP-MINING.

Cet article est un premier noyau pour un nouveau système d'aide à la décision spatiale pour la surveillance épidémiologique EPISOLAP-MINING.Il mérite d'être approfondi pour être plus complet. L'approche peut se généraliser à divers autres cas d'analyse spatiale impliquant une problématique de décision multicritères.

√

Références

A.Bensaid, M. Barki, O. T. K. B. A. M. (2007). L'analyse multicritère comme outil d'aide à la décision pour la localisation spatiale des zones à forte pression anthropique: le cas du département de n'aama en algérie. *Revue Télédétection 7*, 679–696.

Brans, J. et Vincke (1985). A preference ranking organization method: the promethee method. *Management Science 31*, 647-656.

C.Boulemia, G. e. O. (2006). Eléments de proposition à la mise en place d'une base de données urbaines dans une collectivité locale. *dans les actes du Congrès de l'AUGC*, 679–696.

D. Hamdadou, K. L. e. B. B. (2007). Proposal for a decision-making process in regional planning : Gis, multicriterion approach, choquet's integral and genetic algorithms. *ERIMA07, European Research on Innovation Management Alliance, Biarritz, France*, 51–59.

D.Hamdadou (2008).
. Thèse de doctorat, Université d'Oran, Algérie.

D.Hamdadou, K.Bouamrane, A. M. (2014). Un système interactif d'aide multicritères à la décision pour le diagnostic industriel. *International Conference on engineering of Industrial Safety and Environment "ICISE'14" 4*, 3–4.

F.Joerin (1997). *Décider sur le territoire : proposition d ?une approche par utilisation de SIG et de méthodes d'analyse multicritère.* Thèse de doctorat, Université de Lausanne.

Hamdadou, D. et T. Libourel (2009). Couplage approche multicritère et négociation pour l'aide à la décision en aménagement du territoire. *SAGEO 2009, Spatial Analysis and GEOmatics–Colloque de Géomatique et d'Information Spatiale*.

K. Bouamrane, D. Hamdadou, K. Y. K. G. (2012). Un système interactif d'aide multicritères à la décision pour le diagnostic industriel. *Towards a decision support system, application : itinerary road modification for road transport of hazardous materials.Int.J.Decision Sciences, Risk and Management 4*, 3–4.

M.A. Hamadouche, M.Khaladi, A. Y. D. M. Sig, télédétection et analyse multicritère : vers un outil de gestion et de préservation de la biodiversité du parc national de l'ahaggar (algérie). *la sixième session du Congrès International*, 647–656.

Mottier, V. (1999). *Ún composant logiciel pour les systèmes informatisés de gestion des réseaux d'assainissements.* Thèse de doctorat, Ecole polytechnique fédérale de Lausanne, Suisse.

ö M. Bouziani (2000). *Les pathologies infectieuses : Aspects épidémiologiques et prophylactique.*

öAgadir, F., Ali Pacha, S., Anane, T.,Baough, L. Benkara, A.,Boulahbal,F., Caulet, P. Halassa, S.A., Hannoun. D., Haouichat, L'Hadj., M.,Larbaoui, D., Nafti, S., Ougani, D., Yala et D., Zidouni,(2011). *Manuel de la lutte antituberculeuse à l'usage des personnels médicaux.*

Vanina, G. (1997). *Ćombiner analyse spatiale et épidémiologie pour l'aide à la décision dans la lutte contre la tuberculose en Guyane française.* Thèse de doctorat, l'université d'Orléans. Disipline : Environnement et Santé.

Zemri, F.A., H. D. B. K. (2012). Towards a spatio-temporal interactive decision support system for epidemiological monitoring. coupling solap and datawarehouse. *procedding du Colloque sur l'Optimisation et les Systèmes d'Information*.

Summary

The present study aims to integrate Multi Criteria Analysis Methods (MCAM) to a decision support system based on SOLAP technology, modeled and implemented in other work. The current research evaluates on the one part the benefits of SOLAP in detection and location of epidemics outbreaks and discovers on another part the advantages of multi criteria analysis methods in the assessment of health risk threatening the populations in the presence of the risk (presence of infectious cases) and the vulnerability of the population (density, socio-economic level, Habitat Type, climate...) all that, in one coherent and transparent integrated decision-making platform. We seek to provide further explanation of the real factors responsible for the spread of epidemics and its emergence or reemergence. In the end, our study will lead to the automatic generation of a risk map which gives a classification of epidemics outbreaks to facilitate intervention in order of priority.

Index

A

Arenas, Helbert 9
Atanassova, Iana 21

B

Bertin, Marc 21
Braud, Agnès 45

C

Cruz, Christophe 9

D

Dennouni, Nassim 3
Devogele, Thomas 1
Dolques, Xavier 45

G

Gashteovski, Kiril 33
Gomes Da Silva,Alzennyr 33
Guyet, Thomas 33

H

Hamdadou, Djamila 57
Harbelot, Benjamin 9
Huchard,Marianne 45

K

Kauppinen, Tomi 21

L

Lancieri, Luigi 3
Le Ber,Florence 45

M

Masson,Véronique 33

N

Nica, Cristina 45

P

Peter, Yvan 3

Q

Quiniou, René 33

S

Slama, Zohra 3

Z

Zeitouni, Karine 57
Zemri, Farah Amina 57