SemMEP : Nouvelle approche sémantique pour la détection des communautés dans un réseau social

Sami Ben Amor, Lotfi Ben Romdhane, Mounira Harzallah

SemMEP : Nouvelle approche sémantique pour la détection des communautés dans un réseau social

Sami Ben Amor[1], Lotfi Ben Romdhane[1] et Mounira Harzallah[2]

[1] Groupe de recherche MARS, Université de Sousse,
ben_amor_sami@outlook.com, lotfi.ben.romdhane@gmail.com
[2] Data User Knowledge, LINA, Université de Nantes
mounira.harzallah@univ-nantes.fr

Résumé : Plusieurs travaux ont porté sur la détection des communautés dans les réseaux sociaux. La majorité d'entre eux considère seulement la structure d'un réseau en négligeant la richesse sémantique des informations associées à ses utilisateurs et aux liens entre eux. D'autres approches se sont focalisées sur ses aspects sémantiques. Récemment des nouvelles approches ont proposé une modélisation conjointe de ces deux aspects. Dans cet article, nous proposons une nouvelle approche favorisant l'aspect sémantique d'un réseau social tout en tenant compte de sa structure. Une nouvelle fonction de qualité et un nouvel algorithme SemMEP pour l'optimiser sont définis, utilisant les informations sémantiques d'un réseau dans plusieurs étapes du processus de détection, à l'aide des mesures sémantiques.

Mots-clés : Réseau social, analyse sémantique, ontologie, détection des communautés.

1 Introduction

Les réseaux sociaux, tels que Facebook, Twitter et Flickr, sont devenus un moyen de communication très important utilisé par le grand public ainsi que par les professionnels. En raison de l'explosion de ces réseaux et de la richesse de leur contenu, l'intérêt pour leur analyse a augmenté au cours des dernières années. Plusieurs travaux se sont intéressés à la détection des communautés dans ces réseaux. La majorité d'entre eux considère seulement la structure d'un réseau et néglige la richesse des informations liées à ses acteurs et aux liens entre eux. D'autres s'intéressent seulement à la sémantique de ces informations et ignorent l'aspect structurel d'un réseau. Pour faire face à ce problème, on a associé au lien entre deux acteurs un poids défini par une combinaison linaire de l'intensité structurelle entre eux et la similarité sémantique de leurs informations (Dang & Viennet, 2012) (Cruz et al. 2013) (Zhang et al. 2015). Cependant, ce poids n'a pas été évalué d'une façon homogène pour tous les couples d'acteurs. La majorité de ces approches applique des algorithmes classiques de détection des communautés dont le plus utilisé est celui de Louvain (Blondel et al. 2008).

Les informations dans ce type de réseaux ne sont pas faciles à prendre en compte pour la détection des communautés. En effet, on a besoin, en premier lieu, d'en extraire les plus pertinentes, d'identifier leur sémantique et de déterminer la similarité de celles associées à deux nœuds à l'aide des mesures de similarité. En deuxième lieu, on a besoin de savoir les intégrer dans une modélisation d'un réseau social et de bien les prendre en compte tout au long du processus de détection des communautés. Récemment, des nouvelles approches utilisent une ontologie pour définir la sémantique de ces informations. Erétéo et al. (2011) et Leprovost et al. (2012) considèrent seulement les concepts d'une ontologie qui généralisent ceux qui annotent un réseau social. Wan et al. (2014) attribuent seulement trois valeurs

possibles à la similarité sémantique des concepts qui annotent deux nœuds (1, 0.5, 0). Dans ces différentes approches, les aspects sémantiques d'un réseau social ne sont pas pris en compte que dans la première étape du processus.

Nous avons développé l'approche SemMEP proposant une nouvelle fonction de qualité « SemEP » mesurant la qualité d'un partitionnement du point de vue structurelle et sémantique. SemMEP, définie par analogie avec l'approche structurelle MWEP, utilise dans plusieurs étapes du processus de détection les résultats des mesures sémantiques appliquées à l'ontologie qui annote un réseau. Dans cet article, nous présentons notre approche SemMEP et son expérimentation sur le réseau de « Karaté » suivant plusieurs scénarios d'annotation.

2 SemMEP : Maximisation Sémantique de l'Equilibre et de la Pureté

Notre approche SemMEP est une extension de l'approche MWEP pour la détection des communautés dans un réseau social modélisé par un graphe pondéré (Zardi & Ben Romdhane, 2013b). MWEP a donné des bons résultats de la qualité de partitionnement et de la rapidité d'exécution. Dans cette approche, on a défini la fonction de qualité «WEP» et l'algorithme MWEP pour l'optimiser. Ce dernier est composé principalement de deux phases : la pureté et l'équilibre. La phase de pureté reflète l'attachement d'un nœud du réseau à une communauté en fonction de la force de ses relations avec ses voisins. Pour décider si un nœud v est pur par rapport à une communauté C_i et donc l'associer à cette communauté, on compare sa compatibilité à C_i, noté $comp(v, C_i)$ qui est définie par la somme des poids w des arêtes qui le relient avec ses voisins directs de C_i et sa compatibilité maximale noté $comp_{\max}(v)$ qui est le maximum entre la somme des poids de ses voisins libres et directs ($PVLD(v)$) et les valeurs de ses compatibilités à toutes les communautés d'une partition P :

$$comp(v, C_i) = \sum_{v' \in Ci} w(v, v') \quad (1)$$

$$comp_{\max}(v) = \max\{comp(v, C_i), \forall C_i \in P; PVLD(v)\} \quad (2)$$

Si $comp(v, C_i) = comp_{\max}(v)$ alors v est pur par rapport à C_i.

La deuxième phase est celle de l'équilibre qui vise à avoir une interaction entre les nœuds plus forte à l'intérieur d'une communauté C_i qu'à l'extérieur. Pour cela une comparaison entre la séparabilité et la compacité d'une communauté est faite :

Si $séparabilité_{moy}(C_i) < compacité(C_i)$ alors la communauté C_i est équilibrée.

où $séparabilité_{moy}(C_i) = moyenne\{ séparabilité(C_i, C_j) > 0, \forall C_j \in P\}$ (3)

$$séparabilité(C_i, C_j) = \sum_{v \in Ci,\, v' \in Cj} w(v, v') \quad (4)$$

$$compacité(C_i) = \frac{1}{2} \sum_{v, v' \in Ci} w(v, v') \quad (5)$$

Cependant, MWEP ne tient pas compte des informations dans un réseau social. En plus, dans sa première phase, on peut décider qu'un nœud v n'est pas pur par rapport à une communauté si il a un nombre élevé de voisins libres et directs ($VLD(v)$) dont la somme des poids ($PVLD(v)$) est supérieure à celle des poids de ses voisins qui se trouvent dans la communauté formée ($comp(v, C_i)$), bien qu'on peut avoir une similarité forte entre v et ses voisins de C_i. Par exemple, dans la figure 1, le nœud 5 ne sera pas ajouté à C_i bien qu'il est sémantiquement beaucoup plus proche de ses voisins de C_i que de ses voisins libres directs car $PVLD(5) > comp(5, C_i)$. En plus, dans la deuxième phase, on peut avoir à fusionner deux communautés qui ne sont pas proches sémantiquement, car la compacité d'une

communauté est comparée avec la moyenne de ses séparabilités avec les communautés qui y sont attachées.

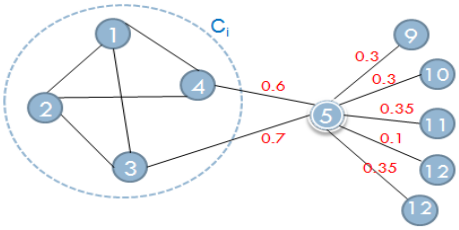

FIGURE 1 – Limite sémantique de MWEP

Par analogie avec MWEP, nous proposons une nouvelle approche SemMEP qui prend en compte plus la sémantique des informations liées aux nœuds que leur degré de connexion et cela en regroupant ensemble les nœuds très proches sémantiquement à condition qu'ils soient connectés. Dans cette approche, nous définissons la fonction de qualité «SemEP» qui dépend de deux nouvelles notions : la pureté sémantique et l'équilibre sémantique. Notre approche s'applique aux réseaux sociaux ayant des liens symétriques ne représentant pas une forte intensité structurelle, par exemple des liens d'amitié dans Facebook.

2.1 Formalisation sémantique d'un réseau social

Dans notre approche, nous représentons un réseau social comme un graphe annoté non orienté $G=(V,E,A,W)$ où V est l'ensemble de ses nœuds, E est l'ensemble de ses arêtes et A est l'ensemble des annotations des nœuds. W est une fonction qui modélise la similarité sémantique de deux nœuds qui sont obligatoirement liés. W est défini comme suit :

$$\mathrm{W}(v_i,v_j)=\mathrm{Sim}(v_i,v_j)*\delta(v_i,v_j);\forall\, v_i,v_j\in V \qquad (6)$$

où $\mathrm{Sim}(v_i,v_j)$ est la similarité sémantique des informations associées aux nœuds v_i et $\mathrm{v_j}$, $\delta(v_i,v_j)$=1 si v_i et v_j sont liés par une arête, 0 sinon.

2.2 La fonction de qualité « SemEP »

Soit une partition $P=\{C_1,...,C_k\}$ d'un réseau social représenté par $G=(V,E,A,W)$, nous définissons la fonction de qualité SemEP (Equilibre et Pureté Sémantique) comme suit :

$$SemEP(P)=\frac{1}{2}\left[SemP(P)+SemE(P)\right] \qquad (7)$$

La pureté sémantique ($SemP$) reflète l'attachement sémantique d'un membre du réseau à une communauté C_i. Pour décider si un nœud v de V est pur sémantiquement par rapport à C_i, sa similarité à C_i notée $Sim(v,C_i)$, définie par la similarité entre v et ses voisins directs de C_i, est comparée à sa similarité maximale (notée $Sim_{\max}(v)$) qui est le maximum entre sa similarité avec ses VLD, notée $\mathrm{Sim(v,C_{VLD})}$, et sa similarité avec chacune des autres communautés différentes de C_i :

$$Sim_{\max}(v)=\max\{Sim(v,C_i),\forall C_i\in P;Sim(v,C_{VLD})\} \qquad (8)$$

Si $Sim(v,C_i)=Sim_{\max}(v)$ alors v est pur sémantiquement par rapport à C_i. En cas où la similarité maximale est atteinte pour plusieurs communautés, celle qui a plus de liens avec v est favorisée. $SemP(C_i)$ est définie comme suit :

$$SemP(C_i)=\frac{\left|SemP(\mathrm{C_i})\right|}{\left|C_i\right|} \qquad (9)$$

$$avec \ \left|SemP(C_i)\right| = \left|\left\{ Sim\,(v, C_{\mathrm{i}}) = Sim_{\max}(v)\right\}\right| \qquad (10)$$

Enfin, la pureté sémantique d'un partitionnement est définie comme suit :

$$SemP(P) = \frac{1}{|P|}\sum_{C_i \in P} SemP(C_i) \qquad (11)$$

Pour vérifier si C_i est équilibrée sémantiquement, sa similarité avec chacune des autres communautés attachées à elle, noté $Sim(C_i, C_j)$ est comparée à un seuil α. Cependant, cette condition seule est insuffisante car on peut avoir deux communautés similaires mais les nœuds qui les connectent n'ont pas d'intérêt à être fusionnés, car ils sont distants sémantiquement. Il faut donc en plus comparer la similarité des deux ensembles des nœuds qui relient deux communautés (séparabilité) qui font l'objet d'une fusion, notée $Sim(C_{ij}, C_{ji})$ à un seuil β. Ci est équilibrée sémantiquement (notée $SemE(C_i)$) si $Sim(C_i, C_j) < \alpha$ ou $Sim(C_{ij}, C_{ji}) < \beta$. L'équilibre sémantique d'un partitionnement P, noté $SemE(P)$, est défini comme suit :

$$SemE(P) = \frac{\left|SemE(C_i)\right|}{|P|} \qquad (12)$$

Afin d'évaluer notre approche, nous avons adapté l'indicateur de performance CI (Connectivité Index) défini dans (Zardi & Ben Romdhane, 2013a) en proposant l'indicateur FCS (Force de Connectivité communautaire Sémantique). Il est défini avec la force de connectivité communautaire sémantique de C_i (notéé $FCS_{Com}(Ci)$) qui est la force de densité sémantique de C_i. $FCS_{Com}(Ci)$ est définie en fonction de la compacité de C_i et sa séparabilité moyenne, comme suit :

$$FCS_{Com}(C_i) = \frac{Compacité(C_i) - separabilité_{moy}(C_i)}{Compacité(C_i) + separabilité_{moy}(C_i)} \qquad (13)$$

Plus la compacité de la communauté est importante et sa séparabilité moyenne est faible plus sa force de connectivité sémantique est grande quel que soit la taille de cette communauté. *FCS* d'une partition est la moyenne de $FCS_{Com}(C_i)$.

2.3 Algorithme SemMEP

SemMEP commence le processus de détection des communautés par une partition initiale dans laquelle chaque communauté contient un seul nœud. Initialement tous les nœuds sont considérés libres. Ensuite, la somme des poids connectant chaque nœud à ces voisins directs représentant un degré de poids, est calculée et les nœuds sont triés selon l'ordre décroissant de ce degré. Ces étapes représentent la phase d'initialisation de MWEP et de SemMEP. Ensuite, SemMEP comporte deux nouvelles phases pour la recherche de la pureté et de l'équilibre sémantiques, définies par analogie à la démarche algorithmique de MWEP.

Algorithme : SemMEP
Données : Un graphe Annoté et pondéré G=(V,E,A,W), une ontologie O
Résultat : Un ensemble des communautés P={C_1,…,C_k}
Début :
1. $P \leftarrow \emptyset$
2. **Pour** i=1 à |V| **faire** $C_i \leftarrow v_i$ **Fin Pour**
3. Trier les nœuds dans *Libre* selon leurs degré de poids
4. **Tant qu'ils** existent des nœuds *Libre* **faire**
   Sélectionner le premier nœud v dans Libre
   si $\exists$ C$\in$P tel que v est pur sémantiquement par rapport à C **alors**
   Ajouter v à C et Supprimer v du vecteur Libre **sinon**

```
    Créer une nouvelle communauté Cn contenant v et ses voisins libres
    Supprimer les nœuds non purs sémantiquement de Cn
    Ajouter à Cn les voisins purs de chaque nouveau membre et Ajouter Cn àP
    Supprimer les membres de Cn du vecteur Libre
    Fin Tant que
5.  Tant qu'il existe des communautés non équilibrées sémantiquement faire
      Fusionner chaque communauté non équilibrée avec la plus similaire
      Déplacer les sommets qui deviennent impurs vers les groupes dans
        lesquels ils seront purs sémantiquement
    Fin Tant que
Fin.
```

3 Expérimentation

Pour expérimenter notre méthode, nous avons considéré le réseau « Karaté » que nous avons annoté avec une ontologie qui définit 4 centres d'intérêt (Cinéma, Sport, Informatique et Musique), selon 4 scénarios d'annotation (S1, S2, S3, S4). Nous avons appliqué des mesures de similarité sémantique adéquates sur cette ontologie, pour mesurer la similarité de deux concepts ou de deux groupes de concepts qui annotent des nœuds ou des communautés (Blanchard et al. 2008) (Harzallah et Berio, 2015). Nous avons ensuite appliqué notre approche et l'approche sémantique de Dang et Viennet (2012) (notée dans la suite ADV) pour la détection des communautés sur ces scénarios. ADV utilise l'algorithme de Louvain pour optimiser la fonction de qualité « Modularité » notée Q, et la densité et l'entropie comme indicateurs de performance. La table1 illustre les résultats de cette expérimentation.

TABLE 1 – *Résultats de SemMEP et de ADV*

Approche / Mesures	ADV				SemMEP			
	S1	**S2**	**S3**	**S4**	**S1**	**S2**	**S3**	**S4**
NC	3	4	4	3	2	4	3	4
Q/SemEP	0.53	0.58	0.58	0.55	1	1	1	1
Densité	0.80	0.73	0.73	0.80	0.87	0.73	0.78	0.71
Entropie	0	0.01	0.01	0.37	0	0	0.07	0.29
FCS	0.68	0.61	0.61	0.78	1	0.70	0.78	0.60

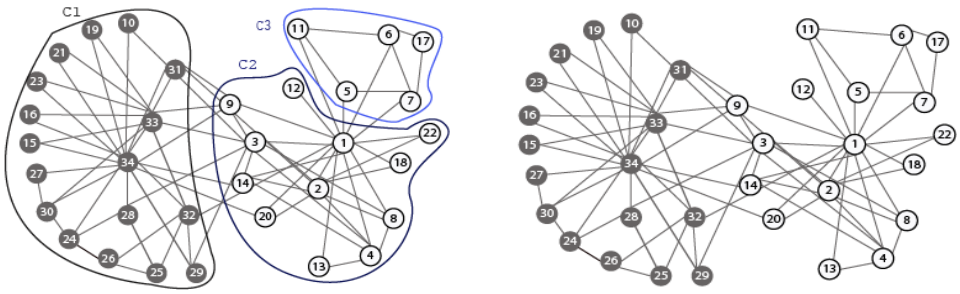

FIGURE 2 – (a) Partition de S1 avec ADV (b) Partition de S1 avec SemMEP

Dans S1, le réseau est annoté par deux centres d'intérêt : Musique (nœud grisé) et Sport (nœud transparent) dans la figure 2. Pour S1, SemMEP a identifié 2 communautés, la première s'intéresse à la musique et la deuxième au sport (figure 2- b), ADV a déterminé 3 communautés illustrées par (figure 2- a). SemEP est optimale alors que Q= 0.53. Les valeurs de la densité et de FCS sont plus élevées pour SemEP que pour ADV. Dans S2, nous avons utilisé le résultat de la figure (2-a) et nous avons annoté les membres associés aux nœuds : 25, 26, 28 et 32 de C1 par le concept « Jazz » qui est une spécialisation de « musique » et les

nœuds associés aux membres de C3 par « cyclisme » qui est une spécialisation de « sport ». L'annotation des autres nœuds n'a pas changé par rapport à S1. Comme attendu, SemMEP a déterminé une partition avec 4 communautés : les nœuds 25, 26, 28 et 32 de C1, C2, C3 et les autres nœuds de C1. ADV a déterminé presque les mêmes communautés avec la différence que le nœud 29 annoté par la musique a migré vers la communauté qui s'intéresse au Jazz, ce qui a rendu cette communauté non strictement homogène. S3 est identique à S2 mais les seuils de similarités $\alpha et \beta$ ont été baissés de 1 à 0.5. Dans ce cas SemMEP a déterminé 3 communautés en fusionnant celles qui s'intéressent au sport et au cyclisme. S4 est une annotation aléatoire du réseau. Pour ces différents scénarios, les résultats de SemMEP pour les différentes mesures sont meilleurs que ceux de ADV, sauf pour la densité (S4). Tous ces résultats montrent que notre modèle SemMEP est plus intéressant que celui de ADV : il a donné un partitionnement plus homogène sémantiquement que celui de ADV.

4 Conclusion

Nous avons proposé une nouvelle approche SemMEP pour la détection des communautés dans un réseau social qui prend en compte plus les aspects sémantiques dans un réseau social que ses aspects structurels en regroupant ensemble les acteurs qui sont proches sémantiquement. Notre approche évalue la sémantique des informations dans un réseau avec une ontologie et des mesures sémantiques appliquées aux concepts de cette ontologie qui annotent les nœuds du réseau. En plus, elle a intégré ces aspects sémantiques dans différentes étapes de processus de détection. Des expérimentations préliminaires ont montré que notre approche donne des résultats plus intéressants que ceux de ADV. Dans nos travaux futurs, nous allons la valider sur des graphes à grand échelle.

Références

BLANCHARD E, HARZALLAH M et KUNTZ P. (2008). A generic framework for comparing semantic similarities on a subsumption hierarchy. *18th European Conference on Artificial Intelligence (ECAI)*, pp 20-24.

BLONDEL V. D., GUILLAUME J.-L., LAMBIOTTE R. et LEFEBVRE E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no 10, p. P10008 (12pp).

CRUZ J. D., BOTHOREL C. et POULET F. (2013). Community detection and visualization in social network: integrating structural and semantic information. *ACM transaction on intelligent systems and technology (TIST)*, vol. 5, n°1, pp. n°11.

DANG T. & VIENNET E. (2012). Community detection based on structural and attribute similarities. *In International conference on digital society (icds)*, p. 7–14.

ERETEO G., GANDON F. et BUFFA M. (2011). SemTagP: Semantic Community Detection in Folksonomies. *IEEE/WIC/ACM International Conference on Web Intelligence*, Lyon.

HARZALLAH M. & BERIO G.(2015). A Unified Framework for Semantic Comparison of Objects: Extension to Semantic Graphcomparison. *KES 2015*: 547-556

LEPROVOST D, ABROUK L. et GROSS-AMBLARD D. (2012). Discovering implicit communities in web forums through ontologies. *Web Intelligence and Agent Systems*, 10(1) :93–103.

WAN T, WEI LIU W& ZONGTIAN LIU, LIU Z. (2014). A community discovering method based on event network for topic detection. *Advanced Communication Technology (ICACT)*.

ZARDI H. & BEN ROMDHANE L. (2013a). An O(n2) algorithm for detecting communities of unbalanced sizes inlarge scale social networks. *Knowl.-Based Syst.* 37: 19-36

ZARDI H. & BEN ROMDHANE L. (2013b). WMEP: Efficiently Mining Community Structures in WeightedLarge Scale Social Graphs. *In The first International Conference on Reasoning and Optimization in Information Systems (ROIS'2013),* Hammam-Sousse, Tunisia, Sept 6–7, 2013.

ZHANG F., LI J., LI F, XU M, XU R. et HE X. (2015). Community Detection Based on Links and Node Features in Social Networks. *Lecture Notes in Computer Science*, pp 418-429.