Sujet de thèse

Comparaison et coopération d'approches en analyse de concepts formels pour les données relationnelles

Laboratoire d'accueil : Laboratoire ICube, équipe SDC, Strasbourg, en collaboration avec IRISA, équipe LACODAM, Rennes

Direction de la thèse : Florence Le Ber (ICube), Sébastien Ferré (IRISA)

Encadrants : Xavier Dolques (ICube), Peggy Cellier (IRISA)

Contact : envoyer CV, relevés de notes et lettre de motivation à florence.leber@icube.unistra.fr Sebastien.Ferre@irisa.fr

Financement : ANR SmartFCA

Contexte : Dans les données disponibles pour l'analyse, beaucoup ont un caractère relationnel : données spatiales, temporelles, ou décrivant des liens entre individus. Les méthodes traditionnelles ne sont pas adaptées à ce type de données, qui nécessitent des approches spécifiques, incluant des techniques d'agrégation. Parmi ces approches, l'*analyse relationnelle de concepts et l'analyse conceptuelle de graphes* sont dérivées de l'analyse de concepts formels (ACF) [1], qui est une méthode mathématique de classification, largement appliquée sur différents types de données et dans de nombreux domaines (par exemple [2,3]). Elle consiste, à partir d'une table (appelée contexte) décrivant des objets par des attributs, à construire un treillis de concepts, i.e. des couples (extension ; intension) d'ensembles fermés décrivant les objets et les attributs qui les définissent.

L'analyse relationnelle de concepts (ARC) [4] considère deux types de contextes, des contextes objets-attributs et des contextes objets-objets décrivant les relations entre objets. L'ARC étend les contextes objets-attributs par des attributs relationnels de la forme qrC, où q est un quantificateur, r une relation et C un concept issu du co-domaine de r. Le résultat de l'ARC est une famille de treillis (un par contexte objets-attributs) reliés entre eux par ces attributs relationnels : un concept d'un treillis représente un groupe d'objets caractérisé par des attributs simples et des attributs relationnels renvoyant à des concepts d'un autre treillis.

L'analyse conceptuelle de graphes (Graph-FCA) [5] a pour contextes des hypergraphes où les nœuds sont les objets et où les hyperarcs sont étiquetés par des attributs. Un hyper-arc unaire a(o) correspond à la description d'un objet par un attribut, comme dans l'ACF. Un hyper-arc binaire a(o1,o2) correspond à une relation 'a' de o1 vers o2, comme les attributs relationnels dans RCA. Les relations n-aires sont représentées par des hyperarcs n-aires a(o1,…,oN). Un concept de graphe représente un ensemble de tuples d'objets (extension) qui peuvent être vus comme les réponses exhaustives à une requête conjonctive (intension), par exemple (x,y) ← a1(x,z), a2(y,z), et où cette requête exprime tout ce que ces tuples ont en commun.

Cette thèse s'inscrit dans le cadre de l'ANR SmartFCA, qui regroupe 5 équipes françaises travaillant dans le domaine de l'ACF et dont l'objectif est de mettre à disposition une plateforme rassemblant les différentes variantes de cette méthode. Plusieurs ingénieurs seront affectés au développement de cette plateforme.

Objectifs de la thèse : Cette thèse a pour but de mener une comparaison théorique et expérimentale des deux approches ARC et Graph-FCA, de proposer des éléments pour faire coopérer les deux approches, et de définir un guide méthodologique d'usage (modélisation des données, valeurs des paramètres, choix des algorithmes, etc.). Les résultats, algorithmes et guide méthodologique, seront intégrés dans la plateforme développée dans le cadre du projet ANR SmartFCA.

Les liens entre les deux approches ont déjà été abordés [6,7,8] et la thèse doit approfondir ces travaux. Il s'agira dans un premier temps d'étudier et de comparer les deux approches, à partir des outils existants, en les testant sur des jeux de données relationnels fournis par les partenaires du projet. On s'intéressera en

particulier à proposer un modèle déclaratif de l'ARC qui est actuellement définie de manière itérative. On s'intéressera aussi à la coopération entre l'ARC et Graph-FCA par la définition des structures de données permettant de les rendre interopérables.

Le caractère explosif des approches fondées sur l'ACF conduit à utiliser des algorithmes ne calculant qu'une sous-partie des concepts ou des treillis : AOC-poset [9], approches exploratoires, calcul de voisinages, estimation des résultats à partir du choix des paramètres [10,11] ... Ces variantes seront aussi étudiées et permettront de définir un cadre méthodologique d'utilisation de l'ARC et de Graph-FCA incluant ces différentes options ainsi que des éléments pour guider leur usage. Le travail sera mené en coopération avec un ingénieur chargé des développements dans la plateforme.

Apports attendus :

- Avancées théoriques sur les méthodes ACF
- Développements méthodologiques
- Expérimentations et validation sur des données réelles

Profil recherché :

- Master 2 en Informatique ou équivalent
- Formation en logique, représentation de connaissances et programmation
- Curiosité, capacité à appréhender différents domaines et à interagir avec les experts de ces domaines

Références :

[1] Ganter, B., Wille, R. Formal concept analysis - mathematical foundations. Springer (1999)

[2] Priss, U. Formal concept analysis in information science. ARIST 40(1), 521–543 (2006)

[3] Alam, M., Coulet, A., Napoli, A., Smaïl-Tabbone, M. Formal Concept Analysis Applied to Transcriptomic Data. CLA 2012, Málaga, Spain

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

[5] S. Ferré and P. Cellier. Graph-FCA: An extension of formal concept analysis to knowledge graphs. Discrete and Applied Mathematics, 273:81–102 (2020).

[6] C. Nica, A. Braud, and F. Le Ber. RCA-Seq: an Original Approach for Enhancing the Analysis of Sequential Data Based on Hierarchies of Multilevel Closed Partially-Ordered Patterns. Discrete Applied Mathematics, 273:232–251, 2020.

[7] S Ferré, P Cellier. How Hierarchies of Concept Graphs Can Facilitate the Interpretation of RCA Lattices? CLA 2018.

[8] P. Keip, S. Ferré, A. Gutierrez, M. Huchard, P. Silvie, P. Martin. Practical Comparison of FCA Extensions to Model Indeterminate Value of Ternary Data. CLA 2020, 197-208

[9] X. Dolques, F. Le Ber, M. Huchard , C. Grac. Performance-friendly rule extraction in large water data-sets with AOC posets and relational concept analysis, International Journal of General Systems, Taylor & Francis (2016)

[10] Braud, A., Dolques, X., Huchard, M., Le Ber, F. Generalization effect of quantifiers in a classification based on relational concept analysis. Knowledge-Based Systems 160, 119–135 (2018)

[11] A. Ouzerdine , A. Braud, X. Dolques, M. Huchard , F. Le Ber. Adjusting the exploration flow in Relational Concept Analysis -- An experience on a watercourse quality dataset, Advances in Knowledge Discovery and Management, Springer (2022).