Résumé de la thèse “Mining Tractable Sets of Graph Patterns with the Minimum Description Length Principle”

Francesco Bariatti*

*LIACS, Leiden University, Leiden, The Netherlands
f.bariatti@liacs.leidenuniv.nl

Dans nombreux domaines il est courant de trouver des données structurées sous la forme d'un ensemble d'entités reliées entre elles par des relations. Par exemple, en chimie et en biologie, les molécules peuvent être exprimées comme des atomes reliés par des liaisons ; en linguistique, les phrases peuvent être exprimées comme des mots reliés par des relations de dépendance ; dans le web sémantique, les connaissances peuvent être exprimées sous forme d'entités nommées reliées par des relations sémantiques. Ces données sont représentées sous forme de graphes : des structures de données où les “sommets” (les entités) sont interconnectés par des “arêtes” (les relations). Les sommets et arêtes peuvent aussi être étiquetés, afin de préciser les attributs des entités et relations correspondantes.

Ces données peuvent révéler de la connaissance utile à l'utilisateur, cependant leur analyse par un humain devient de plus en plus difficile à mesure que la taille du jeu de données augmente. En pratique, il n'est pas rare de trouver des jeux de données dont les graphes comportent des millions ou des milliards de sommets reliés par autant d'arêtes. Afin d'aider les utilisateurs, des approches automatisées sont nécessaires pour rendre les données plus faciles à traiter. Les approches de *fouille de motifs* aident l'utilisateur à s'attaquer à cette tâche en extrayant des structures locales à partir des données. En particulier, de nombreuses approches ont été proposées pour traiter les données de type graphe. Cependant, un problème courant est l'*explosion du nombre de motifs* : même sur des petits jeux de données, les approches classiques de fouille génèrent de très grandes quantités de motifs (des millions ou des milliards). Dans ce cas, la fouille de motifs n'est d'aucune utilité pour l'utilisateur, car l'analyse de la grande quantité de motifs extraits devient une tâche aussi difficile que celle de l'analyse des données initiales. Afin de réduire le nombre de motifs extraits, plusieurs techniques ont été proposées, comme l'utilisation de représentations condensées pour réduire le nombre de motifs affichés à l'utilisateur ; l'intégration de contraintes dans le processus de fouille ; et l'échantillonnage aléatoire de l'espace des motifs. Ces méthodes permettent souvent de réduire le nombre de motifs extraits de plusieurs ordres de grandeur, mais cela n'est souvent pas assez pour que les motifs puissent être analysés par un utilisateur humain (des centaines de milliers de motifs peuvent encore rester).

Plus récemment, des approches ont été proposées qui utilisent le principe *Minimum Description Length* (MDL) pour générer et sélectionner des ensembles de motifs suffisamment *petits* pour permettre une analyse humaine et suffisamment *descriptifs* des données pour permettre d'en extraire de la connaissance significative. Le principe MDL provient du domaine de la théorie de l'information et est souvent résumé par la formule suivante : “le modèle qui décrit le mieux les données est celui qui les compresse le plus”, ce qui signifie qu'un modèle

adapté aux données devrait permettre de les décrire avec une quantité minimale d'informations par rapport à un modèle qui n'est pas adapté. Le principe MDL a été appliqué au problème de la sélection de motifs en traitant les ensembles de motifs comme des "modèles" qui sont utilisés pour encoder les données. Les approches basées sur le principe MDL ont montré leur efficacité sur de nombreux types de données : données transactionnelles, bases de données relationnelles, séquences, matrices, etc. Peu d'approches MDL existent pour les graphes et elles imposent généralement des limites sur le type de motifs extraits.

Dans cette thèse, nous proposons des approches qui utilisent le principe MDL afin de générer et sélectionner des *petits ensembles* de motifs *descriptifs* de type graphe à partir de données de type graphe, afin d'aider les analystes humains à extraire de la connaissance significative des données. La fouille de motifs dans les graphes présente non seulement les défis habituels de la fouille de motifs —tels qu'un grand espace de recherche qui nécessite une stratégie d'exploration efficace— mais présente également des défis spécifiques dus à la nature des graphes. En premier lieu, détecter les occurrences d'un motifs dans les données est un problème NP-complet. Deuxièmement, les données de type graphe ont une composante structurelle importante. Savoir qu'un motif est *présent* dans les données n'est pas suffisant. Savoir *comment le motif se connecte au reste des données* est une information importante qui révèle également de la connaissance sur les données, et qui doit pouvoir être communiquée à l'utilisateur. Dans cette thèse, nous instancions le principe MDL dans un contexte de fouille de motifs de graphes. Nous proposons des mesures basées sur MDL pour évaluer des ensembles de motifs, sans imposer des limites sur la forme de ces derniers. Nous introduisons la notion de *ports*, qui permet de décrire les données de type graphe comme une composition d'occurrences de motifs de type graphe sans aucune perte d'information, ce qui est fondamental dans les approches MDL. De plus, nous montrons que cette notion met en valeur les interactions entre différents motifs. Nous proposons des approches utilisant ces notions pour extraire un petit ensemble *de taille humaine* de motifs descriptifs à partir de données de type graphe. Pour chacune de ces approches nous proposons des algorithmes heuristiques, permettant de produire des résultats en un temps raisonnable, et ne demandant pas un paramétrage extensif par l'utilisateur. Nous évaluons toutes nos contributions expérimentalement sur des jeux de données provenant de différents domaines, y compris du web sémantique. Nous proposons aussi un outil pour *visualiser interactivement* les résultats de nos approches, permettant à l'utilisateur de les manipuler afin de mieux les comprendre.