PROPOSITION DE SUJET DE THESE

Campagne 2018

Laboratoire L3i, Université de La Rochelle

Sujet de la thèse :

Analyse de séquences – Méthodes génériques combinant analyse formelle des concepts et théorie des patrons ; Etude expérimentale en analyse de trajectoires et fouille de texte.

Résumé du travail proposé :

La bibliothèque Galactic développée au sein du laboratoire L3i fournit des méthodes efficaces d'analyse de données tabulaires classiques. L'objectif de cette thèse sera d'en implémenter une extension générique pour des données complexes sur des bases théoriques établies, puis de l'instancier pour des données de type séquences. Il s'agira ensuite de mener une étude expérimentale de nouvelles méthodes d'analyse de séquences pour les deux cas d'usages que sont l'analyse de trajectoires et de la fouille de textes.

Mots clés :

Fouille de séquences ; Analyse Formelle des Concepts, structure de patrons

Informations complémentaires :

Encadrant(s) :

- Karell Bertet (directeur de thèse) <kbertet@univ-lr.fr>
- Christophe Demko <cdemko@univ-lr.fr>

Equipe Laboratoire L3i:

- ☐ **Images et Contenus**
- ☐ **Dynamique des systèmes et adaptativité**
- ☒ **Modèle et Connaissance**

Domaine d'application stratégique :

- ☐ **E-éducation**
- ☒ **Environnement et développement durable**
- ☒ **E-culture**
- ☒ **Valorisation de contenus numériques**

Date de début du contrat : Septembre 2018

Durée du contrat : 3 ans

Contexte de l'étude:

L'analyse de données complexes, et en particulier de données séquentielles, correspond à l'un des problèmes phares actuels en data mining, l'objectif principal étant d'extraire des sous-motifs caractéristiques dans des séquences de symboles. Ainsi, en fouille de texte, l'extraction de sous-séquences de termes fréquents et caractéristiques décrivant un document est une alternative pertinente à l'approche fréquentielle classique TF-IDF qui n'intègre pas cette notion de regroupements de termes [5]. L'approche séquence est également une solution classique à l'analyse de trajectoires, problématique importante du laboratoire L3i actuellement qui s'intéresse à des trajectoires de déplacements virtuels que représentent les traces d'utilisateurs sur le web, de déplacements dans le monde réel effectuées par des visiteurs d'un musée, ou encore de déplacements d'animaux dans leur environnement naturel. Une approche classique consiste alors à représenter une trajectoire sous la forme d'une séquence d'états.

Dans le domaine de l'analyse formelle de concepts (AFC) [2], de nombreux algorithmes sont proposés pour extraire des motifs représentatifs qui peuvent s'organiser en concepts ou sous forme de règles d'association, à partir de données décrites par des ensembles d'attributs. Ces algorithmes, qui reposent sur des fondamentaux algébriques, permettent de pallier le problème de la sur-représentativité ou encore de la redondance des informations extraites en garantissant des propriétés de minimalité et de canonicité, on parle alors de générateurs minimaux ou encore de bases minimales de règles. La théorie des patrons [3] permet d'étendre ces algorithmes à des données complexes sous certaines conditions mathématiques, conditions vérifiées en particulier par les séquences. Cependant, les outils actuels intégrant l'AFC et la théorie des patrons ont été développés dans un cadre applicatif dédié, et ne peuvent être utilisés de façon générique.

Dans le cadre du projet Galactic du laboratoire L3i, nous avons développé une bibliothèque qui implémente les algorithmes de l'AFC en mettant en avant la notion d'opérateur de fermeture [1]. Cette bibliothèque a été conçue de façon générique, et s'adapte particulièrement bien à une extension de la théorie des patrons pour l'intégration de données complexes, l'objectif étant de fournir aux chercheurs de la communauté les outils nécessaires à la mise en place de nouvelles méthodes d'analyse de données complexes.

Description du sujet :

L'objectif général de cette thèse sera de mettre en place la version 2 de la bibliothèque Galactic pour des données complexes, et en particulier des séquences. Puis de proposer des méthodes d'analyse de séquences dans les deux cadres applicatifs que sont l'analyse de trajectoires et la fouille de texte.

Le premier objectif de cette thèse sera de finaliser l'intégration générique de données complexes à la bibliothèque Galactic selon la théorie des patrons, travaux amorcés récemment. Il s'agira ensuite de mettre en place des données de type séquence et graphe, c'est-à-dire d'en fournir une classe descriptive, ainsi que les opérateurs nécessaires à leur manipulation par les algorithmes de l'AFC, à savoir les opérations « sous-séquence » et « plus grandes sous-séquences communes » pour des séquences, et les opérations de « sous-graphe » et « plus grands sous-graphes communs » pour des graphes.

En ce qui concerne les séquences, les algorithmes existants dans la bibliothèque Galactic permettront ainsi d'extraire des sous-séquences minimales et génératrices, ou encore des implications minimales et canoniques entre des sous-séquences en prémisses et en conclusion. Le second objectif sera de proposer de nouvelles méthodes d'analyse de séquences dans deux contextes applicatifs :

- Analyse de trajectoires de déplacements, le but sera d'en générer des résumés sous forme de tableau de bord. L'extraction de bases canoniques d'implications entre des sous-séquences serait une solution pertinente dans ce contexte. Les données considérées seront celles obtenues dans le cadre du projet régional déposé : « Dispositif d'Analyse des

Traces numériques pour la valorisation des Territoires Touristiques (DA3T) », un travail en collaboration avec une thèse dans le cadre de ce projet est envisagé

- Représentation de textes par des regroupements minimaux de sous-termes consécutifs. Dans le domaine de la fouille de textes, une problématique majeure consiste à extraire des descripteurs représentatifs des documents dans un objectif d'indexation ou de classification. Ces descripteurs sont établis sur la base d'un ensemble de termes représentatifs des documents, obtenus par des prétraitements de Traitement Automatique des Langues. Des filtrages par des mesures telles que le TF-IDF permettent de sélectionner les termes les plus représentatifs. Lorsque l'on cherche à obtenir des groupes de termes consécutifs, cette problématique du filtrage est plus grande encore. L'utilisation de méthodes qui ne soient pas uniquement fréquentielles est une des solutions envisagées actuellement. Les générateurs minimaux de sous-séquences seraient une alternative pertinente.

Une migration de la bibliothèque Galactic en Python est envisagée, à la fois pour améliorer les temps de calculs, et pour permettre une plus grande généricité :

- Généricité des données pour des données complexes via la théorie des patrons
- Généricité des stratégies d'exploration des données complexes, la notion de sous-séquence pouvant par exemple se dériver selon plusieurs stratégies selon que l'on considère des sous-séquences strictes ou non.
- Généricité des données selon qu'elles soient stockées en mémoire, dans une base de données, ou accessibles via un flux de données.

Un troisième et dernier objectif de cette thèse sera de participer à cette migration, pour laquelle un support ingénierie est envisagé par un projet SAAT en cours de dépôt.

Enfin, cette thèse s'inscrit dans une volonté de fournir aux membres du laboratoire des outils pour mettre en place des méthodes efficaces d'analyses de données complexes reposant sur la théorie des patrons et l'analyse formelle des concepts. Pour cela, les outils permettant de manipuler des données séquences et graphes seront mis en place, et un accompagnement pédagogique est envisagé sous forme de groupes de travail internes au laboratoire.

Prérequis et contraintes particulières :

Formation informatique et algorithmique. De bonnes connaissances en mathématiques sont recommandées.

Références bibliographiques :

[1] K. Bertet, C. Demko, J.-F. Viaud, C. Guérin, Lattices, closures systems and implication bases: A survey of structural aspects and algorithms, Theoretical Computer Science, 2016, ISSN 0304-3975, http://dx.doi.org/10.1016/j.tcs.2016.11.021.

[2] B. Ganter and R. Wille. Formal Concept Analysis, Mathematical foundations. Springer Verlag, Berlin, 1999.

[3] B. Ganter and S.O. Kuznetsov. Pattern structures and their projections. In LNCS of International Conference on Conceptual Structures (ICCS'01), pages 129-142, 2001.

[4] Multilingual Event Extraction for Early Epidemic Detection G. Lejeune, R. Brixtel, A. Doucet et N. Lucas. Artificial Intelligence in Medicine p. 131-143, 2015