

Evaluation de modèles mathématiques pour le problème de la distance d'édition entre graphes

Mostafa Darwiche^{1,2}, Donatello Conte¹, Romain Raveaux¹, Vincent T'kindt²

¹ Laboratoire d'Informatique, Université François Rabelais Tours, France
`{mostafa.darwiche,romain.raveaux,donatello.conte}@univ-tours.fr`

² Laboratoire d'Informatique, ERL-CNRS 6305, Université François Rabelais Tours, France
`tkindt@univ-tours.fr`

Mots-clés : *problème de la distance d'édition, appariement de graphes, programmation linéaire en nombres entiers.*

1 Introduction

Les représentations structurelles et surtout celles qui se basent sur les graphes deviennent de plus en plus populaires dans plusieurs domaines connexes à la Recherche Opérationnelle (RO) comme la vision par ordinateur, la chimie, la biologie. En effet, les graphes sont capables de représenter des objets tels qu'une image, un texte ou une molécule chimique ([5], [7]). Les nœuds et les arêtes des graphes peuvent posséder un ou plusieurs attributs (des valeurs numériques, catégories, ...). L'étude des corpus de graphes permet de construire des regroupements de graphes similaires (clustering) ou alors d'associer des graphes à des modèles de références (classification). Dans les deux cas, une mesure de (dis)similarité entre graphes est indispensable. Le problème de la distance d'édition entre graphes (DEG) fait partie des problèmes d'appariements de graphes tolérants à l'erreur et fournit une mesure de dissimilarité entre deux graphes ([6]). L'idée de base est de transformer un graphe G en un graphe G' en effectuant un ensemble d'opérations d'édition. Ces opérations sont des substitutions, suppressions, insertions de nœuds ou d'arêtes et un coût est associé à chaque opération. Le problème est de trouver un ensemble d'opérations dont le coût total est minimum. Le problème DEG est un problème d'optimisation NP-difficile ([8]). Depuis quelques années, des chercheurs se sont intéressés à l'utilisation d'outils de la RO pour résoudre ce problème. Dans la littérature, plusieurs formulations mathématiques modélisent le problème DEG. Un modèle quadratique a été proposé dans ([2]) ainsi que deux modèles linéaires en nombres entiers dans ([3], [4]). Dans ce travail, une évaluation de performance des modèles linéaires en nombres entiers est présentée tant dans le cadre d'algorithmes exacts que d'heuristiques. Cela nous permet d'analyser le comportement de ces modèles et d'essayer de les améliorer.

2 Modèles mathématiques pour le problème DEG

Soient deux graphes G comme source et G' comme cible. La distance d'édition $d_{\lambda_{min}}(G, G')$, entre G et G' est définie par :

$$d_{\lambda_{min}}(G, G') = \min_{\lambda \in \gamma(G, G')} \sum_{e_i \in \lambda} c(e_i) \quad (1)$$

où $\gamma(G, G')$ est l'ensemble de tous les ensembles d'opérations d'édition possibles pour transformer G en G' . Pour un ensemble $\lambda = \{e_1, \dots, e_n\} \in \gamma(G, G')$, $c(e_i)$ est le coût de l'opération e_i .

Le modèle de Justice et Hero ([3]) repose sur l'idée de trouver une permutation qui minimise la norme L_1 de la différence entre les matrices d'adjacence de G et G' .

Le second modèle étudié, proposé par Lerouge et al. ([4]), utilise une formulation plus standard en RO. Ainsi la modélisation du problème repose sur des variables booléennes x_{ij} valant 1 si le nœud i du graphe G est apparié avec le nœud j du graphe G' . De la même manière, des variables booléennes sont introduites pour l'appariement des arêtes. Les opérations de suppression et d'insertion sont représentées à l'aide d'appariements à un nœud fictif ϵ . L'évaluation des deux modèles est réalisée sur la base de données Mutagenicity([1]) qui représente des molécules chimiques et qui contient 7 sous-ensembles de graphes non orientés. Il y a 10 graphes possédant le même nombre de nœuds dans chaque sous-ensemble (de 10 jusqu'à 70 nœuds). Pour chaque paire de graphes (en excluant les paires de graphes identiques), le problème est résolu, par les deux modèles. Les conditions d'arrêt sont : temps limite de 900 seconds et 2 GB de mémoire RAM. Les modèles sont résolus par le solveur CPLEX 12.6.0.

Nb. de nœuds	10		20		30		40		50		60		70	
	CPU	Opt	CPU	Opt	CPU	Opt	CPU	Opt	CPU	Opt	CPU	Opt	CPU	Opt
Justice et Hero	0.14	90	1.12	90	156.71	83	275.28	77	501.47	57	803.96	15	828.59	8
Lerouge et al.	0.20	90	29.69	90	823.46	12	874.43	4	900.08	0	900.11	0	900.14	0

TAB. 1 – Résultats d'évaluation des modèles. Le CPU est calculé en secondes, Opt est le nombre d'instances résolues à l'optimalité.

D'après le tableau 1, il est évident que le modèle de Justice et Hero est plus rapide en moyenne que le modèle de Lerouge et al. D'autre part, sauf pour les deux premiers sous-ensembles (10 et 20 nœuds), le second modèle résout moins d'instances à l'optimalité que le premier modèle. Lors de la conférence nous présenterons une analyse détaillée de ces modèles. Nous présenterons également des résultats concernant l'utilisation de techniques de fixation de variables et de leur apport potentiel pour la résolution exacte du problème.

Références

- [1] Zeina Abu-Aisheh, Romain Raveaux, and Jean-Yves Ramel. A graph database repository and performance evaluation metrics for graph edit distance. In *Graph-Based Representations in Pattern Recognition - 10th IAPR-TC-15.Proceedings*, pages 138–147, 2015.
- [2] Sébastien Bougleux, Luc Brun, Vincenzo Carletti, Pasquale Foggia, Benoît Gaüzère, and Mario Vento. Graph edit distance as a quadratic assignment problem. *Pattern Recognition Letters*, pages –, 2016.
- [3] Derek Justice and Alfred Hero. A binary linear programming formulation of the graph edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(8) :1200–1214, Aug 2006.
- [4] Julien Lerouge, Zeina Abu-Aisheh, Romain Raveaux, Pierre Héroux, and Sébastien Adam. *Exact Graph Edit Distance Computation Using a Binary Linear Program*, pages 485–495. Springer International Publishing, Cham, 2016.
- [5] Jianzhuang Liu and Yong Tsui Lee. Graph-based method for face identification from a single 2d line drawing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10) :1106–1119, Oct 2001.
- [6] Kaspar Riesen. Graph edit distance. In *Structural Pattern Recognition with Graph Edit Distance*, pages 29–44. Springer International Publishing, 2015.
- [7] Peter Willett, John M. Barnard, and Geoffrey M. Downs. Chemical similarity searching. *Journal of Chemical Information and Computer Sciences*, 38(6) :983–996, 1998.
- [8] Zhiping Zeng, Anthony KH Tung, Jianyong Wang, Jianhua Feng, and Lizhu Zhou. Comparing stars : on approximating graph edit distance. *Proceedings of the VLDB Endowment*, 2(1) :25–36, 2009.