Dynamique des communautés par prédiction d'interactions dans les réseaux sociaux

Blaise Ngonmang[*****], Emmanuel Viennet[*]

[*]Université Paris 13, Sorbonne Paris Cité,
L2TI (EA 3043), F-93430, Villetaneuse, France. firstname.lastname@univ-paris13.fr
[**]UMI 209 UMMISCO, Université de Yaoundé I, B.P. 337 Yaoundé, Cameroun
[***]LIRIMA, Equipe IDASCO, Faculté des Sciences,
Département d'Informatique,
B.P. 812 Yaoundé, Cameroun

Résumé. Dans cet article, nous proposons une approche générale de prédiction des communautés basée sur un modèle d'apprentissage automatique pour la prédiction des interactions. En effet, nous pensons que, si on peut prédire avec précision la structure du réseau, alors on a juste à rechercher les communautés sur le réseau prédit. Des expérimentations sur des jeux de données réels montrent la faisabilité de cette approche.

1 Introduction

Les réseaux sociaux sont dynamiques par nature. La détection de communautés a longtemps considéré uniquement une vue statique : la capture du réseau à un instant t. Récemment, des travaux sur la dynamique des communautés ont vu le jour. Certains auteurs essayent de suivre l'évolution des communautés durant plusieurs tranches de temps Palla et al. (2007), d'autre proposent de mettre à jour les communautés existantes en fonction des nouveaux événements qui se produisent (ajout ou suppression de nœuds et/ou de liens) Nguyen et al. (2011). Enfin les derniers essayent de trouver des communautés consistantes sur plusieurs tranches de temps Aynaud et Guillaume (2011).

Un des problème non encore exploré dans la littérature sur la dynamique des communautés est la prédiction : connaissant l'évolution du réseau jusqu'au temps t, peut-on prédire les communautés au temps $t + 1$? Dans cet article, nous proposons une approche générale de prédiction des communauté basée sur la prédiction des interactions dans les réseaux complexes. Dans cette approche, étant donné l'évolution du réseau jusqu'au temps t, les interactions sont prédites pour le temps $t + 1$ et les communautés sont ensuite calculées sur ce réseau prédit. L'hypothèse qui soutient cette démarche est la suivante : si on est capable de prédire l'évolution du réseau avec précision, alors on peut utiliser le réseau prédit pour d'autres tâches de prédiction (ici la prédiction des communautés).

Dans la suite de cet article, nous présentons d'abord la prédiction des interactions (section 2) puis nous présentons son évaluation et son application à la prédiction des communautés (section 3). Nous terminons par des conclusions et perspectives (section 4).

2 Prédiction des interactions

Le problème de prédiction des interactions peut être défini comme suit : étant donné un réseau dynamique $G = (G_1, ..., G_n)$ dont les tranches de temps sont non cumulatives (les liens correspondent aux interaction de la tranche de temps uniquement) quelle sera la structure du réseau (G_{n+1}) correspondant à la tranche de temps $n + 1$? Ce problème peut être vu comme une généralisation de la prévision de lien : on ne se limite pas aux liens non existants mais on vérifie aussi que les liens existants resterons présents. De ce fait, les mêmes classes de méthodes peuvent être utilisées pour le résoudre. Dans ce qui suit nous présentons une solution basée sur la similarité et une solution par apprentissage supervisée. Dans les approches proposées, le temps joue un rôle important.

2.1 Modèle basés sur la similarité

Dans cette approche une fonction de similarité est définie et sa valeur est calculée pour chaque paire de nœuds potentielle. Un seuil est ensuite choisi pour décider de la création des interactions. La forme générale de la mesure de similarité proposée est :

$$Sim(i,j) \quad = \quad \sum_{t \in T} f(t) \quad \times \quad (\alpha W[i,j] \quad + \quad \beta g(neigh(i), neig(j)) \quad + \quad \theta h(i,j)) \quad (1)$$

Dans l'équation 1, les fonctions f et g et les paramètres α et β sont à définir. W est la matrice des poids. f est la fonction temporelle, elle permet de prendre en compte l'âge des interactions (donner plus d'importance aux relations récentes par exemple). g est la fonction de similarité topologique qui mesure la proximité dans le graphe social. h est la fonction de similarité entre les attributs (lorsqu'ils sont disponibles). Enfin $neigh(i)$ est une fonction de voisinage (les voisins, les voisins et leurs voisins, la communauté par exemple).

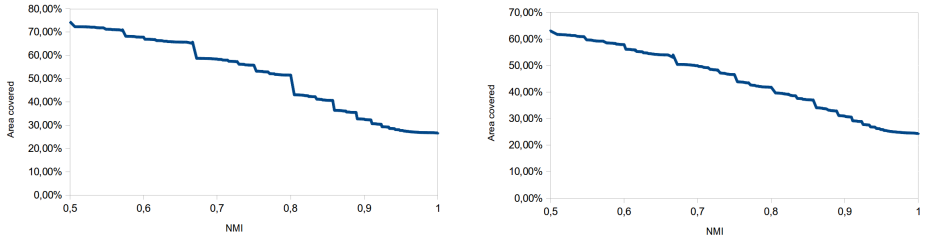
Ce modèle est très intuitif. Cependant, il ne peut pas modéliser une large classe de relations possible entre les variables d'entrée et la variable à prédire. Pour cette raison, nous proposons dans la suite un modèle plus général basé sur l'apprentissage supervisé.

2.2 Modèle d'apprentissage supervisé

Pour l'approche supervisée, nous proposons de procéder comme suit : pour chaque tranche de temps t de la période d'apprentissage, les attributs suivants dont calculés pour chaque paires de nœuds candidate : le nombre de voisins communs, le nombre interactions pour cette tranche de temps, le degré de chaque nœud, le coefficient de clustering de chaque nœud, le score de similarité entre les attributs (si disponibles). Les classes réelles sont obtenues sur la période de test. Une méthode d'apprentissage supervisé peut alors être utilisée pour construire le modèle de prédiction. Les expérimentations sont basées sur les Machines à Vecteurs de Support (SVM). Cette méthode est plus générale et plus flexible si on veut ajouter d'autres attributs.

	DBLP	Facebook wall
Modèle aléatoire	0,50	0,50
Modèle basé sur la similarité	0,69	0,84
Approche supervisée sans attributs	**0,87**	**0,92**
Approche supervisée avec attributs	**0,88**	-

TAB. 1 – *Évaluation (AUC) des modèles de prédiction d'interactions*



(a) (b)

FIG. 1 – *Évaluation de la prédiction des communautés locales.*

3 Evaluation et discussion

Les jeux de données utilisés pour les évaluations sont *DBLP* et *Facebook wall*. DBLP est un réseau de collaborations entre auteurs indexés sur `http://dblp.uni-trier.de/`. Pour chaque tranche de temps (années), une interaction existe entre deux auteurs s'ils ont au moins une publication commune. Les liens sont pondérés par le nombre de publications communes. Le jeu de données Facebook wall (`http://konect.uni-koblenz.de/networks/facebook-wosn-wall`) est un réseau construit à partir d'un sous-ensemble d'utilisateurs de la New Orleans. Pour chaque année, une interaction existe entre deux utilisateurs si l'un deux a publié sur le mur de l'autre. Les liens sont pondérés par le nombre de publications.

En raison du déséquilibre entre les classes, l'aire sous la courbe ROC (Receiver Operating Characteristic) notée AUC (Area Under Curve) est utilisée pour évaluer les performances des approches de prévision des interactions.

La table 1 présente les résultats de l'évaluation de la prédiction des interactions. Dans le jeu de données DBLP, on a à disposition les titres des articles et les noms des conférences ou journaux dans lesquels ils sont publiés. Pour chaque auteur et chaque année, on peut donc construire un vecteur TFIDF(Term Frequency Inverse Document Frequency) relatif aux mots contenus dans les titres de ses publications et les noms des conférences et/ou journaux dans lesquels il a publié. Ce sont ces vecteurs qui sont utilisés comme attributs.

Enfin, les communautés sont calculées en utilisant l'algorithme décrit dans Ngonmang et al. (2012) et évaluées en utilisant l'Information Mutuelle dans sa version Normalisée (NMI).

Les résultats de l'évaluation du jeu de données DBLP sont présentés sur la figure 3 (a). On peut constater que pour plus de 30% des nœuds on a une prédiction parfaite avec une valeur de

$NMI = 1$. Plus de 50% des nœuds produisent un NMI supérieur à $0,8$ et enfin, plus de 70% des nœuds produisent un $NMI > 0,6$. Un constat similaire peut être fait pour les résultats sur Facebook wall's présentés à la figure 3 (b).

Conclusions et perspectives

Récemment, de nombreux travaux sur la détection des communautés dans les réseaux dynamiques ont commencés. Un des problèmes encore non exploré est la prédiction des communautés. Dans cet article, nous avons dans un premier temps proposé des modèles pour la prédiction des interactions (un basé sur la similarité et l'autre par apprentissage supervisé). Ensuite, nous avons utilisé ces modèles pour prédire les communautés. Des tests sur des jeux de données réels montre la faisabilité de notre approche.

En perspective, nous pensons prendre en compte l'arrivée de nouveaux nœuds.

Remerciement

Ce travail est partiellement financé par le projet FUI français AMMICO.

Références

Aynaud, T. et J.-L. Guillaume (2011). Multi-Step Community Detection and Hierarchical Time Segmentation in Evolving Networks. In *Fifth SNA-KDD Workshop Social Network Mining and Analysis, in conjunction with the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2011)*.

Ngonmang, B., M. Tchuente, et E. Viennet (2012). Local communities identification in social networks. *Parallel Processing Letters 22*(1).

Nguyen, N., T. Dinh, Y. Xuan, et M. Thai (2011). Adaptive algorithms for detecting community structure in dynamic social networks. In *INFOCOM, 2011 Proc. IEEE*, pp. 2282 –2290.

Palla, G., A. lászló Barabási, T. Vicsek, et B. Hungary (2007). Quantifying social group evolution. Volume 446, pp. 2007.

Summary

In this paper, we propose a general approach for communities prediction based on a machine learning model predicting interaction in social networks. In fact, we believe that if one is able to predict the structure of the network with a high precision, then one just need to compute the communities on this predicted network to have the prediction of the community structure. Evaluation on real datasets (DBLP and Facebook walls) shows the feasibility of the approach.