UNIVERSITÉ DE LORRAINE

Stage de Master2 Recherche

2023-2024

Identification de communautés sur les réseaux sociaux

1 Encadrant(s)

Encadrant:	Geoffray Bonnin
Equipe et laboratoire:	BIRD, Loria
Contact:	bonnin@loria.fr
Co-encadrant:	Lydia Boudjeloud-Assala
Equipe et laboratoire:	ORPAILLEUR, Loria
Contact:	Lydia.boudjeloud-assala@univ-lorraine.fr

2 Motivation et contexte

L'identification de communautés sur les réseaux sociaux est généralement effectuée selon deux types d'approches. La première consiste en l'exploitation de la similarité entre les individus du réseau social considéré en fonction de leurs caractéristiques (âge, sexe, activité sur un service donné, appréciations musicales, etc.) [WHGD19]. La seconde consiste en l'exploitation du graphe des liens sociaux explicites entre les individus (amis sur Facebook, suivis/suiveurs sur Deezer, etc.) afin d'en extraire des cliques ou des quasi-cliques [MSST07].

L'une des problématiques de ce domaine de recherche est l'évaluation de la pertinence des communautés extraites [EKGB16]. Une solution répandue consiste à vérifier d'une part que les individus au sein de chaque groupe sont fortement similaires (haute similarité intra-cluster), et d'autre part que la similarité entre les individus de groupes différents est faible (faible similarité inter-cluster) [BBA13]. Le problème de cette solution est qu'un score élevé calculé selon ce type de critères ne correspond pas forcément à des communautés pertinentes, et que des communautés extraites très différentes peuvent avoir des scores très proches.

Une manière possible d'amoindrir ce problème serait de croiser deux points de vue différents, chacun correspondant à l'un des deux types d'approches de clustering mentionnées ci-dessus. En effet, ces deux types d'approches utilisant des informations très différentes en entrée, leurs sorties sont souvent très différentes elles aussi. Or, la meilleure version de chacune de ces approches devrait en principe produire des communautés aussi proches des communautés réelles du réseau social considéré que possible, et leurs sorties respectives devraient elles aussi être aussi similaires que possible. La pertinence d'un algorithme correspondant à l'un de ces deux types d'approches pourrait ainsi être évaluée en fonction de la distance entre sa sortie et celle d'un algorithme de l'autre type. L'idée est donc de rechercher une stratégie de dialogue qui, étant donnés deux algorithmes des deux types (similarité et liens sociaux), permette à ces algorithmes de converger vers des sorties aussi similaires que possible.

3 Objectifs

Les objectifs de ce stage porteront ainsi sur la détermination de stratégies de dialogue entre les deux types d'algorithmes et sur la possibilité d'obtenir une convergence. Un bon point de départ est l'article de (Forestier et al., 2010) [FGW10] sur la résolution itérative de conflits entre clusterings.

Nous fournirons au stagiaire une base de données issue du site senscritique, qui permet à ses utilisateurs de rédiger des critiques de films, de livres, de musique, etc., d'attribuer des notes et de suivre d'autres utilisateurs.

Dans un premier temps, le stagiaire devra se familiariser avec d'une part les algorithmes classiques de clustering (K-means, DBscan, etc.) et d'autre part les algorithmes d'extraction de quasi-cliques (Quick, alpha-bêta-cliques, etc.). Des bibliothèques implémentant ces algorithmes seront exploitées dans un second temps pour observer les différences entre les communautés produites en fonction du type d'algorithme et des paramètres choisis (nombre de clusters, densité des clusters, connectivité minimale du voisinage, etc.). Enfin, des stratégies itératives permettant de faire dialoguer les deux types d'approches seront proposées, implémentées et expérimentées.

4 Information suplémentraires

— Stage rémunéré
— Possibilité de poursuite en thèse
— LORIA

Références

[BBA13] Alexandre Blansché and Lydia Boudjeloud-Assala. Processus itératif d'extraction de classes en non supervisée. In *EGC*, pages 9–14, 2013.

[EKGB16] Scott Emmons, Stephen Kobourov, Mike Gallant, and Katy Börner. Analysis of network clustering algorithms and cluster quality metrics at scale. *PloS one*, 11(7) :e0159161, 2016.

[FGW10] Germain Forestier, Pierre Gançarski, and Cédric Wemmert. Collaborative clustering with background knowledge. *Data & Knowledge Engineering*, 69(2) :211–228, 2010.

[MSST07] Nina Mishra, Robert Schreiber, Isabelle Stanton, and Robert E Tarjan. Clustering social networks. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 56–67. Springer, 2007.

[WHGD19] Joyce Jiyoung Whang, Yangyang Hou, David F. Gleich, and Inderjit S. Dhillon. Non-exhaustive, overlapping clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(11) :2644–2659, 2019.