Une approche centrée sur l'utilisateur pour intégrer les acteurs sociaux dans des communautés d'intérêt

A user-centered approach to insert social actors into communities of interest

Nadia Chouchani[1] and Mourad Abed[2]

[1]LAMIH, UMR, CNRS 8021
Valenciennes, France
Nadia.chouchani@uphf.fr
[2]LAMIH, UMR, CNRS 8021
Valenciennes, France
Mourad.abed@uphf.fr

RÉSUMÉ. La détection des communautés d'intérêt est un problème complexe qui a été abordé sous différents angles. Dans ce travail, nous proposons une approche centrée sur l'utilisateur intégrant des profils utilisateurs sociaux dans la détection des communautés dans les réseaux sociaux en ligne. Dans notre approche, nous calculons d'abord l'acquisition explicite des connaissances. En explorant les réseaux égocentriques des utilisateurs, nous pouvons déduire des similitudes implicites des intérêts. Ces similitudes sont estimées en référence à l'homophilie et à l'influence sociale. Cette dernière est utilisée pour améliorer l'analyse des sentiments au sein des communautés. Enfin, nous menons des expériences sur des ensembles de données extraits de réseaux sociaux réels.

ABSTRACT. Detecting communities of interest is a complex problem that has been approached from different angles. In this paper, we propose a user-centric approach integrating social user profiles in the detection of communities in online social networks. In our approach, we first compute the explicit acquisition of knowledge. By exploring the egocentric networks of users, we can infer implicit similarities of interests. These similarities are estimated with reference to homophilia and social influence. The latter is used to improve sentiment analysis within communities. Finally, we conduct experiments on data sets taken from real social networks.

MOTS-CLÉS. Réseaux Sociaux, Communauté d'intérêt, Profil utilisateur, Ontologie, Analyse des sentiments
KEYWORDS. Social Networks, Community of Interest, User Profile, Ontology, Sentiment Analysis

1. Introduction

Les Réseaux Sociaux sont des services basés sur le web dont la fonctionnalité principale est de connecter des personnes ou des entités. Selon Garton, ils sont définis comme "un ensemble d'individus, d'organisations ou d'entités entretenant des relations sociales fondées sur l'amitié, le travail collaboratif et l'échange d'information" (Garton *et al.*, 1997). Pour une meilleure compréhension des Réseaux Sociaux, les chercheurs ont essayé de trouver plus de caractéristiques structurelles de ces réseaux. (Barabasi, Albert, 1999 ; Faloutsos *et al.*, 1999) ont montré que les réseaux réels ne sont pas des graphes aléatoires car ils présentent de grandes hétérogénéités, révélant un niveau élevé d'ordre et d'organisation. En effet, ils ont une structure modulaire avec des nœuds formant des groupes et éventuellement des groupes au sein de groupes. Cette caractéristique s'appelle la structure communautaire (Girvan, Newman, 2002). La classification des nœuds d'un Réseau Social en communautés suppose, d'une part, que les nœuds appartenant à la même communauté aient au moins autant de liens les uns avec les autres qu'avec les autres nœuds appartenant aux autres communautés. D'autre part, et d'un point de vue sémantique, les nœuds partagent des intérêts communs.

En outre, il est fondamental de connaître ce que les membres d'une communauté pensent d'un sujet d'intérêt particulier. Profitant des quantas de données désormais disponibles dans les plateformes des

Réseaux Sociaux, les chercheurs sont en quête de moyens pour analyser automatiquement les opinions exprimées dans les publications diffusées. L'Analyse des Sentiments est le champ du traitement automatique des textes qui étudie les sentiments (Pang, Lee, 2008). Les Réseaux Sociaux permettent d'aller plus loin en identifiant par exemple les leaders d'opinions.

Dans notre travail, nous visons à améliorer la détection des communautés en exploitant conjointement la structure topologique des Réseaux Sociaux et leur sémantique. Les principales contributions sont : (i) la définition de ce que nous appelons un profil utilisateur social représenté à l'aide de l'ontologie *FOAF* [1] (Friend Of A Friend) que nous avons étendu pour comprendre des informations sur l'homophilie et l'influence ; et (ii) une approche en trois étapes centrée sur l'utilisateur pour la détection des communautés utilisant le profil utilisateur social et incorporant la méthode *LDA* pour représenter les sujets d'intérêt, un modèle de prédiction des intérêts et une méthode d'Analyse des Sentiments au niveau d'un utilisateur. Le reste du papier est organisé comme suit. Dans la section 2, nous présentons quelques travaux connexes. Dans la section 3, nous détaillons le profil utilisateur social. Dans la section 4, nous présentons notre approche visant à détecter des communautés d'intérêt en utilisant des modèles de prédiction des intérêts et de classification de la polarité des sentiments. Dans la section 5, nous discutons les résultats des expériences menées. Enfin, nous concluons et décrivons les travaux futurs.

2. Etat de l'art

La détection des communautés est l'un des principaux problèmes dans le domaine de l'Analyse des Réseaux Sociaux. Les méthodes de sa résolution ont fait l'objet de nombreux travaux de recherche depuis le travail fondateur de Girvan et Newman (Girvan, Newman, 2002). Notre revue de la littérature nous a conduit à une catégorisation des méthodes de détection de communautés en trois familles (Chouchani, Abed, 2018) : les méthodes basées sur les caractéristiques structurelles, les méthodes prenant en compte les attributs en plus de la topologie et les méthodes utilisant la structure du réseau avec des phénomènes sociaux tels que l'influence.

Les méthodes basées sur la topologie utilisent l'analyse structurelle des réseaux pour détecter des communautés. L'un des algorithmes les plus populaires est celui proposé par Clauset et al. (Clauset *et al.*, 2004). Les communautés sont détectées selon une approche de regroupement dans les graphes représentant les Réseaux Sociaux, de sorte que la densité des liens est plus élevée au sein des communautés qu'entre celles-ci (Kernighan, Lin, 1970 ; Newman, 2004 ; Blondel *et al.*, 2008 ; Palla *et al.*, 2005). Plusieurs mesures de positions sont proposées et liées aux caractéristiques structurelles afin de définir les groupes cohérents qui constituent les communautés. Cependant, la détection des communautés basée uniquement sur la topologie pose problème car il est difficile d'expliquer la sémantique de sa formation (D. Zhou *et al.*, 2006). En effet, cette dernière résulte de la similitude entre les acteurs sociaux et le simple fait de prendre en compte la structure topologique semble être insuffisant. Les chercheurs ont conclu que la détection des communautés devait prendre en compte les caractéristiques topologiques et sémantiques ensemble. Dans les travaux récents, des efforts ont été déployés dans ce sens.

Des méthodes basées sur la structure et les attributs topologiques s'intéressent aux deux sources d'information (Y. Zhou *et al.*, 2009 ; Xu *et al.*, 2012 ; Bothorel *et al.*, 2015 ; Yang *et al.*, 2013 ; Li *et al.*,

1. http ://www.foaf.com

2008 ; Combe *et al.*, 2012). En effet, le clustering combine les similitudes de structure et d'attributs. Les nœuds des mêmes communautés doivent être fortement connectés et avoir des attributs similaires. Dans ce contexte, les communautés sont définies par des structures cohérentes et des valeurs d'attributs homogènes.

Cependant, les Réseaux Sociaux ont des caractéristiques importantes de la communication humaine. En fait, un utilisateur est considéré comme une entité sociale. Ainsi, les comportements des utilisateurs au sein d'un réseau reflètent la logique de la diffusion de l'information. Ces faits sociaux se traduisent par des phénomènes tels que l'influence et la confiance. Plusieurs approches assument ces caractéristiques et proposent des modèles d'auto-corrélation pour la détection de communautés prenant en compte les caractéristiques structurelles et les phénomènes sociaux en modélisant le graphe social et les cascades d'informations (Barbieri *et al.*, 2013 ; Natarajan *et al.*, 2013 ; Ereteo *et al.*, 2011).

Sur la base de cette nouvelle classification, nous concluons qu'aucune approche ne prend en compte conjointement la structure topologique du Réseau Social, les attributs des utilisateurs et la sémantique. Il sera donc intéressant de tirer parti de tous ces critères pour détecter des communautés de plus en plus cohérentes.

Dans la résolution des problèmes de détection des communautés et d'Analyse des Sentiments, plusieurs travaux ont profité du phénomène d'homophilie (McPherson *et al.*, 2001). Cependant, aucun travail n'a exploité le phénomène d'influence social avec l'homophilie, qui est aussi bien crucial dans les Réseaux Sociaux (Crandall *et al.*, 2008). En effet, il a été démontré que l'homophilie et l'influence expliquent les similarités entre les utilisateurs.

3. Profil utilisateur social

Dans les Réseaux Sociaux, les utilisateurs sont connectés entre eux par diverses relations pour s'exprimer et échanger. Ils se ressemblent et, généralement, interagissent sur des intérêts similaires. D'où la nécessité d'un modèle de données pour représenter les utilisateurs ainsi que l'ensemble de leurs données sociales. Dans tout ce qui suit, nous adoptons la définition suivante des données sociales :

Definition 3.1. *Les données sociales sont les contenus disponibles dans les plateformes de Réseaux Sociaux, publiés ou partagés avec les utilisateurs. Elles comprennent leurs pages de profils, connections, publications, intérêts, etc.*

La modélistaion de l'utilisateur, relevant du domaine des Interactions Homme-Machines, vise à étudier l'extraction, l'analyse et la représentation des données et interactions des utilisateurs pour construire leurs profils. Le processus de construction de ces profils se fait au moyen de méthodes basées sur les données. Pour modéliser les données sociales, nous introduisons un modèle de profil utilisateur social qui est générique ainsi qu'extensible, pour :

— Premièrement, avoir une structure unique et intégrée contenant les différents types de données sociales, structurées ou non, et facilement étendue ;
— Et deuxièmement, pouvoir le réutiliser. Dans notre cas, nous l'exploiterons par la suite dans la personnalisation des applications sociales interactives.

Le processus de modélisation des profils utilisateurs comprend cinq phases (Fayyad *et al.*, 1996) qui sont décrites respectivement dans les sous sections qui suivent.

3.1. *Sélection des données*

Cette phase fait intervenir deux éléments qui sont : les producteurs et les sources de données. Dans notre travail, nous nous focalisons sur les données sociales donc les producteurs sont les utilisateurs des Réseaux Sociaux ; lesquels représentent les sources.

Notons que l'Analyse des Réseaux Sociaux, du point de vue des sciences sociales, est soit centrée sur un utilsateur, soit sur le réseau entier. Dans le premier cas, un réseau **égocentrique** est constitué par un individu, appelé **égo**, et l'ensemble de ses liens directs avec ses "amis" appelés **alters**. Ces derniers peuvent devenir eux-mêmes des égos et identifier à leur tour d'autres alters. Alors que la seconde exploite l'intégralité du Réseau Social. Dans ce cas, des éléments de la théorie des graphes et des mathématiques sont souvent utilisés pour définir des mesures de centralité des acteurs et des groupes.

Dans notre travail, nous procédons par une approche égo centrée, et définissons le réseau égocentrique d'un utilisateur comme le réseau constitué des relations avec ses alters dans son Réseau Social entier ainsi que les relations indirectes avec les "amis" des alters (jusuq'à un niveau maximal spécifié). Quant aux données sociales, elles comprennent les attributs des utilisateurs, leurs relations, leurs publications ainsi que leurs intérêts.

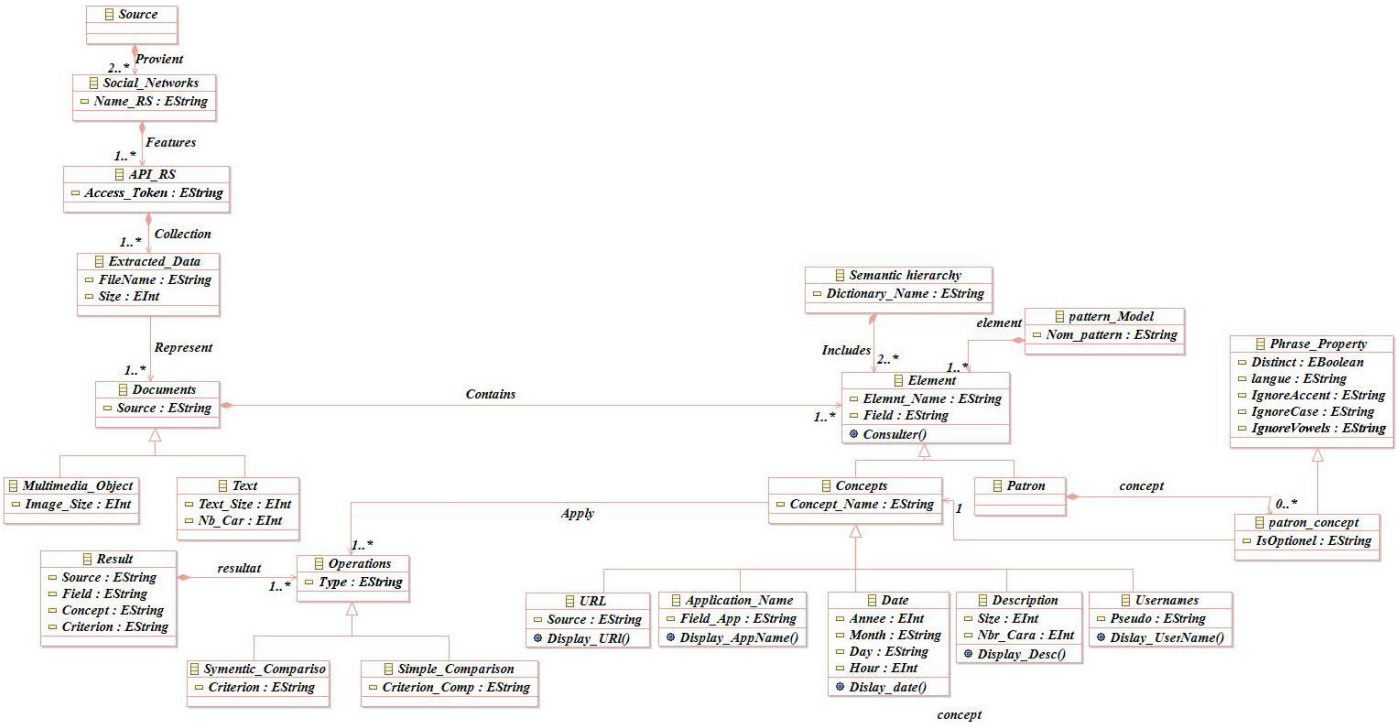

Figure 1. *Diagramme UML du méta-modèle de définition des patterns*

Cependant, nous nous sommes intéressés à extraire seulement les données pertinentes. Concernant les publications, il s'agit de celles qui sont en rapport avec le sujet d'intérêt en question. Une technique de filtrage d'information est donc nécessaire pour extraire les "bonnes" publications. Pour ce faire, nous avons défini un modèle pour la définition des patterns recherchés dans les publications, dont le diagramme *UML* est décrit dans la figure 1. Ce modèle est conçu de telle manière qu'il soit générique. En effet, il peut être appliqué à diverses plateformes de Réseaux Sociaux.

Un pattern (classe "*Pattern*") est composé de concepts (classe "*Concept*"), et dans sa forme la plus simple un concept est un ensemble de mots. Nous avons inclus des concepts spécifiques qui sont souvent utilisés dans les Réseaux Sociaux comme les mots clés et les Hashtags (un Hashtag étant un mot ou une phrase précédé du symbole #). Ce modèle est lié à un dictionnaire externe pour pouvoir rechercher les relations des hiérarchies sémantiques des concepts (comme les synonymes).

3.2. *Pré-traitement des données*

Certains traitements sont appliqués sur les données sélectionnées de la phase précédente. Ils permettent d'éliminer le bruit et de repérer les données incomplètes, incohérentes ou inconsistantes afin de garantir de meilleurs résultats dans les étapes suivantes. Parmi ces traitements, nous citons : le nettoyage des données, la discrétisation, la réduction, la transformation, le transcodage, etc.

En particulier, le processus d'agrégation des données est un moyen de nettoyage dans le cas de données manquantes. Nous visons à modéliser les données sociales de différents sites de Réseaux Sociaux. En effet, l'agrégation des données s'avère indispensable vu qu'un seul utilisateur pourrait avoir plusieurs comptes dans différents sites. Donc une donnée incomplète dans un site peut être disponible dans un autre. Pour ce faire, une identification unique de chaque utilisateur et une collection et récupération des données sont à effectuer.

3.3. *Transformation des données*

L'objectif principal de cette phase est d'organiser les données prétraitées dans une certaine structure, de telle sorte qu'elle soit bien adaptée à l'application d'algorithmes de fouille de données dans l'étape qui suit. En particulier, elles sont structurées suivant un modèle de profil utilisateur. Dans notre travail, nous distinguons les types de données suivants :

— Les données explicites : ce sont celles fournies de manière explicite par les utilisateurs, qui sont en général des éléments des pages de profils construites dans les plateformes des Réseaux Sociaux et leurs relations sociales établies comme les relations d'amitié, etc. ;

— Les données implicites : elles sont calculées en admettant les comportements, les activités, les interactions et les publications des utilisateurs, ou bien inférées par des techniques de prédiction.

— Les données de contexte : très souvent, elles influencent les comportements des utilisateurs. Elles sont essentielles à l'adaptation et la personnalisation des informations et des services. Plus particulièrement, le contexte de l'utilisateur comprend deux dimensions : sociale et personnelle (Tchuente *et al.*, 2012). La première réfère aux liens et connexions sociales c'est-à-dire le voisinage social ; et la deuxième, contient les contextes démographique (âge, genre, nationalité, etc.), psychologique (caractéristiques affectives, sentiments, etc.) et cognitif (centres d'intérêts, préférences, etc.).

— Les données sémantiques : lors du traitement de données textuelles, les données sémantiques enrichissent la sémantique des profils. Citons par exemple les dictionnaires des relations sémantiques hiérarchiques comme la synonymie ; aussi, les thesaurus permettant de classifier les termes.

3.4. *Fouille de données*

Pour une meilleure compréhension des contenus publiés par les utilisateurs, notre but est d'identifier automatiquement les sujets qui les intéressent à partir de leurs publications. En effet, généralement, les

utilisateurs n'expriment pas explicitement leurs sujets d'intérêt. Dans cette perspective, une solution possible consiste à utiliser les "hashtags" qu'ils publient. Cependant, il y a parfois des faibles usages du "hashtag" dans les ensembles de données, ce qui le rend inapproprié à être utilisé comme sujet d'intérêt.

La modélisation automatique des thématiques est par contre couramment utilisée pour analyser de grands volumes de contenus non étiquetés et extraire automatiquement les sujets d'intérêt, parfois appelés structures thématiques latentes. C'est dans ce but que nous appliquons le modèle *LDA* (Latent Dirichlet Allocation) (Blei *et al.*, 2003) ; pour identifier les intérêts latents des utilisateurs à partir d'une collection de documents représentant leurs publications. Il s'agit d'une technique d'apprentissage automatique non supervisé qui traite chaque document comme un vecteur de mots. Sur la base de cette hypothèse, un document est représenté comme une distribution de probabilité sur certains sujets d'intérêt, tandis qu'un sujet est représenté comme une distribution de probabilité sur un certain nombre de mots.

Le modèle a deux paramètres à inférer à partir des données observées, qui sont les distributions des variables latentes θ (document-sujet) et ϕ (sujet-mot). En déterminant ces deux distributions, il est possible d'obtenir les sujets d'intérêt sur lesquels les utilisateurs écrivent. Pour cette inférence, et vue sous l'angle de la maximisation de la log-vraisemblance, nous passons par des heuristiques. Dans notre travail, nous avons recours à *Gibbs Sampling*. Il s'agit d'une méthode de Monte-Carlo. D'abord, elle assigne aléatoirement les sujets. Ensuite, elle calcule les distributions conditionnelles sur des échantillons et, selon une certaine probabilité, assigne les sujets aux mots. Ainsi, cela recommence un grand nombre de fois pour obtenir une bonne approximation des distributions.

Le résultat est représenté en trois matrices :

1. *DT*, une matrice $n \times k$, où $DT_{i,j}$ contient le nombre de fois un mot dans les documents correspondants aux publications d'un utilisateur *i* a été assigné au sujet t_j ;
2. *WT*, une matrice $p \times q$, où $WT_{i,j}$ contient le nombre de fois un mot w_i a été assigné au sujet t_j ;
3. Z, un vecteur $1 \times p$, où Z_i est l'assignement d'un sujet à un mot w_i.

En particulier, nous nous intéressons particulièrement à la matrice DT contenant le nombre de fois un mot dans une publication d'un utilisateur été assigné à un sujet donné. Nous la normalisons sous forme d'une matrice DT' telle que $\|DT'_{i.}\| = 1$ pour toute ligne $DT'_{i.}$. Chaque ligne de DT' représente la distribution de probabilité de l'utilisateur i sur les k sujets, c'est-à-dire chaque élément DT'_{ij} contient la probabilité qu'un utilisateur i est intéressé au sujet j.

3.5. *Représentation des profils utilisateurs*

Les profils des utilisateurs peuvent être représentés suivant différents modes qui ont été proposés : ensembliste, multidimentionnel ou sémantique. Dans notre travail, nous optons pour une représentation sémantique conceptuelle qui considère en même temps la représentation des Réseaux Sociaux. En effet, en s'appuyant sur les technologies du Web sémantique et la théorie classique des graphes, nous visons à tirer parti des deux domaines et mener une analyse sémantique des Réseaux Sociaux.

FOAF est l'un des projets de Web Sémantique les plus importants, comme étant une ontologie pour représenter des quantités considérables de données distribuées sous une forme standard. Ainsi, elle est devenue un vocabulaire largement utilisé pour représenter les Réseaux Sociaux. Se servant du potentiel

de Web Sémantique, *FOAF* permet de fusionner des données du même utilisateur à partir de plusieurs sites de ces réseaux.

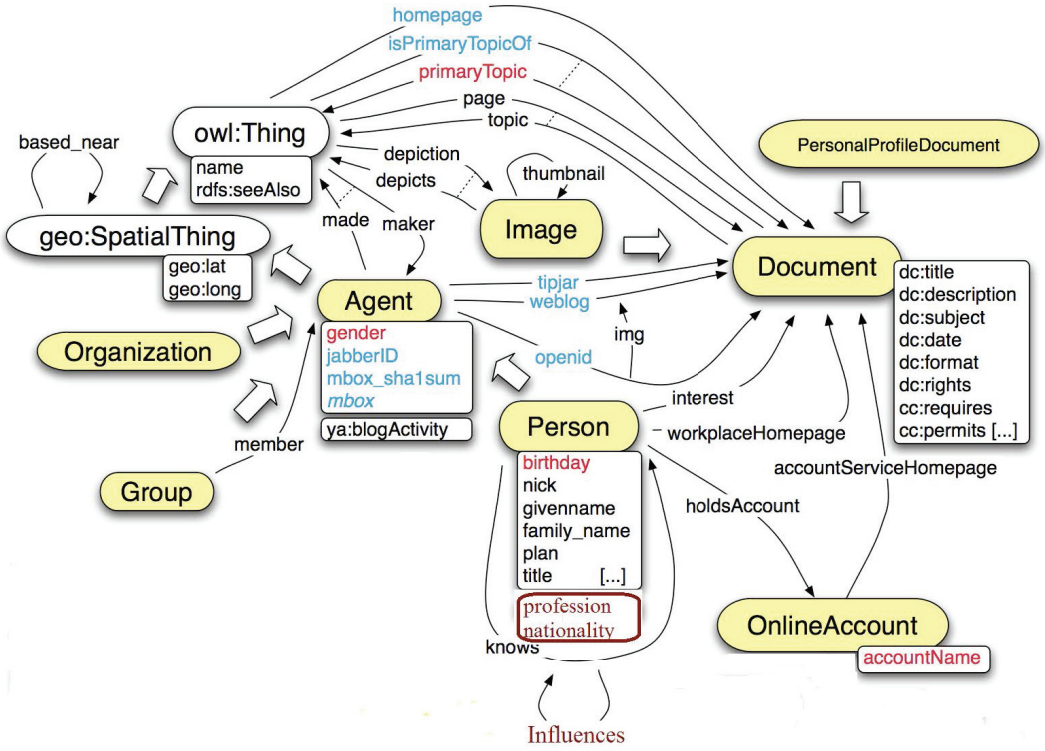

Figure 2. *Ontologie FOAF étendue*

Plus particulièrement, *FOAF* comprend, parmi d'autres, les propriétés suivantes :

— ***foaf :knows*** : permet de relier une personne à une autre qu'elle connaît indiquant un certain niveau d'interaction réciproque. Dans les Réseaux Sociaux, elle peut représenter les liens d'amitiés ou de collaboration entre les utilisateurs ;

— ***foaf :topic_interest*** : permet de relier directement une personne à un sujet d'intérêt ; signifiant qu'elle est intéressée par ce sujet ;

— ***foaf :gender*** : il s'agit généralement d'une chaîne (féminin : "Female") ou (masculin : "male") représentant le genre de la personne ;

— ***foaf :Document*** : ce sont les documents électroniques ou physiques partagés ou publiés par les utilisateurs. Chaque document est caractérisé par une propriété qui est ***foaf :topic*** décrivant son sujet d'intérêt ;

Nous avons mené une extension sur cette ontologie afin de représenter plus d'attributs des utilisateurs comme la profession et la nationalité et plus de relations sociales comme l'influence ; que nous jugeons utiles pour mener le reste de notre travail (voir figure 2).

— ***foaf :influences*** : permet de relier une personne à une autre par une relation d'influence qui est extraite à partir des données sociales ;

— ***foaf :nationality*** : c'est une chaîne qui représente la nationalité d'une personne ;

— ***foaf :profession*** : c'est une chaîne qui représente la profession d'une personne.

4. Solution générale

Symbole	Description
V	Ensemble des utilisateurs du Réseau Social
q	Sujet d'intérêt
V_q	Ensemble des utilisateurs intéressés en q
$N(v_i)$	Ensemble des voisins de v_i
$P(v_i)$	Ensemble des publications de v_i à propos de q

Tableau 1. *Notations*

Nous introduisons des notations que nous utiliserons tout au long du reste du papier (voir tableau 1).

Soit $C(q) = \{v_i \in V\}$ la communauté d'intérêt des utilisateurs intéressés en q. A partir du profilage des utilisateurs, nous en distinguons deux types : actifs et passifs.

— Les utilisateurs actifs : sont ceux qui indiquent explicitement leurs sujets d'intérêt dans l'ensemble de leurs intérêts publiés dans les Réseaux Sociaux, ainsi que le reste de données comme les attributs et les connexions.

— Les utilisateurs passifs : n'indiquent pas leurs intérêts explicitement ou les reportent partiellement mais présentent les autres données.

4.1. *Principe*

Pour répondre à notre problématique, nous proposons une approche sémantique pour la détection des communautés d'intérêt basée sur le contexte et orientée données, en considérant la topologie et la sémantique conjointement.

A partir de la modélisation des utilisateurs et leurs sujets d'intérêt, nous construisons les regroupements autour de ces sujets. Ces ensembles que nous définissons comme étant les communautés d'intérêt. Le principe de l'algorithme peut être résumé de la manière suivante :

1. Pour chaque utilisateur v_i spécifié dans le réseau (représente l'égo), si v_i est actif par rapport à q (c'est-à-dire indique explicitement q comme intérêt), v_i est intégré à $C(q)$. L'ensemble des nœuds actifs repérés ainsi que les liens entre eux forment la communauté d'intérêt ;
2. Pour chaque nœud dans les réseaux égocentriques des nœuds de la communauté qui sont ordonnés selon leurs degrés, intégrer les nœuds actifs par rapport à q à $C(q)$ et inférer parmi les nœuds passifs ceux qui pourraient être intéressés en q et les ajouter à $C(q)$;
3. pour tout nœud v_i de la communauté $C(q)$, déterminer la polarité (+ ou -) de ses sentiments par rapport à q. Ainsi deux sous-communautés $C^+(q)$ et $C^-(q)$ sont distinguées.

4.2. *Définition algorithmique*

A partir du principe précédemment décrit, nous définissons un algorithme formel qui identifie la façon dont la détection de communautés doit être effectuée. Afin de remplir les objectifs identifiés pour répondre à la problématique énoncée, nous développons notre approche en trois étapes qui sont les suivantes :

1. **Formation** : fonction définissant le pattern qui permet d'extraire de manière explicite les entités qui appartiennent à une communauté. Elle prend en paramètres, entre autres, le réseau et le sujet d'intérêt recherché correspondant à la communauté ;
2. **Evolution** : fonction définissant les règles permettant à une communauté à choisir d'intégrer de nouvelles entités. Elle prend en paramètres un nœud candidat et une communauté, et retourne soit vrai si le nœud doit être intégré à la communauté, soit faux sinon ;
3. **Division** : fonction qui garantit l'équilibre social au sein de la communauté. En effet, une communauté peut être éventuellement divisée en deux sous-communautés $C^+(q)$ et $C^-(q)$ contenant, respectivement, les nœuds qui sont intéressés positivement et ceux qui sont intéressés négativement en q.

Algorithm 1 $ENES(v_i, V, q, k)$

$C(q) = \emptyset$
$level = 0$;
$queue = (v_i)$;
while $queue \neq ()$ **do**
  $v = dequeue(queue)$;
  for $v_j \in neighbors(v)$ **do**
    if v_j not labelled as enqueued **then**
      Label v_j as enqueued ;
      $Enqueue(queue, v_j)$;
      if $v_j = EAK(v_j, q)$ *and* $v_j \notin C(q)$ **then**
        $C(q) = C(q) \cup \{v_j\}$;
      end if
    end if
  end for
end while
for $v_j \in C(q)$ **do**
  $LDC(v_j, C(q))$;
end for
for $v_j \in C(q)$ **do**
  if $LDC(v_j, C(q)) < k$ **then**
    $C(q) = C(q) \backslash \{v_j\}$;
  end if
end for
return $C(q)$;

4.3. *Procédure d'implémentation*

4.3.1. *Détails de l'étape 1*

Un algorithme de parcours du réseau égocentrique d'un nœud donné visite séqentiellement tous les nœuds de ce réseau. Lors de l'extraction explicite, un parcours en largeur ***ENES*** (**E**gocentric **N**etwork **E**xplicit **S**earch) est effectué (voir *algorithme 1*). ***ENES*** place d'abord le nœud d'origine dans une file.

A chaque itération, ***ENES*** va visiter le premier élément de la file puis placer tous ses voisins dans la file, s'ils n'y sont pas déjà.

Definition 4.1. ***EAK*** *(**E**xplicit **A**cquisition of **K**nowledge) est une fonction d'acquisition explicite de connaissances basée sur l'ontologie FOAF qui représente le profil utilisateur social. Elle extrait les utilisateurs actifs par rapport au sujet d'intérêt q.*

Cette acquisition se fait à travers la spécification de requêtes basée sur la syntaxe du protocole *SPARQL* pour sélectionner des données d'intérêt (voir le tableau 2).

Mesure	**Définition formelle *SPARQL***
$int_{<topic>}$	$SELECT\ ?name\ ?interest$ $WHERE$ $\{$ $\quad ?x\ foaf:name\ ?name$ $\quad ?x\ foaf:topic_interest\ ?interest$ $\quad FILTER\ (?interest, "topic")$ $\}$
$int_{<person,topic>}$	$SELECT\ ?name\ ?interest$ $WHERE$ $\{$ $\quad ?x\ foaf:name\ ?name$ $\quad ?x\ foaf:topic_interest\ ?interest$ $\quad FILTER\ (?interest, "topic")$ $\quad FILTER\ (?name, "person")$ $\}$
$deg_{<type,length>}(y)$	$SELECT\ ?y\ count(?x)$ $WHERE$ $\{$ $\quad ?x\ \$path\ ?y$ $\quad FILTER(match\ (\$path, star(param[type])))$ $\quad FILTER(pathLength\ (\$path) <= param[length])$ $\}$ $UNION$ $\{$ $\quad ?x\ \$path\ ?y$ $\quad FILTER(match\ (\$path, star(param[type])))$ $\quad FILTER(pathLength\ (\$path) <= param[length])$ $\}$ $GROUP\ BY\ ?y$

Tableau 2. *Définition formelle dans SPARQL des mesures paramétrées sémantiquement*

Afin d'analyser les positions des individus relativement aux autres dans le réseau, des mesures de centralités sont utilisées pour les caractériser. Parmi ces mesures, la centralité de degré est une mesure qui reflète l'activité relationnelle directe d'un individu. Elle calcule le nombre de connexions directes de chaque acteur dans le réseau entier. Celui qui détient la plus grande valeur de centralité de degré est l'acteur qui occupe la position centrale dans le réseau. L'équation de cette mesure est la suivante :

$$CD(v_i) = \frac{d(v_i)}{n-1} \tag{1}$$

Où $d(v_i)$ est le degré de nœud v_i du réseau et $n-1$ est le nombre total de connexions directes.

Nous définissons la **Centralité de Degré Locale** (*LDC*) qui représente la mesure de centralité de degré mais localement dans une communauté. Dans ce cas, $n-1$ de la formule de la centralité de degré représente le nombre total de connexions à l'intérieur de la communauté.

Les utilisateurs de $C(q)$ sont ordonnés selon leurs valeurs de centralités de degrés locales. Ceux ayant des valeurs supérieures à un certain paramètre k sont retenus dans la communauté d'intérêt et les autres sont supprimés, d'où la contraction. Cette définition d'un *k*-backbone du réseau garantit la caractéristique topologique des communautés qui stipule que le nombre de liens à l'intérieur des communautés est supérieur à celui en dehors de ces communautés.

4.3.2. *Détails de l'étape 2*

Sur le plan pratique, la détection des communautés doit s'appuyer sur une structure incrémentatle qui est mise à jour en prenant compte de nouveaux éléments. En particulier, les données explicites publiées par les utilisateurs dans les sites des Réseaux Sociaux ne sont pas assez complètes et ne peuvent pas être considérées comme entièrement connues, correctes et accessibles. Ainsi, il est difficile de s'appuyer sur la seule acquisition explicite de connaissances pour détecter les intérêts réels des utilisateurs et les classifier en communautés. Selon la solution proposée, le problème consite à déterminer les utilisateurs passifs dans le réseau égocentrique d'une personne donnée, qui peuvent être intéressés au sujet *q* en question.

Sur cette base, le problème principal devient, étant donné un utilisateur actif v_i par rapport à un intérêt *q* et un utilisateur passif v_j, déduire si v_i et v_j sont similaires donc ont le même intérêt *q* ou sont dissimilaires.

Notre solution pour résoudre ce problème consiste à concevoir un modèle de prédiction qui peut déduire la similarité d'intérêt entre les utilisateurs en fonction de leurs profils sociaux. La question qui se pose est de savoir quelles caractéristiques sociales dans ces profils déterminent la similarité d'intérêt.

En effet, notre modèle (voir la figure 3) est construit à partir de l'hypothèse que l'environnement social, et plus particulièrement, les personnes proches d'un utilisateur donné, peuvent fournir des informations pour l'inférence des intérêts de cet utilisateur. Par personnes proches, nous référons au réseau égocentrique de l'utilisateur.

Les données sociales massives alimentent l'application de l'algorithme prédictif dont le processus d'inférence des intérêts pour un utilisateur passif est procédé comme suit : comparer un utilisateur passif avec un utilisateur actif ; deux résultats possibles sont attendus : les deux utilisateurs ont un intérêt similaire ou des intérêts dissimilaires. La comparaison entre les utilisateurs se fait par rapport aux su-

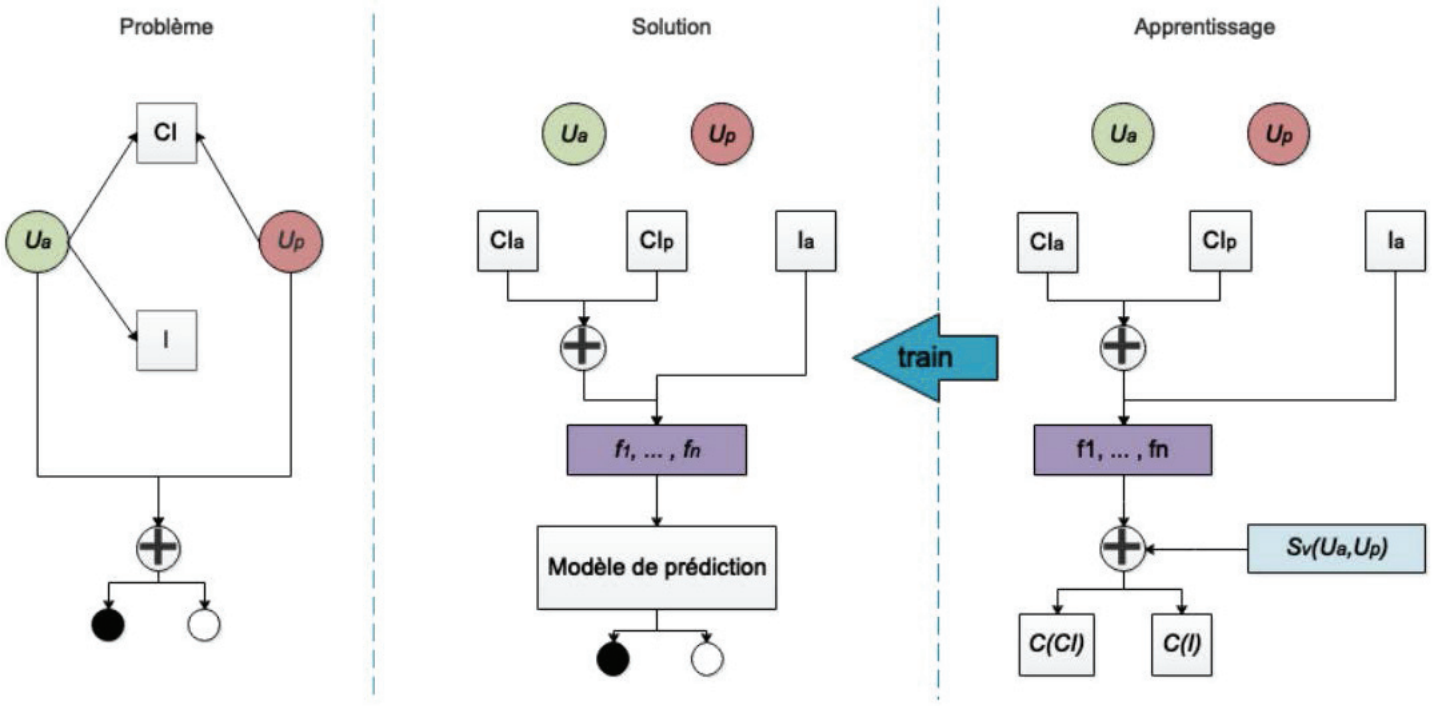

Figure 3. *Modèle de prédiction des intérêts avec : CI informations contextuelles, I les sujets d'intérêt, U_a utilisateur actif, U_p utilisateur passif, f_i caractéristiques sociales, S_v vecteur social.*

jets d'intérêt et aux informations contextuelles qui peuvent être calculées à partir des données sociales représentées dans les profils utilisateurs.

Cette étape repose sur un parcours en profondeur du réseau égocentrique d'un nœud afin d'effectuer une acquisition implicite de connaissances. ***ENIS*** (**E**gocentric **N**etwork **I**mplicit **S**earch) visite un nœud v_i puis choisit comme nœud à visiter par la suite l'un des voisins non visités de v_i (voir *algorithme 2*). Si tous les voisins de v_i ont été déjà visités, ***ENIS*** choisit l'un des voisins du nœud précédent et ainsi de suite. A chaque fois qu'un nœud visité est passif (***EAK*** retourne "Faux"), un algorithme de prédiction est appliqué afin de déterminer l'intérêt potentiel de ce nœud en q.

Algorithm 2 $ENIS(v_i, q)$

Label v_i as visited;
for $v_j \in neighbors(v_i)$ **do**
  if v_j not labelled as visited and $v_j = EAK(v_j, q)$ and $v_j \notin C(q)$ **then**
    if $EAK(v_j, q) = False$ and $v_j \notin C(q)$ **then**
      Prediction(v_j, q)
    end if
    $ENIS(v_j)$;
  end if
end for

D'après notre étude du domaine des Réseaux Sociaux, nous supposons que la similitude entre les utilisateurs peut être expliquée par deux phénomènes qui sont : l'homophilie et l'influence sociale.

Mesures de la similarité d'intérêt Une fois nous connaissons les sujets qui intéressent les utilisateurs, nous passons à mesurer la similarité d'intérêt entre les paires d'utilisateurs. Nous la quantifions sur une agrégation de ces paires par deux mesures :

— la similarité des sujets d'intérêt entre deux utilisateurs, définie comme la différence entre la probabilité que deux utilisateurs soient intéressés au même sujet :

$$sim_t(i,j) = 1 - |DT'_{it} - DT'_{jt}| \tag{2}$$

— et le degré de similarité d'intérêt qui capture les chevauchements d'intérêts défini par la moyenne des similarités d'intérêt de tous les utilisateurs :

$$s = \frac{\sum_{(u,v)\in C} sim_t(u,v)}{\|C\|} \tag{3}$$

L'homophilie : est l'une des régularités empiriques les plus marquantes et les plus robustes de la vie sociale (McPherson *et al*., 2001). En fait, elle explique à quel point les paires d'individus sont similaires en termes de certains attributs comme le genre, la profession et la nationalité. En particulier, pour chaque paire d'utilisateurs actif et passif, respectivement, v_i et v_j, leurs attributs sont utilisés pour extraire les informations contextuelles suivantes :

— Combinaison de genre (***GC***) : prend deux valeurs possibles : 1 si v_i et v_j ont le même genre et 0 sinon ;

— Combinaison professionnelle (***PC***) : mise à 1 si v_i et v_j ont la même profession et 0 sinon ;

— Combinaison de nationalité (***NC***) : mise à 1 si v_i et v_j ont la même nationalité et 0 sinon.

— Distance de connectivité (***CD***) : mesure la distance entre v_i et v_j, elle est mise à 1 s'ils ont un lien entre eux et 0 sinon ;

— Entropie d'intérêt : l'entropie mesure jusqu'à quel point les utilisateurs se concentrent sur des sujets d'intérêt. Donc, nous utilisons l'entropie pour caractériser les intérêts des utilisateurs. Généralement une entropie élevée reflète un utilisateur avec des poids d'intérêts élevés. L'entropie quantifie alors la quantité d'informations sur les intérêts d'un utilisateur à partir de deux éléments : le nombre d'intérêts et leurs poids. Le poids d'un intérêt représente sa popularité. En effet, il semble que deux utilisateurs ont plus de chance de partager un intérêt qui est très populaire. La popularité d'un sujet d'intérêt est le rapport du nombre des utilisateurs qui en sont intéressés et la popularité moyenne de tous les sujets d'intérêt :

$$h_t = \frac{N_t}{(\sum_{t=1}^{n} N_t) \div n} \tag{4}$$

Ainsi, le poids d'un sujet par rapport à sa popularité est défini par :

$$w_t = \frac{h_t}{n} \tag{5}$$

Et l'entropie d'intérêt est donc :

$$H(I_u) = -\sum_{x_i \in I_u} w(x_i) \log w(x_i) \tag{6}$$

En outre, nous supposons que la similitude d'intérêt est corrélée à l'influence sociale entre les utilisateurs dans les Réseaux Sociaux. Dans la suite, nous décrivons le modèle d'influence.

L'Influence sociale : est un phénomène complexe. Son rôle et ses effets ont été largement étudiés dans les domaines de sociologie, marketing, communication et sciences politiques. Dans l'Analyse des Réseaux Sociaux, en particulier dans l'analyse comportementale, nous nous focalisons sur l'étude des relations d'influence. Les principaux défis à prendre en compte lors de la définition du modèle informatique de l'influence sociale sont : comment différencier les influences sociales sous différents angles et comment intégrer différentes informations (distribution des sujets d'intérêt et la structure du réseau) dans un modèle unifié.

Dans notre travail, nous distinguons deux types de métriques dans l'évaluation de l'influence entre les utilisateurs :

— Les activités interpersonnelles : nous utilisons trois activités interpersonnelles dans les Réseaux Sociaux. En effet, les utilisateurs communiquent et interagissent entre eux par des commentaires, des mentions comme "j'aime" que nous désignons par "likes" en anglais, et des partages de contenus mutuels.

— Le degré entrant : c'est le nombre des abonnés d'un utilisateur reflétant sa popularité.

Afin de calculer les mesures d'influence entre deux utilisateurs, étant donné leurs données sociales, nous avons recours au modèle d'influence représenté par un graphe hétérogène normalisé d'influence (Chouchani, Abed, 2019).

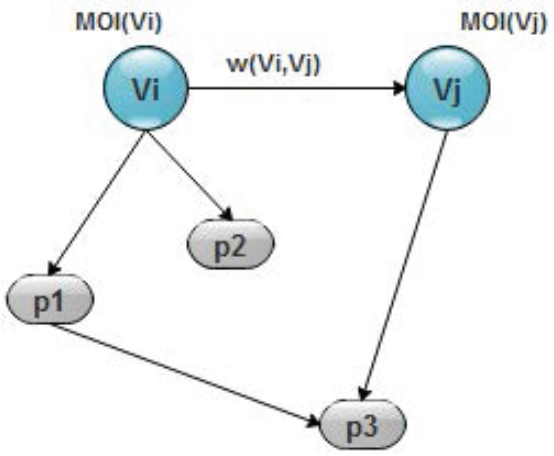

Figure 4. *Exemple de graphe d'influence (HN-DIG)*

Ce graphe, comme le montre la figure 4 copmrend les magnitudes d'influence des acteurs dans leurs réseaux égocentriques ainsi que les mesures d'influence entre eux définies par :

$$w(v_i, v_j) = \frac{ROI(v_i, v_j)}{MOI(v_i)} | v_i, v_j \in V_q$$

Où : v_i et v_j sont deux utilisateurs, ROI est le ratio d'influence et MOI la magnitude d'influence.

Méthode SVM Cette méthode est utilisée dans divers problèmes de classification en cherchant un hyperplan qui distingue deux classes tout en respectant une contrainte qui stipule que la marge entre les classes doit être maximisée. Un modèle de prédiction basé sur *SVM* revient à résoudre le problème d'optimisation suivant :

$$\min L(w) = \frac{1}{2} * \|w\| + \mu \sum_{i=0}^{i=l} \beta_i$$
$$\textit{Subject to} \begin{cases} \beta_i \geq 0 \\ h_i < w, x_i > \geq 1 - \beta_i \end{cases} \tag{7}$$

Où l est le nombre total de paires d'utilisateurs dans l'ensemble d'apprentissage, μ une constante et $\beta_i, i = 1..l$ des variables d'optimisation.

Dans notre travail, la méthode svm est utilisée pour l'apprentissage et la classification du modèle de prédiction des intérêts.

Pour chaque paire *k* d'utilisateurs, nous générons le vecteur social suivant :

$$SV_k(v_i, v_j) =< GC(v_i, v_j), PC(v_i, v_j), NC(v_i, v_j), CD(v_i, v_j), CI(v_i, v_j), infl(v_i, v_j) > \tag{8}$$

Ainsi, nous construisons le modèle de prédiction des intérêts basé sur *SVM*. Pour former ce modèle, nous générons des paires d'utilisateurs en couplant aléatoirement deux utilisateurs.

4.3.3. *Détails de l'étape 3 : Division*

Etant donné un intérêt ***q*** et un graphe hétérogène normalisé ***HNDIGq***, soit $y_i \in \{-1, +1\}$ l'étiquette définie pour chaque utilisateur et publication représentant la polarité de sentiment comme "positive" $(+1)$ ou "négative" (-1) par rapport à q. Soient Y_v le vecteur des étiquettes pour tous les utilisateurs et Y_p celui de toutes les publications.

En particulier, nous distinguons deux catégories d'utilisateurs : les utilisateurs étiquetés pour lesquels les étiquettes de polarité sont connues et les utilisateurs non étiquetés ceux dont les étiquettes de polarité sont inconnues. Étant donné la difficulté de collecter des étiquettes et l'échelle des Réseaux Sociaux, nous travaillons dans un paradigme d'apprentissage semi-supervisé. Nous supposons que seulement un petit groupe d'utilisateurs est déjà étiqueté. Ainsi, notre tâche consiste à prédire les étiquettes de polarité de tous les utilisateurs non étiquetés.

Nous définissons un modèle qui obéit à l'hypothèse de Markov impliquant que la polarité du sentiment d'un utilisateur est déterminée par les polarités de sentiment de ses publications (facteur Utilisateur-Publication) et celles de ses adjacents qui peuvent l'influencer (facteur Utilisateur-Utilisateur).

En se basant sur cette hypothèse, le modèle probabiliste défini est détaillé dans ce qui suit.

$$\begin{aligned} \log P(Y_v) = (\sum_{v_i \in V} [& \sum_{p \in P(v_i), k, l} \mu_{k,l} f_{k,l}(y_{v_i}, y_p) \\ & + \sum_{v_j \in N(v_i), k, l} \lambda_{k,l} h_{k,l}(y_{v_i}, y_{v_j})]) \\ & - \log Z \end{aligned} \tag{9}$$

Où k, l ∈ {-1, + 1} en référence aux étiquettes de sentiment, $\mu_{k,l}$ et $\lambda_{k,l}$ les paramètres d'impact, $f_{k,l}(.,.)$ La fonction qui évalue le facteur Utilisateur-Publication, $h_{k,l}(.,.)$ la fonction du facteur Utilisateur-Utilisateur, y_p l'étiquette du sentiment de la publication *p* et *Z* un facteur de normalisation.

Le facteur Utilisateur-Publication. Les publications d'un utilisateur sont censés fournir des informations sur son opinion. La fonction du facteur Utilisateur-Publication évalue la conformité entre la polarité du sentiment de la publication et le sentiment de l'utilisateur. Ceci est par rapport aux niveaux de confiance tirés des données qui sont initialement étiquetées ou non. Ces niveaux, $\tau_{labeled}$ et $\tau_{unlabeled}$, sont estimés sur la base de l'hypothèse que les étiquettes initiales sont les plus fiables, donc nous avons fixé $\tau_{labeled}$ = 1.0 et $\tau_{unlabeled}$ = 0.125. Notons que cette fonction suppose que la polarité de sentiment de chaque publication doit être classée.

$$f_{k,l}(y_{v_i}, \hat{y}_p) = \begin{cases} \frac{\tau_{labeled}}{\mid P(v_i) \mid} & y_{v_i} = k, \hat{y}_p = l, v_i : labeled \\ \frac{\tau_{unlabeled}}{\mid P(v_i) \mid} & y_{v_i} = k, \hat{y}_p = l, v_i : unlabeled \\ 0 & otherwise \end{cases} \tag{10}$$

Le facteur Utilisateur-Utilisateur. Nous admettons que les relations d'influence sociale entre les utilisateurs peuvent être corrélées avec la similarité des sentiments. La fonction du facteur Utilisateur-Utilisateur évalue la conformité du sentiment d'un utilisateur avec l'opinion de son voisin en se référant à leurs rapports sociaux d'amitié et d'influence.

$$h_{k,l}(y_{v_i}, y_{v_j}) = \begin{cases} \frac{\tau_{relation}}{\mid N(v_i) \mid} + \frac{\tau_{influence}}{\mid N(v_i) \mid} \times \frac{1}{1 - IR(v_j)} & y_{v_i} = k, y_{v_j} = l \\ 0 & otherwise \end{cases} \tag{11}$$

Ces facteurs sont estimés directement à partir de statistiques simples en utilisant les dénombrements des données étiquetées ou non.

Jusqu'à présent, il reste à estimer les valeurs optimales des paramètres $\mu_{k,l}$ et $\lambda_{k,l}$ afin que l'attribution de l'étiquette de polarité d'un utilisateur maximise $logP(Y_v)$. Pour l'apprentissage de ces paramètres, nous utilisons l'algorithme SampleRank (Wick *et al.*, 2009).

Nous visons à apprendre ces paramètres en maximisant $logP(Y)$ en fonction des paramètres $\mu_{k,l}$ et $\lambda_{k,l}$. Pour ce faire, nous utilisons le modèle d'Analyse des Sentiments au niveau utilisateur basé sur les phénomènes d'influence et d'homophilie (Chouchani, Abed, 2019).

5. Résultats expérimentaux et évaluation

Dans cette section, nous réalisons des expériences sur des données de Réseaux Sociaux réels. Nous commençons par décrire l'échantillon des données étudié, quelques statistiques d'observation, puis nous présentons l'analyse de performance.

5.1. *Description de l'échantillon des données*

Nous avons collecté un échantillon de données à partir des sites des Réseaux Sociaux *Facebook*[2] et *Twitter*[3] .

2. https ://www.facebook.com
3. https ://www.twitter.com

Le tableau 3 montre les statistiques de base de toutes les données collectées sur les sujets sélectionnés dans différents domaines : politique et musique (*Donald Trump*, *Hilary Clinton* et *Lady Gaga*). Notons que les informations démographiques des utilisateurs sont extraites de *Facebook*.

Sujet d'intérêt	# utilisateurs	# abonnés	# commentaires	# likes	# amis	# Retweets	# publications
Trump	140	852186	160	8555	259676	1281	328
Clinton	140	338224	176	6375	334979	1626	578
Gaga	140	580656	112	6004	574564	2306	526

Tableau 3. *Statistiques de notre échantillon de données*

Notre objectif est de trouver un grand nombre d'utilisateurs dont les polarités de sentiment sont claires, afin que les labels de référence soient fiables. Nous avons sélectionné un ensemble de profils de célébrités du monde politique et musical, et un ensemble d'utilisateurs qui leur sont opposés. Nous avons manuellement annoté les polarités de sentiment des utilisateurs et leurs publications correspondantes.

Dans la figure 5, nous décrivons la distribution des publications positives et négatives sur les différents sujets d'intérêt.

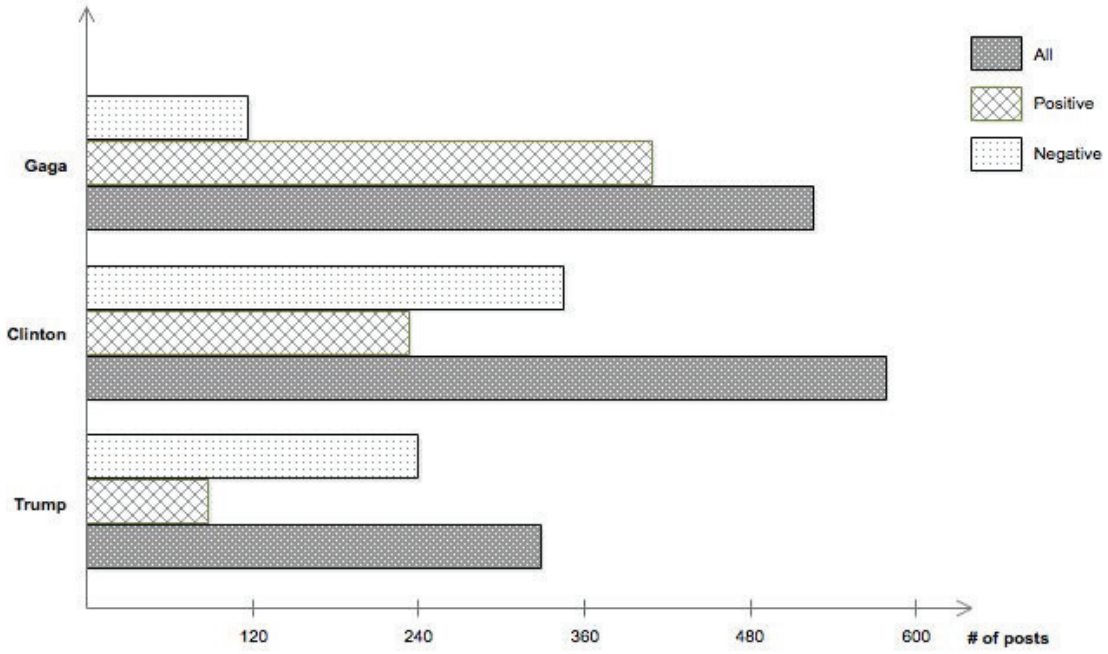

Figure 5. *Distributions des publications positives et négatives*

5.2. *Statistiques d'observation*

5.2.1. *Corrélation entre les caractéristiques sociales et la similarité des intérêts*

Les données collectées contiennent les informations démographiques, les intérêts des utilisateurs et les relations sociales. Une fois que les données collectées provenant de différentes sources sont intégrées, nous générons des paires d'utilisateurs et les classons en deux classes : la classe "0" où les utilisateurs ont des intérêts différents et la classe "1" d'utilisateurs partageant le même intérêt.

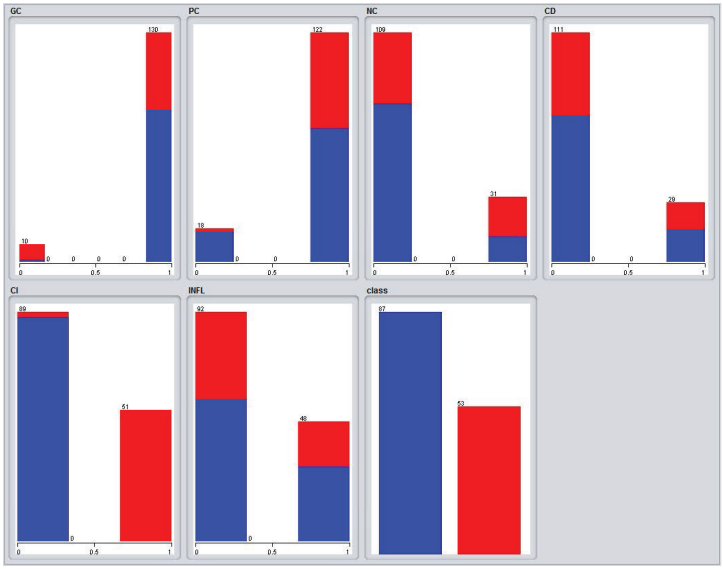

Figure 6. *Corrélation entre les caractéristiques sociales et la similarité d'intérêt*

Dans la figure 5, nous décrivons la corrélation entre les différentes caractéristiques du vecteur social (défini dans l'équation 8). Les zones rouges contiennent des utilisateurs partageant les mêmes intérêts et les zones bleues, des utilisateurs aux intérêts dissemblables. Les résultats obtenus montrent que la caractéristique «intérêts communs» est la plus corrélée à la similitude des intérêts. Par conséquent, nous concluons que plus les utilisateurs ont des intérêts communs, plus ils pourraient être similaires sur un sujet d'intérêt donné.

5.2.2. *Corrélation entre influence sociale et similarité de sentiments*

Afin d'expliquer la corrélation entre influence sociale et similarité de sentiments, nous avons concentré notre étude sur des données extraites de Twitter, avec des sujets d'intérêt qui sont "Trump" et "Clinton".

Nous avons défini deux types de statistiques pour étudier l'interaction entre la polarité de l'étiquette du sentiment d'un utilisateur et les relations d'influence. Ces statistiques sont les suivantes :

— Probabilité que deux utilisateurs influencés l'un par l'autre soit conditionnée par la même polarité de sentiment : cette statistique mesure l'influence conditionnée sur les étiquettes. La figure 7 montre que le sentiment partagé tend à impliquer une influence. En fait, dans les graphiques résultants, les utilisateurs ont plus de chances d'avoir une relation d'influence s'ils partagent une opinion que s'ils diffèrent.

— Probabilité que deux utilisateurs aient la même polarité de sentiment, conditionnée par l'influence réciproque ou non : la seconde statistique mesure les sentiments partagés conditionnés par une relation d'influence. La figure 8 montre que la probabilité que deux utilisateurs soient influencés l'un par l'autre, partageant le même sentiment sur un sujet donné, est beaucoup plus élevée que le hasard.

En résumé, les paires d'utilisateurs dans lesquelles au moins l'un influence l'autre ont tendance à avoir le même sentiment et deux utilisateurs partageant le même sentiment sont plus susceptibles d'être influencées l'un par l'autre que deux utilisateurs ayant des sentiments différents.

Ces observations valident notre hypoyhèse selon laquelle influence et sentiment partagé sont clairement corrélés.

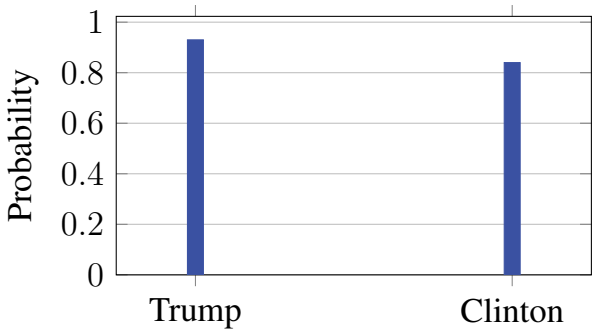

Figure 7. *Probabilité d'influence conditionnée ou non par la même polarité de sentiment*

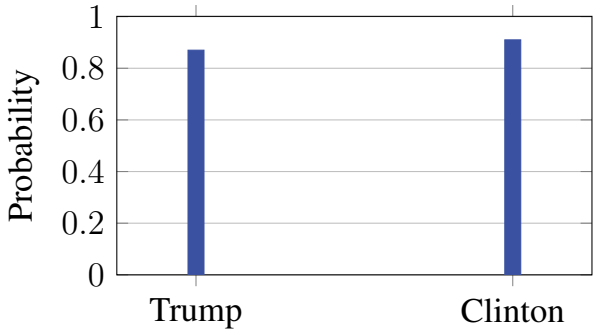

Figure 8. *Probabilité d'avoir la même polarité de sentiment conditionnée par une relation d'influence*

5.3. *Analyse de performance*

Afin de prendre en compte le fait que l'algorithme SampleRank est randomisé du fait de sa dépendance à la fonction d'échantillonnage, nous avons effectué des inférences k (k in 1,3,5,11) fois pour obtenir k prédictions. L'idée est de conserver un vote à la majorité (prédiction) parmi les k étiquettes possibles.

Nous faisons des expériences 10 fois. À chaque fois, les données avec les étiquettes de vérité sont divisées en un ensemble d'entraînement et un ensemble d'évaluation. Le premier ensemble est composé de 50 utilisateurs positifs et de 50 utilisateurs négatifs, choisis au hasard. Le deuxième ensemble est constitué des utilisateurs étiquetés restants. Le rapport des deux ensembles est différent dans différents sujets d'intérêt.

Dans le cadre du modèle, nous avons besoin de l'annotation des étiquettes de tweet obtenue en exécutant la méthode proposée dans (Vo, Zhang, 2015)

Nous comparons deux méthodes de classification des utilisateurs, dont notre proposition, afin d'évaluer nos résultats.

— Modèle de graphe hétérogène (Tan *et al.*, 2011) : il effectue un apprentissage semi-supervisé sur le graphe hétérogène représentant les utilisateurs, les connexions mutuelles et leurs publications. Ensuite, il applique la propagation de croyances en boucle pour obtenir des étiquettes de sentiment au niveau utilisateur.

— Modèle de graphe d'influence hétérogène : nous effectuons notre apprentissage semi-supervisé sur le graphe d'influence hétérogène pour obtenir la classification des utilisateurs.

Afin d'évaluer les résultats obtenus, nous introduisons les résultats de performance pour les différentes méthodes considérées. Nous évaluons les performances en utilisant la précision (P), le rappel (R) et le score F1 (F1) sur chaque sujet.

$$P = \frac{TP}{TP + FP} \tag{12}$$

$$R = \frac{TP}{TP + FN} \tag{13}$$

$$F1 = \frac{2PR}{P + R} \tag{14}$$

Ensuite, l'exactitude (accuracy) est mesurée à l'aide de ces mesures. Son équation est le suivant :

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{15}$$

Où TP est le nombre de vrais positifs, FP le faux positif, FN le faux négatif et TN le vrai négatif en termes de prédictions.

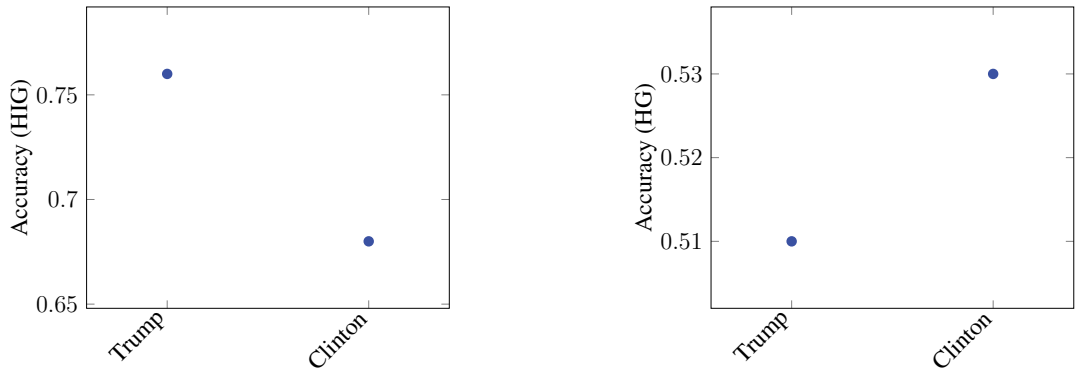

Figure 9. *Mesures d'exactitude pour les deux méthodes*

Les résultats confirment l'amélioration de la performance de la classification des sentiments. La figure 8 montre ces résultats pour les deux méthodes et indique que notre modèle obtient les mesures d'exactitude les plus élevées.

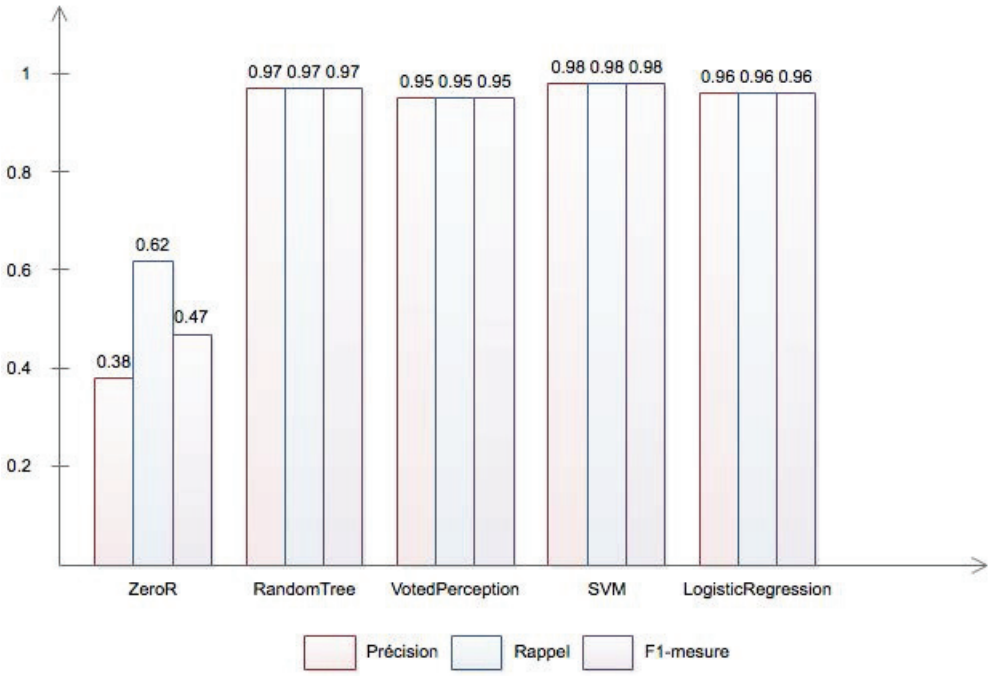

Figure 10. *Comparaison des mesures de présicion, rappel et F1*

La seconde expérience a pour objectif d'évaluer les performances du modèle de prédiction des intérêts. Nous avons effectué la tâche de prédiction en utilisant différentes méthodes d'apprentissage automatique standard. Cinq algorithmes de classification ont été sélectionnés. Ils ont été choisis de manière semi-aléatoire pour leur diversité, leur représentation et leur style d'apprentissage. L'entraînement de ces méthodes se fait par aggrégation de paires d'utilisateurs actifs et passifs parmi l'échantillon des données pour former le vecteur social. Nous avons mené cette expérience et nous avons obtenu les résultats présentés dans la figure 10.

Les résultats obtenus indiquent clairement que la méthode *SVM* fonctionne bien sur les mesures d'exactitude.

6. Conclusion

Dans ce papier, nous avons abordé le problème de la détection des communautés d'intérêt basée sur les Réseaux Sociaux. Jusqu'à présent, la plupart des travaux se sont concentrés sur la structure du réseau plutôt que sur la sémantique. Cependant, notre approche considère conjointement la topologie structurelle et la sémantique, y compris les attributs et les différents types de relations sociales. Il s'agit tout d'abord de proposer un modèle générique du profil utilisateur social pour répondre au problème. Le modèle proposé comprend deux dimensions : la dimension utilisateur et la dimension sociale. Nous ne considérons qu'une partie significative du Réseau Social autour de l'utilisateur : c'est son réseau égocentrique. Il peut être judicieux d'opter pour une représentation d'ontologie afin de prendre en compte différents types de données sociales existantes. Deuxièmement, nous proposons une approche de détection de communautés d'intérêts exploitant le modèle de profil utilisateur social ainsi construit. Cette approche est composée de trois phases. Une phase de "formation" au cours de laquelle une extraction explicite de connaissances dans le réseau égocentrique d'un utilisateur est effectuée. Ensuite, une phase d ' "évolution" est basée sur une extraction implicite de connaissances. Nous avons proposé plus précisément des modèles de prédiction des intérêts des utilisateurs et de détermination des mesures d'influence entre eux. Enfin, une phase de "division", exploitant un graphe d'influence hétérogène que nous avons proposée pour déterminer les polarités des sentiments des utilisateurs de la communauté à l'égard du sujet d'intérêt en question. Il en résulte deux sous-communautés contenant des utilisateurs qui s'intéressent de manières positive et négative à ce sujet.

Plusieurs perspectives sont possibles concernant l'approche proposée, les expériences et les évaluations. En effet, il serait intéressant de prendre en compte l'aspect dynamique des communautés d'intérêts. De plus, nous visons à extraire plus de données des Réseaux Sociaux pour obtenir des résultats plus précis. Les communautés détectées peuvent également être exploitées pour développer des applications personnalisées ainsi que des systèmes de recommandation.

Bibliographie

Barabasi A.-L., Albert R. (1999, October). Emergence of scaling in random networks. , vol. 286, p. 509–512.

Barbieri N., Bonchi F., Manco G. (2013). Cascade-based community detection. In S. Leonardi, A. Panconesi, P. Ferragina, A. Gionis (Eds.), *Wsdm*, p. 33-42. ACM.

Blei D. M., Ng A. Y., Jordan M. I. (2003, mars). Latent dirichlet allocation. *J. Mach. Learn. Res.*, vol. 3, p. 993–1022.

Blondel V., Guillaume J., Lambiotte R., Mech E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech*, p. P10008.

Bothorel C., Cruz J. D., Magnani M., Micenkova B. (2015). Clustering attributed graphs : models, measures and methods. *Network Science*, vol. 3, n° 3, p. 408–444.

Chouchani N., Abed M. (2018, 16 Aug). Online social network analysis : detection of communities of interest. *Journal of Intelligent Information Systems*.

Chouchani N., Abed M. (2019, 05 Feb). Enhance sentiment analysis on social networks with social influence analytics. *Journal of Ambient Intelligence and Humanized Computing*. Consulté sur https://doi.org/10.1007/s12652-019-01234-0

Clauset A., Newman M. E. J., Moore C. (2004). Finding community structure in very large networks. *Physical Review E*, vol. 70, p. 066111.

Combe D., Largeron C., Egyed-Zsigmond E., Gery M. (2012). Getting clusters from structure data and attribute data. In *Asonam*, p. 710-712. IEEE Computer Society.

Crandall D., Cosley D., Huttenlocher D., Kleinberg J., Suri S. (2008). Feedback effects between similarity and social influence in online communities. *KDD '08 : Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 160–168.

Ereteo G., Gandon F., Buffa M. (2011). Semtagp : Semantic community detection in folksonomies. In O. Boissier, B. Benatallah, M. P. Papazoglou, Z. W. Ras, M.-S. Hacid (Eds.), *Web intelligence*, p. 324-331. IEEE Computer Society.

Faloutsos M., Faloutsos P., Faloutsos C. (1999, Aug-Sept.). On power-law relationships of the Internet topology. *SIGCOMM*, p. 251-262.

Fayyad U., Piatetsky-Shapiro G., Smyth P. (1996, novembre). The kdd process for extracting useful knowledge from volumes of data. *Commun. ACM*, vol. 39, n° 11, p. 27–34.

Garton L., Haythornthwaite C., Wellman B. (1997). Studying online social networks. *Journal of Computer-Mediated Communication*, vol. 3, n° 1, p. 0–0.

Girvan M., Newman M. E. J. (2002, June). Community structure in social and biological networks. *PNAS*, vol. 99, n° 12, p. 7821-7826.

Kernighan B., Lin S. (1970). An efficient heuristic procedure for partitioning graphs. *The Bell Systems Technical Journal*, vol. 49, n° 2.

Li H., Nie Z., Lee W.-C., Giles C. L., Wen J.-R. (2008). Scalable community discovery on textual data with relations. In J. G. Shanahan *et al.* (Eds.), *Cikm*, p. 1203-1212. ACM.

McPherson M., Smith-Lovin L., Cook J. M. (2001). Birds of a feather : Homophily in social networks. *Annual Review of Sociology*, vol. 27, p. 415-444.

Natarajan N., Sen P., Chaoji V. (2013). Community detection in content-sharing social networks. In J. G. Rokne, C. Faloutsos (Eds.), *Asonam*, p. 82-89. ACM.

Newman M. E. J. (2004). Fast algorithm for detecting community structure in networks. *Physical Review*, vol. E 69, n° 066133.

Palla G., Derenyi I., Farkas I., Vicsek T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, vol. 435, p. 814-818.

Pang B., Lee L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, vol. 2, n° 1-2, p. 1–135.

Tan C., Lee L., Tang J., Jiang L., Zhou M., Li P. (2011). User-level sentiment analysis incorporating social networks. In C. Apte, J. Ghosh, P. Smyth (Eds.), *Kdd*, p. 1397-1405. ACM.

Tchuente D., Canut M.-F., Jessel N. B., Péninou A., Sèdes F. (2012, avril). Visualizing the relevance of social ties in user profile modeling. *Web Intelli. and Agent Sys.*, vol. 10, n° 2, p. 261–274.

Vo D.-T., Zhang Y. (2015). Target-dependent twitter sentiment classification with rich automatic features. In *Proceedings of the 24th international conference on artificial intelligence*, p. 1347–1353. AAAI Press. Consulté sur http://dl.acm.org/citation.cfm?id=2832415.2832437

Wick M., Rohanimanesh K., Culotta A., McCallum A. (2009). Samplerank : Learning preferences from atomic gradients. In N. I. P. S. N. W. on Advances in Ranking (Ed.), *booktitle.*

Xu Z., Ke Y., Wang Y., Cheng H., Cheng J. (2012). A model-based approach to attributed graph clustering. In K. S. Candan, Y. Chen, R. T. Snodgrass, L. Gravano, A. Fuxman (Eds.), *Sigmod conference*, p. 505-516. ACM.

Yang J., McAuley J., Leskovec J. (2013). Community detection in networks with node attributes. In *Data mining (icdm), 2013 ieee 13th international conference on*, p. 1151–1156.

Zhou D., Manavoglu E., Li J., Giles C. L., Zha H. (2006). Probabilistic models for discovering e-communities. In L. Carr, D. D. Roure, A. Iyengar, C. A. Goble, M. Dahlin (Eds.), *Www*, p. 173-182. ACM.

Zhou Y., Cheng H., Yu J. X. (2009). Graph clustering based on structural/attribute similarities. *Proceedings of the VLDB Endowment*, vol. 2, n° 1, p. 718–729.

