Extraction d'arbres spatio-temporels d'itemsets pour le suivi environnemental

Jérémy Sanhes*, Frédéric Flouvat*, Nazha Selmaoui-Folcher*, Jean-François Boulicaut**

*University of New Caledonia, PPME, BP R4, F-98851 Nouméa, New Caledonia
{jeremy.sanhes, nazha.selmaoui, frederic.flouvat}@univ-nc.nc
**INSA of Lyon, LIRIS, BP R4, F-98851 Lyon, France
jean-francois.boulicaut@insa-lyon.fr

Un nombre croissant de données spatio-temporelles ont été collectées pour étudier des phénomènes naturels (p.ex. risques naturels, changements environnementaux, propagation de maladies infectieuses). Extraire des connaissances pour mieux comprendre la dynamique de propagation de tels phénomènes est une tâche difficile. Par exemple, l'érosion des sols est caractérisée par un ensemble de facteurs interagissants, causant la propagation ou l'apparition de surfaces érodées. Il est important de savoir comment et quels facteurs ont un effet sur ce phénomène. Même si l'influence des facteurs environnementaux (présence de mines, type de sol, météo, feux de forêts, etc.) est connue, l'impact de ces facteurs avec leurs interactions dans l'espace et dans le temps est un problème ouvert. La fouille de données spatio-temporelles vise à proposer des solutions pour mieux comprendre et décrire ces phénomènes complexes. Les travaux existants (Yao, 2003) utilisent typiquement des motifs (p.ex. des séquences, des arbres, ou des graphes) pour modéliser la dynamique de ce type de phénomènes. Ces travaux font principalement face à deux problématiques : les trajectoires d'objets en mouvement (Yao, 2003) et les séquences d'événements (Tsoukatos et Gunopulos, 2001; Wang et al., 2004; Huang et al., 2008; Celik et al., 2008; Mohan et al., 2010). Toutefois, ces travaux ne permettent pas d'étudier les dynamiques de propagation d'un phénomène en fonction de son environnement.

Face à ces limites, nous proposons d'utiliser des arbres spatio-temporels d'itemsets afin de représenter les dynamiques de propagation de phénomènes naturels en fonction de leur environnement proche. Soit une base de données spatio-temporelles composée d'un ensemble de n-uplets, où chaque n-uplet représente l'ensemble des caractéristiques, appelé itemset, d'une zone à un temps donné (p.ex. pluviométrie, bâtiments, état des sols). Un **arbre spatio-temporel d'itemsets** $T = (V, E)$ est un arbre où V est un ensemble d'itemsets et E est un ensemble d'arêtes $(X_t, X_{t'})$, avec $X_t, X_{t'} \in V$ situés dans des zones voisines (par rapport à une relation d'adjacence $\Re_n$) et à des temps successifs.

Nous introduisons aussi une méthode visant à extraire les propagations les plus fréquentes, tout en tirant partie du grand nombre de travaux réalisés sur la fouille d'arbres. La première étape de cette méthode consiste à capturer la dynamique du phénomène à étudier. En effet, dans beaucoup d'applications (études de risques naturels, changements environnementaux, propagation de maladies infectieuses), le phénomène et sa dynamique de propagation ne sont pas clairement identifiés (contrairement aux applications telles que la fouille de trajectoires). Dans cet objectif, nous adaptons l'approche proposée dans Mabit et al. (2011) afin de construire in-

crémentalement un ensemble d'arbres d'itemsets représentant les dynamiques de propagation du phénomène.

La deuxième étape consiste à exploiter les algorithmes existants d'extraction d'arbres fréquents. Cependant, ces algorithmes ne permettent pas de traiter des arbres d'itemsets, mais uniquement des arbres avec une seule caractéristique par noeud. De plus, ces algorithmes nécessitent d'avoir un prédicat anti-monotone, ce qui n'est pas le cas pour les arbres d'itemsets. Nous transformons donc les arbres d'itemsets afin qu'il soient "compatibles" avec les algorithmes classiques de fouille d'arbres. Nous avons plus particulièrement identifié trois stratégies (duplication, décomposition et fusion). Nous avons également intégré des contraintes (spatiales et temporelles) du domaine pour filtrer les motifs les plus intéressants pour les experts, et améliorer les performances de l'extraction en réduisant le nombre de motifs étudiés.

Ce travail a été appliqué au suivi de l'érosion dans une zone montagneuse ainsi qu'à l'étude d'une épidémie de maladie infectieuse. Les résultats préliminaires montrent l'intérêt de cette approche.

Remerciements. Ces travaux ont été en partie financés par le contrat français ANR-2010-COSI-012 FOSTER.

Références

Celik, M., S. Shekhar, J. P. Rogers, et J. A. Shine (2008). Mixed-drove spatiotemporal co-occurrence pattern mining. *IEEE Trans. Knowl. Data Eng. 20*(10), 1322–1335.

Huang, Y., L. Zhang, et P. Zhang (2008). A framework for mining sequential patterns from spatio-temporal event data sets. *IEEE Trans. Knowl. Data Eng. 20*(4), 433–448.

Mabit, L., N. Selmaoui-Folcher, et F. Flouvat (2011). Modélisation de la dynamique de phénomènes spatio-temporels par des séquences de motifs. In *EGC*, pp. 455–466.

Mohan, P., S. Shekhar, J. A. Shine, et J. P. Rogers (2010). Cascading spatio-temporal pattern discovery : A summary of results. In *SDM*, pp. 327–338.

Tsoukatos, I. et D. Gunopulos (2001). Efficient mining of spatiotemporal patterns. *Advances in Spatial and Temporal Databases*, 425–442.

Wang, J., W. Hsu, M. Lee, et J. Wang (2004). FlowMiner : finding flow patterns in spatio-temporal databases. *ICTAI*, 14–21.

Yao, X. (2003). Research issues in spatio-temporal data mining. In *White paper UCGIS*.

Summary

These last years an increasing amount of spatio-temporal data has been collected to study complex natural phenomena (e.g. natural hazards, environmental change, spread of infectious diseases). Extracting knowledge to better understand the dynamic of these phenomena is a challenging task. In this paper, we define a new type of pattern, called complex spatio-temporal tree, to better capture the spatio-temporal properties of natural phenomena. We also propose a method to extract these complex patterns using a "classical" tree mining algorithm. Our approach has been applied to a real dataset dealing with soil erosion monitoring.