

Détection de communautés dans les réseaux socio-sémantiques par point de vue

Juan David Cruz*, Cécile Bothorel*, François Poulet**

*Département LUSSE, Télécom – Bretagne
{juan.cruzgomez, cecile.bothorel}@telecom-bretagne.eu,
<http://www.telecom-bretagne.eu/>

**Université de Rennes 1 – IRISA
francois.poulet@irisa.fr
<http://www.irisa.fr/texmex/>

Résumé. Les algorithmes classiques de détection de communautés dans les réseaux sociaux utilisent l'information structurelle pour détecter des groupes, i.e la topologie du graphe de relations. Toutefois, ils ne prennent en compte aucune information externe qui peut guider le processus et aider à la réalisation des analyses du réseau selon différentes perspectives. La méthode proposée pour détecter des communautés utilise, de façon conjointe, l'information sémantique du réseau, représentée par des "points de vue", et l'information structurelle. Elle permet la combinaison entre les relations sociales explicites, les arêtes du graphe social, et les relations implicites, dites sémantiques, correspondant par exemple à des intérêts ou des usages similaires.

1 Introduction

Un réseau social est un ensemble de personnes qui sont liées selon différentes attaches qui proviennent de différents types d'interaction. Cela signifie que, dans un réseau social, peuvent coexister différentes relations, entre autres, d'amitié, de travail et d'appartenance à des organisations. Mais les membres d'un réseau social sont également décrits par leur appartenance à une entreprise, à une communauté, ou par leurs centres d'intérêts, ou encore par leur usage de contenus.

L'information riche décrivant relations et acteurs forment l'information "sémantique" du réseau social. Considérant ces attributs descriptifs dépassant la nature structurelle d'un réseau social, il est possible d'analyser celui-ci selon différentes perspectives, sans se limiter à la topologie.

2 Les points de vue dans un réseau socio sémantique

Étant donné un réseau socio-sémantique, c-à d. un réseau social augmenté de son information sémantique, nous définissons la notion de "point de vue".

Détection de communautés par point de vue

D'une manière générale, le point de vue peut être représenté comme un ensemble de caractéristiques associées à un sous-ensemble de l'information sémantique du réseau.

Formellement, soit \mathbf{S} l'ensemble d'information sémantique d'un réseau social, et soit $s_i \subset \mathbf{S}, i = 1, 2, \dots, p$ le sous-ensemble d'information lié au point de vue i . Le point de vue PoV_i se formule comme suit :

$$PoV_i = s_i \times V$$

où V est l'ensemble des acteurs du réseau social.

Cela veut dire que, le point de vue est l'association d'attributs/valeurs entre chaque acteur dans le réseau et les caractéristiques sélectionnées par le point de vue.

Un point de vue peut être aussi défini avec les attributs des liens, toutefois ce cas n'est pas traité dans ce papier.

		Point de Vue			
nœuds	Trending topic	France	USA	N/A	
nœud 1	1	0	1	0	} Instance ξ_j
\vdots	\vdots	\vdots	\vdots	\vdots	
nœud m	1	0	0	1	

FIG. 1 – Structure d'un point de vue. Chaque jeu des caractéristiques assignées à chaque nœud s'appelle une instance. Exemple ici pour un réseau Twitter que l'on souhaiterait visualiser selon un thème phare et une composante géographique.

Par exemple, Twitter peut se représenter comme un graphe dirigé, dont les nœuds représentent des personnes, et les arêtes représentent les relations entre eux. La Fig. 1 montre un exemple d'un point de vue créé avec un sujet de mode sur Twitter et la geo-localisation des acteurs du réseau.

3 Détection de communautés

La détection de communautés permet d'identifier différents groupes dans un réseau social. Dans ce cas, l'information structurelle du réseau est utilisée. Boutin (Boutin et Hascoet (2004)), Newman (Newman et Girvan (2004)), Blondel (Blondel et al. (2008)) et Brandes (Brandes et al. (2008)), entre autres, ont proposé des algorithmes qui utilisent l'information du nombre des arêtes en minimisant les arêtes inter-groupes tout en maximisant le nombre des arêtes intra-groupes.

Ces méthodes génèrent de bons résultats pour des réseaux peu denses (degré moyen faible), toutefois, avec des matrices denses, ces algorithmes trouveraient un résultat trivial : chaque nœud constitue à lui-seul un groupe, ou tous les nœuds sont regroupés dans un seul groupe. Ces méthodes n'utilisent pas l'information sémantique du réseau.

3.1 Détection de communautés en utilisant l'information sémantique

Afin d'inclure l'information sémantique en plus de l'information structurelle dans la détection de communautés, nous proposons un système qui procède au clustering en deux phases. La Fig. 2 illustre le fonctionnement général du système.

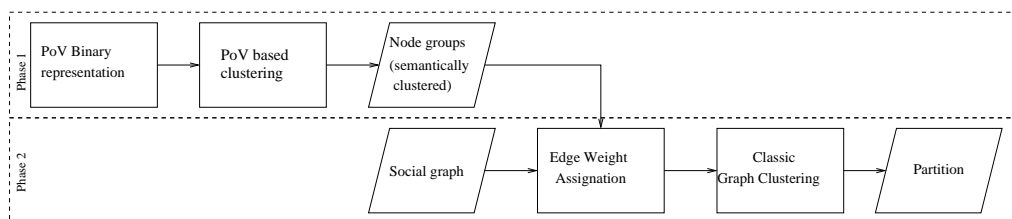


FIG. 2 – Architecture générale du système. Pendant la première phase, les groupes sémantiques sont identifiés. Ensuite, selon ces groupes, les arêtes sont pondérées pour utiliser un algorithme classique de détection de communautés.

Pendant la première phase, avec le point de vue *PoV* sélectionné, sont trouvés, en utilisant la méthode Self-Organizing Maps (Kohonen (1997)), des regroupements de nœuds selon la ressemblance de leurs instances de point de vue.

Pour la deuxième phase, nous affectons à chaque arête du réseau social un poids dérivé de la distance entre chaque groupe résultant de la première phase. Ensuite, un algorithme classique de détection de communautés est utilisé sur le graphe pondéré. Dans ce cas, il est possible d'utiliser l'algorithme classique tout en tenant du biais introduit par le point de vue.

3.2 Exemple d'utilisation

Quelques expériences avec des réseaux sociaux et points de vue ont été exécutées afin de tester le fonctionnement du système présenté en Fig. 2. L'application visée est la visualisation de données de réseautage réelles, telles qu'on peut les imaginer sous Twitter par exemple. Ce qui nous intéresse plus particulièrement est l'identification du changement de la configuration des groupes par la sélection d'un point de vue particulier.

Nous avons utilisé un réseau avec 200 nœuds et 3892 arêtes, et un point de vue composé de cinq caractéristiques. Chaque instance a été générée de façon aléatoire. Le réseau a une forme proche de celle des extractions de Twitter que nous avons réalisées (en terme de distribution de degrés, densité). L'algorithme classique pour détecter des communautés est celui proposé par Blondel (Blondel et al. (2008)), lequel utilise la modularité Q (Newman et Girvan (2004)) comme mesure de qualité.

Expérience	Final Q	Average Intracluster Distance	Standard Deviation
Classic Clustering	0.1885	1.5375	0.0169
PoV Based Clustering	0.4439	0.9954	0.6021

FIG. 3 – Résultats de l'exécution de l'algorithme classique de détection de communautés vers l'exécution de l'algorithme proposé.

La Fig. 3 présente un tableau avec les résultats pour les données décrites ci-dessus. Les résultats préliminaires montrent que la distance moyenne, calculée avec la mesure de distance Euclidienne, de chaque groupe trouvé par notre l'algorithme est moindre que la distance

Détection de communautés par point de vue

moyenne des groupes trouvés par l'algorithme classique : une distance moins grande indique que les nœuds sont plus similaires. Dans la perspective structurale, la modularité est supérieure avec notre méthode.

4 Conclusion

L'information contenue dans un réseau socio-sémantique est liée à certaines caractéristiques des acteurs et des relations. Une telle information permet d'analyser (et d'appréhender par la visualisation) le réseau depuis différentes perspectives, i.e. selon différents points de vue.

Les algorithmes classiques de détection de communautés utilisent l'information de la structure du réseau, et n'utilisent pas l'information sémantique pour influencer le processus de détection.

L'assignation de poids aux arêtes, dérivés des regroupements de nœuds faisant sens dans le contexte d'un point de vue, permettent d'associer les deux types d'information pour trouver des communautés et visualiser un réseau social selon le point de vue choisi.

Le travail à venir inclut l'extraction des caractéristiques depuis des données réelles, et l'expérimentation avec des points de vue créés à partir des liens. Le but de cette expérimentation est d'utiliser conjointement l'information des nœuds et des liens pour générer des points de vue.

Références

- Blondel, V. D., J.-L. Guillaume, R. Lambiotte, et E. Lefebvre (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics : Theory and Experiment* 2008(10), P10008 (12pp).
- Boutin, F. et M. Hascoet (2004). Cluster validity indices for graph partitioning. pp. 376 – 381.
- Brandes, U., M. Gaetler, et D. Wagner (2008). Engineering graph clustering : Models and experimental evaluation. *Journal of Experimental Algorithmics* 12, 1–26.
- Kohonen, T. (1997). *Self-Organizing Maps*. Springer.
- Newman, M. E. J. et M. Girvan (2004). Finding and evaluating community structure in networks. *Phys. Rev. E* 69(2).

Summary

Classic algorithms for community detection in social networks use the structural information to identify the groups in the social network, i.e., the topology of the relationships. However, these methods don't take into account any external information which could guide the clustering process, and which add elements to do further analyses. The method we propose, uses in a conjoint way, the semantic information from the network, represented by the point of view, and the structural information. This is, combining the relationships, expressed by the edges on one hand, and the implicit relations deduced from the semantic information on the other hand.