Une approche de détection des communautés d'intérêt dans les réseaux sociaux: application à la génération d'IHM personnalisées

Nadia Chouchani

Une approche de détection des communautés d'intérêt dans les réseaux sociaux : application à la génération d'IHM personnalisées

Nadia Chouchani

Université Polytechnique HAUTS-DE-FRANCE

Thèse de doctorat

Pour obtenir le grade de Docteur de

l'UNIVERSITE POLYTECHNIQUE HAUTS-DE-FRANCE

Discipline : **INFORMATIQUE**

Présentée et soutenue par Nadia, CHOUCHANI.

Le 07/12/2018, à *Valenciennes*

Ecole doctorale : Sciences Pour l'Ingénieur (ED SPI 072)

Equipe de recherche, Laboratoire : LAMIH - UMR 8201

Une approche de détection des communautés d'intérêt dans les Réseaux Sociaux : application à la génération d'IHM personnalisées

JURY

Rapporteurs

- MOLLI, Pascal. Professeur, Université de Nantes
- EGYED-ZSIGMOND, Elöd. Maître de Conférences, HdR, INSA Lyon

Examinateurs

- MARCAL DE OLIVEIRA, Kathia. Maître de Conférences, HdR, UPHF
- CHAMROUKHI, Faicel. Professeur, Université de Caen (président du jury)

Invité

- ARTIBA, Abdelhakim. Professeur, UPHF

Directeur de thèse

- ABED, Mourad. Professeur, Université Polytechnique Hauts-de-France

Résumé

De nos jours, les Réseaux Sociaux sont omniprésents dans tous les aspects de la vie. Une fonctionnalité fondamentale de ces réseaux est la connexion entre les utilisateurs. Ces derniers sont engagés progressivement à contribuer en ajoutant leurs propres contenus. Donc, les Réseaux Sociaux intègrent également les créations des utilisateurs ; ce qui incite à revisiter les méthodes de leur analyse. Ce domaine a conduit désormais à de nombreux travaux de recherche ces dernières années. L'un des problèmes principaux est la détection des communautés.

Les travaux de recherche présentés dans ce mémoire se positionnent dans les thématiques de l'analyse sémantique des Réseaux Sociaux et de la génération des applications interactives personnalisées. Cette thèse propose une approche pour la détection des communautés d'intérêt dans les Réseaux Sociaux. Cette approche modélise les données sociales sous forme d'un profil utilisateur social représenté par un ontologie. Elle met en oeuvre une méthode pour l'Analyse des Sentiments basées sur les phénomènes de l'influence sociale et d'Homophilie. Les communautés détectées sont exploitées dans la génération d'applications interactives personnalisées. Cette génération est basée sur une approche de type *MDA*, indépendante du domaine d'application. De surcroît, cet ouvrage fait état d'une évaluation de nos propositions sur des données issues de Réseaux Sociaux réels.

Mots clés : Réseaux Sociaux, Communauté d'intérêt, Profil utilisateur, Ontologie, Analyse des Sentiments

Abstract

Nowadays, Social Networks are ubiquitous in all aspects of life. A fundamental feature of these networks is the connection between users. These are gradually engaged to contribute by adding their own content. So Social Networks also integrate user creations ; which encourages researchers to revisit the methods of their analysis. This field has now led to a great deal of research in recent years. One of the main problems is the detection of communities.

The research presented in this thesis is positioned in the themes of the semantic analysis of Social Networks and the generation of personalized interactive applications. This thesis proposes an approach for the detection of communities of interest in Social Networks. This approach models social data in the form of a social user profile represented by an ontology. It implements a method for the Sentiment Analysis based on the phenomena of social influence and homophily. The detected communities are exploited in the generation of personalized interactive applications. This generation is based on an approach of type *MDA*, independent of the application domain. In addition, this manuscript reports an evaluation of our proposals on data from Real Social Networks.

Key words : Social Networks, Community of interest, User profile, Ontology, Sentiment Analysis

Remerciements

Merci à tous ceux qui m'ont aidé au cours des trois dernières années.

Tout d'abord, je voudrais remercier mon directeur de thèse, le professeur Mourad ABED, pour son soutien durant mon doctorat. Sans lui rien n'aurait été possible. Il a su, à moult reprises, m'encourager et me motiver.

J'adresse mes vifs remerciements à M. Pascal MOLLI, professeur à l'Université de Nantes, et à M. Elöd EGYED-ZSIGMOND, maître de Conférences, HdR, à l'INSA de Lyon, d'avoir accepté de juger mon travail et m'avoir faite l'insigne honneur d'être rapporteurs de mon mémoire de thèse.

Mes remerciements vont également à Mme Kathia MARCAL DE OLIVEIRA, maître de conférences, HdR, à l'Université Polytechnique des Hauts de France, et M. Faicel CHAMROUKHI, professeur à l'Université de Caen, d'avoir bien voulu examiner ce travail de thèse.

Je tiens également à exprimer ma gratitude à tous les membres de notre laboratoire *LAMIH*, que j'ai eu l'honneur de connaître.

Un grand merci à tous mes amis en France, vous étiez et vous resterez toujours ma deuxième famille.

Enfin, je ne saurais terminer sans remercier mes parents et toute ma famille qui ont toujours été au plus près de moi en veillant à ma réussite.

Table des matières

Table des figures

Liste des tableaux

INTRODUCTION

Sommaire

I Contexte et problématique

La vertu civique de la seconde guerre mondiale a contribué à l'apogée de la vie associative et de la démocratie participative. Ces domaines ont connu par la suite un déclin prononcé, qui a été expliqué par les transformations du capital social. "L'idée centrale de la théorie du capital social est que les Réseaux Sociaux ont de la valeur. (..) Le capital social se rapporte aux relations entre individus, aux Réseaux Sociaux et aux normes de réciprocité et de confiance qui en émergent" [Putnam, 2002]. De ce fait, plus la vertu civique est insérée dans un réseau dense de relations sociales, plus elle est efficace. La connectivité sociale apparaît donc comme une ressource dont la richesse réside dans la densité et la qualité des relations, qui est différente de celle que les individus seuls, peuvent s'approprier. Ces relations obéissent aux règles de la vie sociale. Elles produisent en conséquence le capital social qui profite autant aux individus qu'aux communautés. Alors, comment la densité des réseaux tissés entre ses membres et le degré de partage des intérêts expliquent-ils et résument-ils le capital social, c'est-à-dire les richesses détenues par une communauté d'individus ? La réponse à cette question permet d'extraire des ressources riches en informations dans le contexte d'un Réseau Social. Ceci constitue, désormais, un défi sans précédent et une occasion pour déterminer des connaissances qui peuvent être exploitables dans une variété de domaines tels que le marketing, la santé, les sciences sociales voire la défense.

Les Réseaux Sociaux en ligne ont connu un nouvel élan grâce à la prolifération croissante des objets connectés. Effectivement, ce phénomène n'est pas récent ; mais cette révolution est passée à la vitesse supérieure impulsée par l'IoT (Internet of Things). Quelques exemples de ces réseaux sont Twitter[1], Facebook[2] et LinkedIn[3]. Au cours des deux dernières décennies, ils ont gagné en popularité car ils ne sont plus bornés par les limites géographiques des Réseaux Sociaux conventionnels. Leur infrastructure a soutenu une énorme explosion de données centrées sur le réseau et dans une grande variété de scénarios. Assurément, cette ère des Réseaux Sociaux donne un essor innovant à la société numérique telle que nous la connaissons. Les utilisateurs y participent davantage au moyen de la production et du partage des données. Ces dernières sont de deux types : (i) les données de liaison qui sont essentiellement la structure du réseau et les communications entre les entités ; et (ii) les données de contenu qui contiennent les textes, les images et d'autres données multimédia.

Concrètement et afin de mesurer le capital social, certains aspects sociaux doivent être considérés tels que la vitalité des structures, les comportements et les attitudes des individus. C'est dans ce contexte que l'Analyse des Réseaux Sociaux a suscité de nombreux travaux. Ils se sont concentrés essentiellement sur les aspects sociaux, structurels et cognitifs du Réseau Social ; et moins concentrés sur les problèmes qui se posent dans le contexte de l'interaction entre les aspects structurels et les données sémantiques du réseau. De prime abord, ces travaux ont montré que la densité d'interaction est plus importante pour les entités qui font partie des mêmes sous-groupes. Subséquemment, cette organisation a été popularisée sous le nom de structure communautaire, un nom tiré de l'observation de ce type de structure dans les Réseaux Sociaux. Afin de la définir approximativement, un certain consensus a été établi stipulant que la structure communautaire est une décomposition d'un système complexe en sous-ensembles d'éléments densément connectés avec ou sans recouvrement entre eux. Cette densité peut être justifiée par le phénomène d'Homophilie [McPherson et al., 2001], qui traduit la tendance qu'ont les individus à se lier avec d'autres individus aux caractéristiques similaires. Dans le cas où des personnes se rassemblent autour d'un sujet d'intérêt commun, il s'agit de ce que nous appelons une communauté d'intérêt. La détection de cette structure est, néanmoins, problématique. Ainsi, et compte tenu de l'importance du problème, la quantité des algorithmes de détection de communautés a explosé. Il est donc rapidement devenu difficile de comparer ces algorithmes dans un contexte de liberté donnée autant pour la définition de la structure recherchée que pour l'approche utilisée pour l'extraire. Cependant, il devient de plus en plus clair qu'une

1. http ://www.twitter.com
2. http ://www.facebook.com
3. http ://www.linkedin.com

approche unifiée combinant les informations sur le contenu avec l'analyse structurale du réseau est nécessaire pour progresser dans ce domaine. Cette intégration n'a pas encore été étudiée à fond et soulève de nouvelles questions liées à la façon de tirer le meilleur parti de l'ensemble des données disponibles. En fait, ce problème est au coeur des efforts récents pour **détecter des communautés topologiquement bien connectées et sémantiquement cohérentes et significatives**. Vu sous cet angle, de nouvelles approches s'intéressent à la détection de communautés d'un double point de vue : la topologie du réseau et la sémantique des données. En ce qui concerne la sémantique, la mine de données fournie par les Réseaux Sociaux peut être exploitée grâce aux techniques de fouille de données (Data Mining). Ce domaine offre des techniques d'exploitation de données nécessaires pour analyser des données sociales volumineuses, complexes et qui changent fréquemment. En particulier, ces plateformes contiennent beaucoup de textes. Ces derniers sont porteurs des opinions de leurs auteurs qui sont souvent convaincantes et peuvent servir de base aux choix et décisions d'autres personnes dans le réseau. La découverte et la reconnaissance des expressions positives et négatives des opinions des utilisateurs sur divers sujets d'intérêt sont définies comme étant l'Analyse des Sentiments. Cependant analyser automatiquement les sentiments des textes est toujours difficile parce qu'ils sont, éventuellement, courts et contiennent de l'argot, des expressions informelles, des émoticônes, des erreurs typographiques et beaucoup de mots introuvables dans les dictionnaires. Ces caractéristiques influencent négativement la performance de leur classification. Afin de pallier ce problème, certaines tentatives ont exploité les données sur les interactions des utilisateurs dans les plateformes des Réseaux Sociaux dans le but d'améliorer la classification des sentiments. En prime, de telles interactions peuvent conduire les différents acteurs à s'influencer mutuellement en termes de comportement. Dès lors, les contributions pionnières d'anthropologues comme Simmel, confirment que parmi les mécanismes qui rendent possible l'interaction entre les individus c'est le phénomène psychosociologique de l'influence. Il est considéré comme l'origine des actions réciproques. "Les hommes influent les uns sur les autres, les uns font ou souffrent quelque chose, présentent telle manière d'être ou de devenir parce que d'autres sont là et s'expriment, agissent ou éprouvent des sentiments" [Simmel, 1908]. L'analyse de l'influence sociale vise à mesurer qualitativement et quantitativement l'influence d'une personne sur une autre. Comme les Réseaux Sociaux deviennent plus répandus dans les activités de millions de personnes au jour le jour, la recherche et les applications sur l'influence sociale continuent à se développer. En effet, la place centrale qu'occupe l'influence dans la sociologie des groupes impose d'y accorder un grand intérêt.

Concisément, un problème crucial dans l'analyse des Réseaux Sociaux et des données comportementales est de former des groupes d'utilisateurs ayant des intérêts similaires

sous la forme de communautés. Ces structures peuvent être exploitées pour fournir des indications sur la conception et l'amélioration des services sociaux avec une signification pratique. En effet, s'il devient plus facile de publier des contenus sur les plateformes des Réseaux Sociaux, l'accès à ces contenus est cependant rendu plus difficile aux utilisateurs compte tenu du nombre de plus en plus important et de la diversité des informations susceptibles de les intéresser. Ceci pose, en général, des problèmes de surcharge cognitive à l'utilisateur qui aura davantage du mal à retrouver les informations correspondant à ses attentes. Conséquemment, si la croissance des données dans les plateformes sociales constitue une riche source d'informations pour la détection des communautés, c'est également un défi de recherche intéressant qui émerge dans de nombreux cas. Afin de surmonter ces défis, les opportunités de Big Data sont exploitées dans l'analyse des Réseaux Sociaux en améliorant les applications et les services basés sur ces plateformes. Ceci favorise le déploiement de systèmes de plus en plus personnalisés pour enrichir et multiplier les expériences des utilisateurs. En effet, une fois structurées dans les communautés d'intérêt, les données sociales peuvent être exploitées pour développer des fonctionnalités qui adaptent les informations en fonction des besoins des utilisateurs.

II Objectifs de la thèse

La brève introduction précédente permet à peine d'entrevoir la complexité et la richesse du problème de détection communautaire et de son exploitation dans diverses applications. En effet, malgré la taille toujours grandissante des travaux traitant cette problématique, plusieurs problèmes ouverts subsistent. Nous verrons dans cette thèse dans quelle mesure nous pouvons améliorer le processus de détection de communautés et l'exploiter dans la génération des applications sociales personnalisées. Nos objectifs sont multiples :

1. Pour ce qui est de notre travail, nous nous sommes intéressés à définir une nouvelle approche pour la détection communautaire dans les Réseaux Sociaux en considérant les données structurales et les contenus conjointement.

    - Il est clair que pour pouvoir coupler les deux types d'informations, il est nécessaire de définir un modèle sémantique des données sociales ; dans ce cadre il faut bien définir les données sociales. L'idée est de construire un profil utilisateur "social" générique et extensible, qui peut être exploité par la suite dans diverses applications ;
    - Les données étudiées comprennent les utilisateurs ainsi que leurs contenus. Il est donc indispensable de représenter des réseaux hétérogènes et en profi-

ter pour déduire les éventuelles corrélations entre ces types de données. En effet, nous distinguons les liens de type utilisateur-utilisateur et ceux de type utilisateur-contenu. En particulier, nous visons à étudier le phénomène d'influence sociale afin de le modéliser et le considérer dans la détection communautaire et l'Analyse des Sentiments. Cette dernière sera effectuée grâce à un modèle de prédiction des sentiments au niveau d'un utilisateur ;

- Proposer une méthode pour l'Analyse des Sentiments en exploitant le modèle d'influence afin de montrer la corrélation entre les polarités des sentiments et les relations d'influence ;
- Mettre en place un modèle prédictif des intérêts des utilisateurs en appliquant les techniques prédictives sur les volumes de données rendus énormes grâce à la participation des utilisateurs dans la génération des contenus dans les plateformes des Réseaux Sociaux ;

2. Il est également intéressant d'exploiter les communautés des utilisateurs détectées pour la génération des applications sociales personnalisées.

   - Au lieu de la personnalisation individuelle, proposer une méthode de personnalisation collective basée sur les communautés d'intérêt ;
   - La génération des applications sociales est basée sur une approche de type *MDA* (*Model Driven Architecture*) en définissant les *DSL*s (*Domain Specific Language*) pour la détermination des données pertinentes à partir des flux de données dans les Réseaux Sociaux et pour la synthèse de réponses personnalisées aux utilisateurs selon leurs communautés d'intérêt.

La méthodologie de recherche utilisée dans cette thèse est décrite dans la figure .1.

III Guide de lecture

Cette thèse est divisée en cinq chapitres.
Le chapitre 1 présente le contexte de travail en introduisant quelques définitions et terminologies. Nous abordons les Réseaux Sociaux et leurs analyses, leurs modes de représentations, leurs caractéristiques ainsi que les communautés d'intérêt. En outre, nous présentons deux notions clés de cette thèse : l'Analyse des Sentiments et les applications sociales ainsi que leur génération automatique. Ce faisant, nous aurons une traçabilité plus claire des notions utilisées et une identification des concepts potentiellement utiles pour notre contribution.

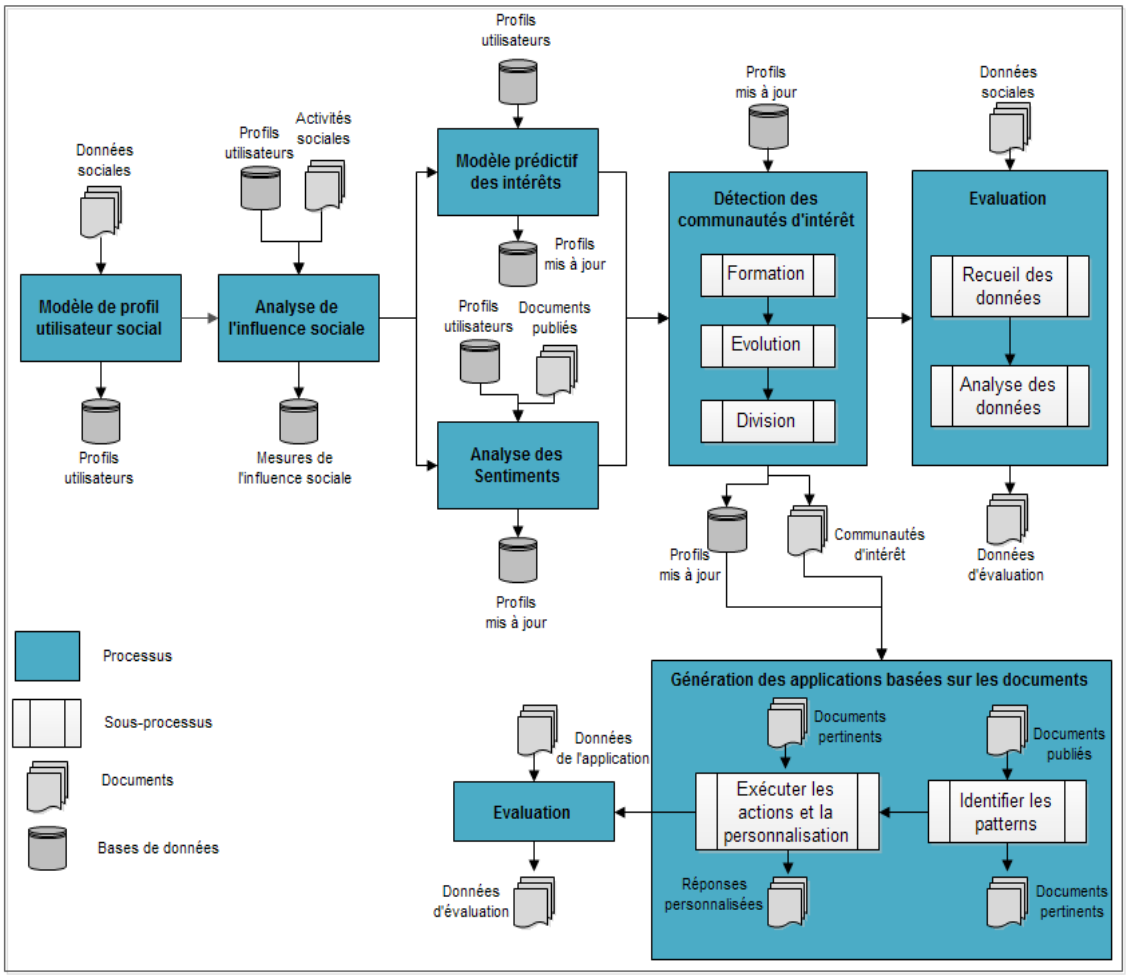


FIGURE .1 – Méthodologie de recherche

Dans le chapitre 2, nous présentons une analyse critique des travaux de littérature liés aux problématiques de détection communautaire, Analyse des Sentiments et génération des applications sociales. Afin de bien étudier ces travaux et pour positionner notre travail parmi eux, nous répertorions des critères sur lesquels se base notre classification. Cette étude est suivie d'une synthèse permettant de classifier les méthodes de détection communautaire étudiées selon différentes perspectives. Nous nous sommes intéressés à mettre l'accent sur leurs forces et leurs faiblesses ; et compte tenu des limites rencontrées, nous présentons par la suite notre proposition.

Le chapitre 3 expose les contributions apportées pour la détection communautaire. Une approche centrée utilisateur est définie en se basant sur une définition d'un profil utilisateur social. Cette approche est composée d'algorithmes de dérivation de connaissances explicite et implicite à partir des profils construits. Plus particulièrement, elle est composée de 3 étapes qui sont : Formation, Evolution et Division. Nous détaillons également un

modèle prédictif des intérêts et un modèle d'influence sociale. Plus particulièrement, la dernière étape repose sur une méthode d'Analyse des Sentiments que nous proposons en profitant du modèle d'influence.

Dans le chapitre 4, nous présentons l'architecture *MDA* de notre approche dont le but est de prendre en compte la personnalisation du contenu dans la conception et la génération semi-automatique des applications sociales. Le modèle du profil utilisateur social et les communautés d'intérêt constituent les éléments clés de cette approche. En effet, il s'agit d'une personnalisation collective en fonction des communautés auxquelles un utilisateur appartient. Nous présentons les différentes étapes à suivre en utilisant des *DSL*s appropriés.

Finalement, le chapitre 5 concerne la mise en pratique de nos contributions en proposant des études de cas permettant de détecter des communautés d'intérêt dans les Réseaux Sociaux et les exploiter dans la personnalisation d'une application sociale.

Ce mémoire se termine par une conclusion qui, en revenant sur les grandes thématiques qui nous aurons guidés tout au long de cette lecture, porte sur le bilan de notre recherche, sur l'ensemble des contributions apportées par cette thèse et finalement ses limites. Cette conclusion donne également l'occasion d'exprimer les perspectives de nos travaux de recherche.

CHAPITRE 1

Notions préliminaires

Sommaire

Ce chapitre est consacré à la présentation d'un panorama sur les Réseaux Sociaux ainsi que leurs modes de représentation, propriétés et différents phénomènes sous-jacents, les communautés d'intérêt, les méthodes d'analyse du Big Data Social et la génération des applications sociales basées sur ces réseaux. En particulier, dans la première section nous donnons un aperçu sur les Réseaux Sociaux et leur analyse en vue de la détection des communautés d'intérêt. La deuxième section explorera ces réseaux comme étant des vecteurs du Big Data. Alors, elle présente diverses méthodes et applications pour l'analyse de ce qui est appelé Big Data Social. La dernière section quant à elle est dédiée à la description des systèmes sociaux interactifs basés sur ces réseaux et leur génération.
L'ensemble de ce chapitre présente donc la terminologie utilisée et les concepts fondamentaux nécessaires au positionnement de nos travaux et à la compréhension du présent mémoire.

1.1 Les Réseaux Sociaux

De nos jours, les Réseaux Sociaux (*RS*) sont devenus une partie très importante et omniprésente de notre vie quotidienne. Ce terme de *RS* est souvent utilisé pour référer à des différents services en ligne qui sont associés à une catégorie générale de situations d'interactions sociale et professionnelle. Alors qu'est-ce qu'un Réseau Social ? Comment est-il représenté ? Quelles sont ses caractéristiques et propriétés ? Cette section tente de répondre à toutes ces questions.

Les Réseaux Sociaux sont des services basés sur le web dont la fonctionnalité principale est de connecter des personnes ou des entités. Ils sont définis, selon Garton et al. [Garton et al., 1997], 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". [Wasserman and Faust, 1994] les décrivent comme des ensembles finis d'acteurs et les relations définies entre ces acteurs. En général, ils permettent aux individus : (i) de construire un profil public ou semi-public dans un système délimité, (ii) d'articuler une liste d'autres utilisateurs avec lesquels ils partagent une connexion, et (iii) de voir et de parcourir leur liste de connexions et celles d'autres personnes dans le système [Boyd and Ellison, 2007]. Ces médias sont devenus le composant central du Web 2.0 [Auvinen, 2012]. Un grand nombre d'applications de médias sociaux existe sous diverses formes qui peuvent être classées de multiples façons. Ils existent des réseaux dits généralistes comme Facebook, Twitter et MySpace[4] conçus pour discuter. D'autres sont destinés au partage comme Youtube[5] ou Flicker[6]. Nous retrouvons aussi des réseaux de services, politiques ou encore professionnels.

Afin de comprendre le fonctionnement du monde social au sein de ces plateformes, l'Analyse des Réseaux Sociaux, parfois appelée analyse néo-structurale, appréhende les logiques relationnelles entre les acteurs sociaux à des échelles micro ou méso. Elle est fondée sur une vision structurale s'attachant à décrire les interdépendances entre les acteurs afin de simplifier leur représentation. La capacité à représenter de façon simplifiée la complexité d'un système social représente la force de cette analyse structurale. En effet, les Réseaux Sociaux sont des systèmes complexes ayant de nombreux éléments en interaction qui sont essentiellement les utilisateurs, les communautés et les contenus générés.

4. http ://www.myspace.com/
5. http ://www.youtube.com/
6. http ://www.flickr.com/

1.1.1 Modes de représentation

Jacob Levy Moreno était l'un des pionniers ayant proposé une représentation en construisant des sociogrammes des affinités ou des rejets entre des personnes [Moreno, 1933]. Dans cette représentation visant à modéliser les interactions sociales, les acteurs sont désignés par des points et les relations entre eux par des flèches. [Harary et al., 1965] ont proposé une meilleure formalisation des sociogrammes, au sens mathématique, grâce à la théorie des graphes. Cette représentation graphique est un outil puissant d'analyse. En effet, l'application de la théorie des graphes à l'Analyse des Réseaux Sociaux a permis une meilleure représentation graphique et qualification des propriétés structurales.
Dans ce qui suit, nous introduisons les notions utilisées par la théorie des graphes pour l'Analyse des Réseaux Sociaux :

- Un **noeud** ou **sommet**, **entité** ou **acteur**, est l'unité de base d'un réseau pouvant désigner différents objets. Souvent, ils représentent des personnes, des organisations ou bien des groupes ;
- Une **arête** ou **lien** est une connexion entre deux noeuds ;
- Un **arc** est une arête orientée ;
- Une arête (ou arc) est **pondérée** lorsqu'un poids lui est attribuée ;
- Un **chemin** est une séquence d'arêtes reliant deux sommets, sa longueur est le nombre des arêtes ;
- Le **degré** d'un noeud est le nombre d'arêtes ou arcs qui lui sont incidents ;
- L'**ordre** d'un graphe est le nombre de sommets qui le constituent, soit noté N ;
- La **taille** d'un graphe est le nombre de liens ou d'arcs, soit noté P.

Définition 1 (Graphe)**.** *Un graphe $G = (V, E)$ est un ensemble V de N noeuds (de l'anglais vertices) et un ensemble E de P liens (edges) connectant les noeuds entre eux. Les noeuds sont référés à l'aide de la notation v_i, v_j... Et les liens sont référés à l'aide de la notation $e_{i,j}$, qui dénote un lien connectant les noeuds v_i et v_j.*

Les relations entre les sommets, telles que les relations d'amitié ou de collaboration, sont représentées par des arêtes. Nous parlons également d'arcs lorsque les liens sont orientés. Donc, il est possible de distinguer des graphes orientés et des graphes non orientés selon le fait de prendre en compte ou pas la direction des liens. Notons que les graphes

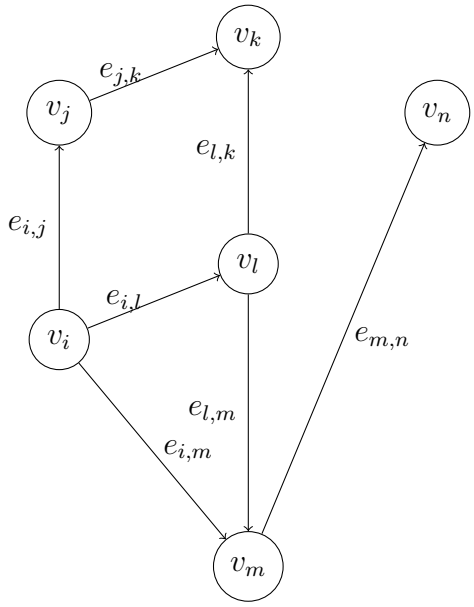


FIGURE 1.1 – Graphe orienté et pondéré d'un Réseau Social constitué de six noeuds

peuvent tout de même être pondérés en attribuant des poids aux arêtes ou bien aux arcs. La figure 1.1 montre un graphe orienté et pondéré.

Néanmoins, les Réseaux Sociaux sont souvent des graphes d'ordre et taille élevés. Par conséquent, une modélisation par matrices a été adoptée afin de pouvoir mener des calculs plus poussés sur ces réseaux comme le calcul de certains indicateurs tels que la densité et la centralité [Aggarwal, 2011]. Différents types de matrices sont distingués : les matrices d'adjacence, les matrices d'incidence, les matrices des degrés, et les matrices Laplaciennes.

Définition 2 (Matrice d'incidence)**.** *La matrice d'incidence M est une matrice binaire* $N \times P$ *où l'élément* $m_{i,\alpha}$ *de M est égal à* 1 *si l'arc* α *admet le sommet i comme origine, et* 0 *autrement.*

Définition 3 (Matrice d'adjacence)**.** *La matrice d'adjacence A est une matrice* $N \times N$ *dont les colonnes et les lignes sont associées aux noeuds du graphe* $G = (V, E)$*. L'élément* a_{ij} *de cette matrice représente le poids de l'interaction entre les noeuds* v_i *et* v_j*.*

Définition 4 (Matrice des degrés)**.** *La matrice des degrés est une matrice diagonale avec le degré des nœuds du graphe sur les diagonales.*

Définition 5 (Matrice Laplaciennes)**.** *La Matrice Laplacienne d'un graphe (ou matrice de Kirchhoff) est la différence entre sa matrice des degrés et sa matrice d'adjacence.*

Plus récemment, et afin de prendre en compte les caractéristiques des acteurs du réseau et de leurs relations, une généralisation de la notion de Réseaux Sociaux a conduit à la définition de celle de réseaux d'informations [Sun and Han, 2012]. Ceci a conduit à la définition de graphes d'information [Moser et al., 2007] ou encore appelés graphes d'attributs [Zhou et al., 2009]. Ces attributs peuvent être quantitatifs ou qualitatifs, structurés ou non structurés. Généralement, il s'agit d'un vecteur de données textuelles, numériques ou de n'importe quel type. Ils représentent les propriétés des noeuds comme les informations personnelles des individus (sexe, âge, nationalité, profession, etc.) ou d'autres caractéristiques d'un groupe et de leurs relations. Un exemple de ces réseaux est celui des bases de données bibliographiques professionnelles (telle que *DBLP*[7]), des éditeurs (telle que *ACM Digital Library*[8]) ou celles crées par des utilisateurs (telle que *Mendeley*[9]). Ils peuvent modéliser des auteurs de publications scientifiques ayant des attributs temporels comme les années de publication ou des attributs textuels référant à leurs intérêts en termes de domaines de recherche.

De surcroît, une myriade de données est constamment générée. En effet, les Réseaux Sociaux ont accru la participation des utilisateurs à la production et au partage de données. La nécessaire prise en compte d'autres méthodes sémantiques est par conséquent de plus en plus évoquée. En fait, des méthodes étendues d'Analyse des Réseaux Sociaux exploitent des modèles et formalismes du Web Sémantique pour représenter les réseaux et les interactions sociales. Effectivement, l'analyse sémantique des Réseaux Sociaux présente des intérêts indéniables vu la richesse des données sociales. Ces dernières incluent non seulement les données relationnelles mais aussi des données personnelles comme les intérêts des personnes impliquées, leurs informations démographiques ainsi que leurs activités et publications. Outre les représentations canoniques déjà évoquées, des représentations enrichies sont donc fournies afin de tirer profit de la sémantique assurant des meilleures expressivité et flexibilité. Les chercheurs ont mis beaucoup d'efforts dans le développement de nombreux modèles génériques. Toutefois, les sites des Réseaux Sociaux représentent les données à leurs propres formats. Par conséquent, des standards permettant de décrire les données, les consolider et connecter leur sémantique sont requis, ainsi que des protocoles communs pour y accéder. Dans cette optique, les modèles du Web Sémantique répondent au problème de la représentation et de l'échange de données sur le web avec un riche modèle de graphe (*RDF*), des modèles de définition de schéma (*RDFS* et *OWL*) et un protocole avec un langage de requête pour accéder aux données (*SPARQL*).

7. http ://dblp.uni-trier.de/
8. http ://dl.acm.org/
9. http ://www.mendeley.com/

- ***RDF*** : (***R***esource ***D***escription ***F***ramework) est un format de données sous forme de graphe dirigé pour représenter des informations sur le Web, permettant de décrire de façon naturelle la majorité des données traitées par les machines [Berners-Lee et al., 2001]. Il s'agit d'un modèle conceptuel standard d'informations mises en œuvre dans les ressources disponibles sur le Web. La principale caractéristique de ce modèle est de faciliter la fusion de données avec différents schémas sous-jacents. Ces ressources sont décrites à l'aide de triplets : "sujet", "prédicat" et "objet". Le sujet représente la ressource à qualifier ; le prédicat étant une relation bien définie qui peut être appliquée au sujet pour un type d'objet donné. Le sujet est qualifié à travers du prédicat pour former l'objet qui peut être une autre ressource ou un scalaire en fonction de la définition du prédicat lui-même.

- ***RDFS*** : (***RDF S***chema) complète *RDF* en fournissant un vocabulaire de base pour décrire les ressources et les structurer afin d'en construire des ontologies. La notion de "Class" permet de décrire des concepts ou objets à instancier. De plus, le prédicat "subClassOf" permet d'ordonner les différentes instances entre elles.

- ***OWL*** : (**W**eb **O**ntology **L**anguage) constitue une extension des deux vocabulaires précédents. Il s'agit d'un langage de modélisation de données pour décrire les données *RDF* en y apportant plus d'expressivité lors de la conception d'ontologies. Ce langage permet d'exprimer les relations entre les objets en utilisant différents vocabulaires. En particulier, il spécifie les manières plus précises d'exprimer les dépendances ou cardinalités entre deux ressources. Ces dernières sont caractérisées en utilisant deux mécanismes : 1) "data properties" permettant de spécifier les propriétés scalaires des ressources ; et 2) "object properties" spécifiant les relations entre les concepts, pourvu que leurs types correspondent au domaine et rang de la relation. Par exemple, *OWL* permet de créer des ontologies décrivant les réseaux de taxonomies et de classifications.

- ***SPARQL*** : (**S**PARQL **P**rotocol **A**nd **R**DF **Q**uery **L**anguage) est le langage de requêtes sémantiques pour les graphes *RDF* et un protocole d'accès aux données. Ce langage exprime des requêtes sur diverses sources de données. Les résultats de ces requêtes peuvent être des ensembles de résultats ou des graphes *RDF*. Il définit différents protocoles pour envoyer les requêtes et leurs résultats sur le web.

Dans ce contexte, des modèles sont proposés en se basant sur la modélisation ontologique, basée sur des standards coordonnés par *W3C*[10] (*World Wide Web Consortium*), utilisant des représentations enrichies de la description des ressources *RDF* pour l'analyse sémantique des interactions sociales.

Définition 6 (Ontologie)**.** *Une ontologie est un ensemble de primitives représentatives permettant de modéliser un domaine de connaissance. Les primitives de représentation sont typiquement des classes, des attributs (ou des propriétés) et des relations (ou des relations entre les membres de classe). Les primitives de représentation incluent des informations sur leur signification et des contraintes sur leur application logiquement cohérente [Gruber, 2009].*

Les ontologies permettent de représenter formellement les connaissances, de décrire le raisonnement sur ces connaissances et de les partager et les réutiliser par plusieurs applications [Staab and Studer, 2004]. Elles incluent généralement une organisation hiérarchique des concepts pertinents (principes, idées, catégorie d'objet, notions potentiellement abstraites) et des relations qui existent entre ces concepts ainsi que des règles et axiomes qui les contraignent.

Solution	**Information démographique**	**Relations**	**Contenus**	**Comptes des utilisateurs**	**Intérêts des utilisateurs**
FOAF [Brickley and Miller, 2004]	x	x	x	x	x
SIOC [Breslin et al., 2005]	-	-	x	x	x
GUMO [Heckmann et al., 2005]	x	-	-	-	-
Relationship [Davis and Vitiello]	-	x	-	-	-

TABLE 1.1 – Modèles de Web sémantique pour la représentation des Réseaux Sociaux

Plusieurs ontologies existent déjà pour représenter les Réseaux Sociaux en ligne, comme le montre le tableau 1.1. En particulier, et afin de consolider les données sociales hétérogènes, un modèle de représentation unifiée est nécessaire. Dans cette vue, *FOAF*

10. http ://www.w3.org

semble être l'ontologie de représentation des Réseaux Sociaux la plus exhaustive (voir *tableau* 1.1). En outre, elle permet d'exporter les réseaux pour les utiliser par d'autres services et applications. *FOAF* décrit les personnes, les liens entre elles, ce qu'elles créent et ce qu'elles font. Dans cette ontologie, les utilisateurs sont représentés par des instances de "*FOAF :Person*". Chaque instance (utilisateur) est décrite par un large ensemble de propriétés (voir *figure* 1.2).

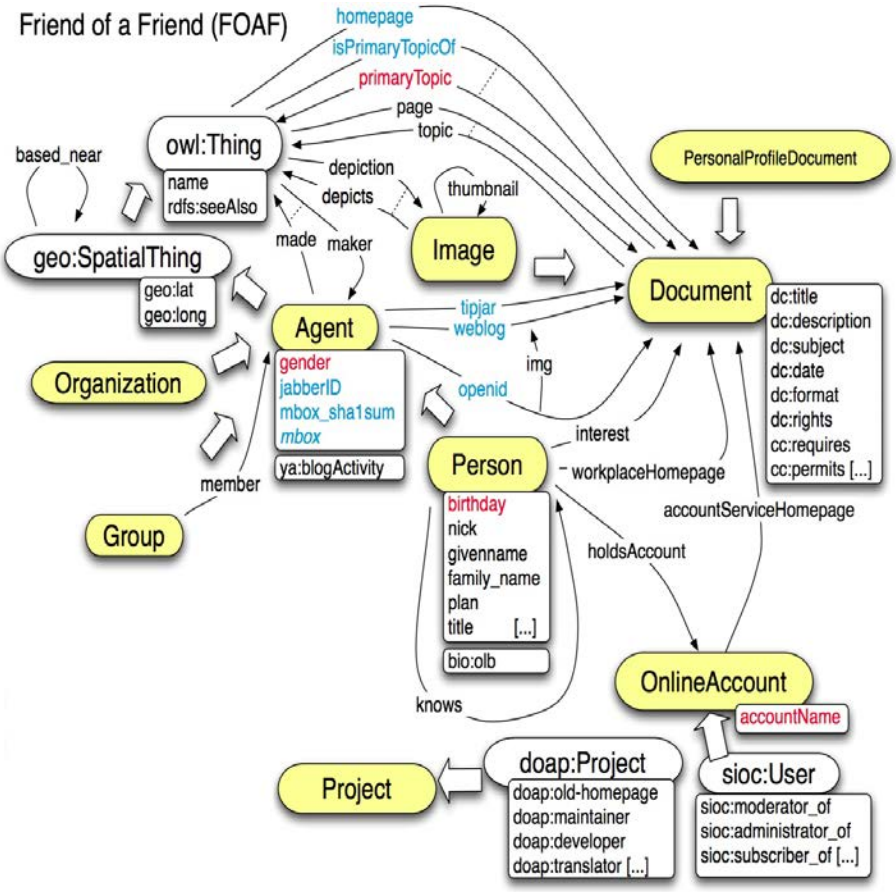


FIGURE 1.2 – Ontologie FOAF

1.1.2 Propriétés des Réseaux Sociaux

L'étude des Réseaux Sociaux a révélé que ceux-ci partagent des propriétés formelles communes telles que l'effet de «petits mondes», la distribution des degrés en loi de puissance et la structure communautaire.

1.1.2.1 Effet de "petits mondes"

C'est *Stanley Milgram*, qui en 1967, a développé son hypothèse stipulant que chacun est relié à n'importe quel individu par une courte chaîne de relations sociales [Milgram, 1967]. Ce phénomène a été testé dans le contexte de données de la messagerie de *MSN* [Leskovec and Horvitz, 2008]. Comme une vérification de cette règle de "*six degrés de séparation*", il a été démontré que la longueur moyenne d'un chemin entre deux utilisateurs est de 6.6. Cette règle a prouvé l'idée que les individus peuvent se connaître directement ou indirectement en se connectant par des connaissances communes. Le plus court chemin entre deux sommets dans un réseau de taille N est de l'ordre de $log(N)$. Donc, la longueur des plus courts chemins n'augmente que très peu lorsque la taille du réseau augmente. Ce phénomène semble être certainement correct, puisque les statistiques sur les paramètres des graphes ont montré que les diamètres de la plupart des Réseaux Sociaux sont relativement faibles. Le diamètre est défini comme la distance la plus longue parmi tous les plus courts chemins possibles.

1.1.2.2 Distribution des degrés en loi de puissance

Dans les Réseaux Sociaux, ils existent beaucoup de sommets de faibles degrés et très peu de sommets de forts degrés. En effet, la distribution des sommets suit une loi de puissance $p(k) \sim k^{\gamma}$, où k est le degré et γ l'exposant de la loi de puissance. Cet exposant est en pratique entre 2 et 3 et représente la vitesse de décroissance de la courbe des degrés. Plus γ est grand, plus la probabilité d'obtenir des sommets de forts degrés est petite.

1.1.2.3 Structure communautaire

Pour une meilleure compréhension des Réseaux Sociaux, les chercheurs ont essayé de trouver plus de caractéristiques structurelles de ces réseaux. [Barabasi and Albert, 1999] et [Faloutsos et al., 1999] ont montré que les réseaux réels ne sont pas des graphes aléatoires car ils présentent de grandes hétérogénéités, révélant un haut niveau d'ordre et d'organisation. En outre, la répartition des arêtes n'est pas seulement globale, mais aussi localement non homogène, avec des concentrations élevées dans des groupes spéciaux de sommets et de faibles concentrations entre ces groupes (voir *figure* 1.3). Cette caractéristique des réseaux réels s'appelle la structure communautaire [Girvan and Newman, 2002]. En effet, les graphes représentant les Réseaux Sociaux du monde réel présentent une structure modulaire avec des nœuds formant des groupes et éventuellement des groupes au sein de groupes. [Armstrong and Hagel, 1999] distinguent quatre types de communautés en ligne : les communautés de transaction, les communautés de relations, les communautés

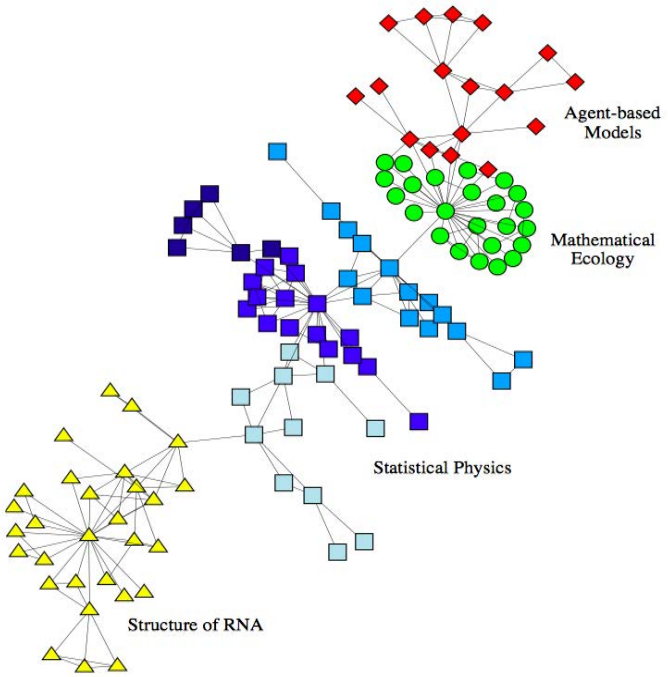


FIGURE 1.3 – Exemple d'un Réseau Social de collaboration entre scientifiques : il s'agit du graphe de collaboration de scientifiques de différentes disciplines (représentées par différentes formes) de l'institut Santa Fe aux Etats Unis. Le graphe contient 271 noeuds représentant les scientifiques de l'institut durant les deux années 1999 et 2000. Un lien est mis entre deux auteurs s'ils sont co-auteurs d'au moins un article de recherche durant ces deux années. Chaque communauté représente l'ensemble des chercheurs d'une discipline donnée [Girvan and Newman, 2002]

de fantaisie et les communautés d'intérêt. Ces dernières semblent homogènes de point de vue des intérêts partagés entre ses membres. En effet, "La socialisation est la forme qui se réalise suivant d'innombrables manières différentes, grâce à laquelle les individus, en vertu d'intérêts – sensibles ou idéaux, momentanés ou durables, conscients ou inconscients, causalement agissant ou téléologiquement stimulants - se soudent en une unité au sein de laquelle ces intérêts se réalisent" [Simmel, 1917]. L'analyse structurale en sociologie, et selon [Degenne and Forsé, 2004], tente de comprendre comment les structures sociales émergent des interactions tout en prouvant que ces dernières subissent des contraintes exercées par les structures. La décomposition d'un Réseau Social en communautés suppose que les noeuds appartenant à la même communauté ont au moins autant de liens entre eux qu'avec les autres noeuds appartenant à d'autres communautés. D'autre part, de point de vue sémantique, les noeuds partagent les mêmes sujets d'intérêts. Ces derniers sont définis comme suit.

Définition 7 (Sujet d'intérêt)**.** *Un sujet d'intérêt ou encore appelé thématique est un sujet unique clairement identifiable dans un ou plusieurs documents.*

Dans notre travail, nous considérons les définitions suivantes des communautés d'intérêts.

Définition 8 (Définition structurale)**.** *Une communauté est un ensemble de noeuds fortement liés entre eux et faiblement liés aux autres noeuds.*

Cette définition est basée sur la structure du Réseau Social. L'idée est que le nombre d'arêtes à l'intérieur des communautés est supérieur à celui en dehors de ces communautés.

Définition 9 (Définition sémantique)**.** *Une communauté est un ensemble de noeuds qui partagent le même sujet d'intérêt.*

Une communauté est constituée d'un ensemble d'individus qui ont des affinités communes ou interagissent plus souvent entre eux qu'avec les autres tissant ainsi des liens plus forts. Pour Simmel, une communauté implique une certaine égalité, un attribut commun [Simmel, 1917]. Donc les nœuds qui forment des communautés sont plus étroitement connectés entre eux que par les noeuds extérieurs à la communauté.

Au cours des dernières années, de nombreux chercheurs se sont intéressés au problème de détection de communautés. L'intérêt de cette détection est multiple, et nous pouvons citer à titre d'exemple les applications suivantes :

- **Recommandation** : L'identification des communautés de clients ayant des intérêts similaires dans les réseaux de vente et achat entre clients et produits, comme par exemple Amazon[11], permet de développer des systèmes de recommandation plus efficaces [Reddy et al., 2002], dans le but de mieux répondre aux besoins des clients et améliorer les opportunités d'affaires.

- **Prédiction de liens** : Il s'agit de déduire quelles nouvelles interactions entre les membres d'un Réseau Social qui sont susceptibles de se produire dans un proche avenir [Liben-Nowell and Kleinberg, 2003]. Des informations sur les interactions futures peuvent être extraites de la topologie du réseau, en particulier, la structure communautaire [Valverde-Rebaza and Andrade Lopes, 2014].

- **Propagation des épidémies** : Il s'avère crucial de comprendre la manière dont une épidémie se développe une fois qu'elle est apparue afin de pouvoir la contrôler.

11. http ://amazon.com/

Les modèles récents de la propagation des épidémies prennent en compte les distances effectives induites par les moyens de communication en plus des distances géographiques. Certains modèles ont montré que la structure communautaire est un déterminant important du comportement des processus de percolation sur les réseaux tels que la propagation d'une épidémie. En effet, la structure de la communauté peut à la fois imposer et inhiber les processus de diffusion [Stegehuis et al., 2016].

- **Diffusion de l'information** : Une caractéristique majeure des Réseaux Sociaux est la diffusion d'information, telles que les rumeurs, les nouvelles et les opinions. En effet, les processus de cette diffusion sont désormais affectés par la structure communautaire [Lin et al., 2015]. Plus spécifiquement, la détection de communautés dynamiques liées à des sujets tendances reliant les créateurs de contenu et ceux qui les diffusent, permet une diffusion rapide de l'information [Recalde, 2017].

- **Prédiction des fonctions cellulaires** : En biologie, les réseaux d'interaction protéine-protéine (*PIN*) se caractérisent par une organisation modulaire remarquable qui reflète les associations fonctionnelles entre les protéines. Ainsi, un groupe de protéines qui collaborent sur la même fonction cellulaire correspondent aux communautés. La détection de ces communautés et l'analyse des *PIN* s'avèrent donc être un outil précieux pour la prédiction fonctionnelle [Brun et al., 2003].

- **Détection des organisations criminelles** : La compréhension des hiérarchies au sein des organisations criminelles et la découverte des membres qui jouent un rôle central sont nécessaires pour soutenir les organismes d'application de la loi. Dans ce contexte, l'étude des réseaux criminels utilisant des traces de communication à partir d'enregistrements d'appels téléphoniques révèle la structure de la communauté sous-jacente qui sera exploitée par les enquêteurs judiciaires [Ferrara et al., 2014].

- **Analyse combinatoire des données** : Certaines méthodes de Recherche Opérationnelle pour l'analyse combinatoire des données visent à regrouper un grand nombre de caractéristiques de population (appelées modèles) dans un petit nombre de familles représentatives. Ces familles sont identifiées comme des communautés dans un graphe où les sommets sont les motifs [Darlay et al., 2010].

Dans ce qui suit, nous essayons de comprendre les origines de formation des communautés d'intérêts. Notons que le comportement du réseau dépend de l'interaction humaine. L'information est diffusée par les communautés qui connectent des personnes partageant

les mêmes idées. Ces groupes contiennent des personnes qui échangent des pensées et qui partagent des intérêts communs. “Similarity breeds connection” [McPherson et al., 2001]. En d’autres termes, la formation des communautés repose sur la similitude des acteurs sociaux. Les gens sont semblables à leurs voisins dans un Réseau Social pour deux raisons distinctes : d’abord, ils ressemblent à leurs amis actuels en raison d’une influence sociale, et deuxièmement, ils ont tendance à former de nouveaux liens vers d’autres qui sont déjà comme eux. Parmi les acteurs, il y a des influents qui influencent leurs voisins et qui ont un effet sur le comportement des autres. L’influence sociale est alors une autre forme d’interaction humaine qui explique la formation des communautés.

1.1.3 Homophilie

Il a été démontré que l’homophilie est omniprésente dans les Réseaux Sociaux [McPherson et al., 2001]. L’analyse des propriétés structurelles de ces réseaux met en évidence ce concept. Il s’agit de l’un des phénomènes fondamentaux qui régissent le jeu des affinités [Maisonneuve, 1966]. En effet, il illustre la notion de proximité et montre qu’un contact entre des personnes similaires survient à un taux plus élevé que chez les personnes dissemblables [McPherson et al., 2001]. En fait, intuitivement, les gens forment des liens avec les personnes qui leurs sont similaires [Gueorgi and Duncan, 2009]. Donc, ce phénomène explique la composition des groupes en communautés en termes de similitude des membres correspondants (voir *figure* 1.4). Même la confiance et la solidarité seraient plus faciles à s’établir entre des personnes similaires que des personnes dissimilaires. Le résultat est que les réseaux personnels des gens sont homogènes en ce qui concerne de nombreuses caractéristiques socio-démographiques, comportementales et intra-personnelles. Ce phénomène est alors une variable clé dans la formation de groupes cohésifs, qui sont les communautés d’intérêt, dans les Réseaux Sociaux.

[Lazarsfeld and Merton, 1954] ont distingué deux types d’homophilie : l’homophilie de statut, dans laquelle la similarité est basée sur un statut informel, formel ou attribué, et l’homophilie valuée, qui repose sur des valeurs, des attitudes et des croyances. L’homophilie de statut comprend les dimensions socio-démographiques majeures qui stratifient la société - caractéristiques attribuées comme la race, l’origine ethnique, le sexe ou l’âge, et ont acquis des caractéristiques similaires, l’éducation, l’occupation ou les comportements. L’homophilie valuée comprend la grande variété d’états internes présumés pour façonner notre orientation vers le comportement futur :

- éthique ;
- genre ;

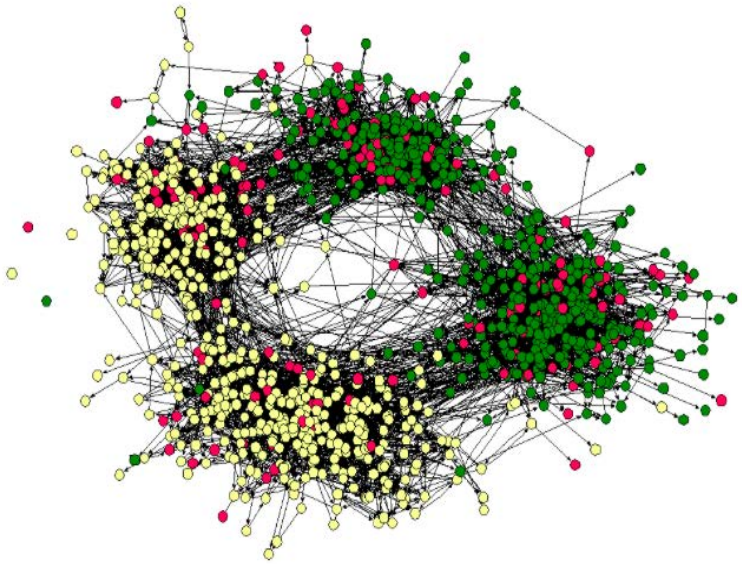


FIGURE 1.4 – Démonstration du phénomène d'Homophilie : dans un réseau d'amitiés au lycée, les noeuds sont connectés si les élèves sont amis et colorés par origine avec des noeuds jaunes et verts formant deux groupes distingués. Des élèves minoritaires (noeuds colorés en rouges) existent dans les deux groupes [Moody, 2002]

- âge ;
- religion ;
- éducation, occupation et classe sociale ;
- positions dans le réseau ;
- comportement ;
- attitudes, idées et aspirations.

1.1.4 Influence sociale

Dans les Réseaux Sociaux, les personnes ressemblent à leurs amis en raison d'une certaine influence sociale. Cette dernière est un changement du comportement d'une personne en adoptant des comportements exposés par leurs voisins. Un problème central pour l'influence sociale est de comprendre l'interaction entre la similarité et les liens sociaux [Leenders, 2002]. Ce phénomène engendre de grands groupes cohérents qui forment des communautés. Ces dernières, réciproquement, jouent un rôle important en influençant

les comportements des utilisateurs. Par conséquent, la compréhension et l'étude de l'influence sociale dans les Réseaux Sociaux au niveau communautaire sont importantes. L'influence intracommunautaire des utilisateurs dépend de leur importance dans la communauté. En effet, les théories sociologiques indiquent que les positions des utilisateurs dans les Réseaux Sociaux ont un impact important sur leur influence sociale. Ainsi, il faut mettre l'accent sur les utilisateurs qui sont les noyaux de la communauté, appelés leaders. Le leadership est une forme commune de l'influence. Les leaders d'une communauté sont considérés comme ayant une plus grande influence que les autres membres. Donc, pour comprendre ce phénomène, il est crucial de comprendre les comportements des leaders.

Notons que les propriétés de l'influence sociale sont décrites comme suit :

- **Propagative** : l'information est caractérisée par sa nature propagative. En effet, elle peut être transmise d'un individu à un autre dans un Réseau Social, créant ainsi des chaînes d'influence.
- **Composable** : le long des chaînes sociales, la propagation de l'influence permet à un individu d'influencer un autre qui peut ne pas être directement lié à celui-ci. Cependant, lorsque plusieurs chaînes influencent un membre, il doit composer l'ensemble des informations d'influence.
- **Mesurable** : une valeur d'influence permet de mesurer le niveau d'influence par un nombre réel continu. Elle peut être également représentée par une incertitude (plus faible, faible, moyenne, forte, plus forte).
- **Subjective** : l'influence est caractérisée par sa nature subjective. En effet, son calcul peut être personnalisé où les opinions et les préférences de l'influenceur ont un impact sur la valeur d'influence calculée.
- **Asymétrique** : l'influence est asymétrique. Si par exemple, une personne x influence une personne y mais cela n'implique pas forcément que y influence x.

1.2 Le Big Data Social

Actuellement, le Big Data représente un enjeu important pour divers domaines de recherche tels que les Réseaux Sociaux, l'Apprentissage Automatique, la Fouille de données et le Web sémantique. La définition de ce concept a évolué d'un modèle $3V$ à un modèle $7V$. En effet, en 2001, Laney l'a défini comme des ressources d'information à haut volume, à haute vélocité et à grande variété qui exigent des formes innovantes et rentables de traitement de l'information pour améliorer la compréhension et la prise de

décision [Laney, 2001]. Cette définition a été mise à jour en 2012 pour spécifier que les données volumineuses sont des ressources d'information volumineuses, à grande vitesse et/ou de grande variété qui requièrent de nouvelles formes de traitement pour améliorer la prise de décision et l'optimisation des processus [Beyer and Laney, 2012]. Selon ces définitions, les trois caractéristiques de base sont : "Volume", "Variété" et "Vitesse". Un modèle $4V$ a été étendu par la suite pour inclure une quatrième caractéristique qui est la "Valeur". La "Véracité" a été incorporée pour donner le modèle $5V$, à quoi s'ajoutent la "Visualisation" et la "Variabilité" donnant finalement le modèle $7V$

En particulier, une véritable Big Data émerge du web social [Khan et al., 2014]. En effet, des quanta de données se diffusent dans les Réseaux Sociaux reflétant la façon dont les gens interagissent entre eux. Ces immenses flots de données sont aussi caractérisés par leur rapidité ce qui esquisse l'essence du Big Data. Forums de discussions, plateformes vidéo, réseaux de partage ou généraliste, autant de plateformes aux contenus riches peuvent être exploitées. Ainsi, le Big Data Social provient de la combinaison des efforts des deux domaines de Réseaux Sociaux et de Big Data. Il est donc question d'analyser des grandes quantités de données sociales provenant de plusieurs sources distribuées. Elles incluent les commentaires des blogs, les tweets, les photos ou encore les messages Facebook. En tant que tel, ces données sont difficiles à traiter en utilisant les technologies existantes [Constantiou and Kallinikos, 2015]. Une alternative aux solutions traditionnelles fondées sur les bases de données et leur analyse est fournie par le Big Data. Elle vise à analyser les données afin d'exploiter leur valeur en tant que connaissances et ne concerne pas seulement leur stockage et accès.

Par conséquent, le Big Data Social peut être défini comme étant les processus et méthodes conçus pour fournir des connaissances pertinentes provenant de sources de données des Réseaux Sociaux qui peuvent être caractérisées par leurs différents formats et contenus, leur grande taille et la génération d'information en ligne [Orgaz et al., 2016]. Cependant, l'extraction des connaissances à partir des sources non structurées contenues dans les Réseaux Sociaux, s'avère une tâche extrêmement difficile qui n'est pas complètement résolue. L'un des principaux défis consiste à identifier les données précieuses parmi les grands ensembles de données hétérogènes ; et les analyser pour en extraire des connaissances utiles. Parmi les applications du Big Data Social, il serait intéressant d'expliquer les comportements des utilisateurs en déterminant et suivant des métriques pour le faire. De plus, il est possible de profiter des informations sur les rôles joués par les individus au sein des communautés, comme les personnes influentes.

Trois grands domaines de recherche sont alors en interaction : les Réseaux Sociaux en tant que sources naturelles de données ; le Big Data comme paradigme de traitement parallèle et massif ; et l'analyse des données sous la forme d'un ensemble d'algorithmes

et de méthodes utilisés pour analyser et extraire des connaissances. L'intersection de ces domaines représente le concept des applications sociales dont l'objectif principal est l'extraction et l'exploitation des connaissances issues des données sociales.

Dans ce qui suit, nous fournissons une description de quelques méthodes de base liées à l'analyse du Big Data Social.

1.2.1 De l'analyse des Réseaux Sociaux à l'analyse prédictive

Les analyses prédictives sont le résultat pratique du Big Data. En effet, les modèles prédictifs sont plus efficaces grâce à la puissance de calcul associée au Big Data. Ces analyses impliquent l'application de la fouille de données (*Data Mining*), de l'apprentissage automatique (*Machine Learning*) et de la modélisation statistique pour produire des modèles prédictifs d'observations futures ainsi que des méthodes appropriées pour déterminer le pouvoir prédictif de ces modèles en pratique [Shmueli and Koppius, 2011].

En particulier, et afin d'anticiper l'évolution des structures communautaires, des analyses prédictives peuvent être exploitées dans les Réseaux Sociaux. En effet, ces derniers s'avèrent des immenses sources de données là où les comportements des individus sont liés aux structures des réseaux dans lesquelles ils s'insèrent.

Une fois les données sont collectées et agrégées, les modèles prédictifs sont construits à partir des analyses statistiques, faisant appel à des algorithmes sophistiqués, et d'apprentissage automatique. La démarche consiste à prédire l'apparition d'un évènement ou une variable par rapport à des variables explicatives (prédicteurs). Ces variables ne sont pas forcémet dépendantes du temps. Les algorithmes utilisés dépendent des distributions des données ; lesquels couvrent les modèles paramétriques (régression linéaire, analyse discriminante...) et les modèles non-paramétriques (arbres de décision, réseaux de neurones, plus proches voisins...). La question suivante est de retenir le meilleur modèle en termes d'efficacité et de précision dans la situation donnée.

La classification supervisée est une méthode prédictive de la fouille de données. Son objectif est de définir des règles permettant de classer des objets dans des classes définies à priori à partir de variables qualitatives ou quantitatives caractérisant ces objets. Notons que la classification fait souvent appel à l'Apprentissage Automatique. Chaque paradigme d'apprentissage se caractérise par un modèle souvent paramétrique, une façon d'interagir avec l'environnement correspondant, une fonction de coût à minimiser (sauf exception) et un algorithme pour adapter le modèle en utilisant les données issues de l'environnement de façon à optimiser la fonction de coût. Un échantillon dit d'apprentissage dont le classement est connu, est utilisé pour l'apprentissage des règles de classement. Afin de comparer les différents modèles de classification supervisée, un échantillon de test est

utilisé. En effet le taux de bon classement sera évalué sur cet échantillon n'ayant pas servi à estimer les règles de classement.

La validation croisée est une méthode basée sur l'échantillonnage pour l'évaluation de la fiabilité d'un modèle. La technique "*k-fold cross-validation*" se base sur la division de l'échantillon de données original en k échantillons parmi lesquels un est utilisé comme échantillon de test et les $(k-1)$ restants constitueront l'ensemble d'apprentissage. Par la suite, chaque ensemble servira tour à tour d'échantillon test. L'estimation de l'erreur de prédiction est enfin basée sur le calcul de la moyenne des k-erreurs quadratiques moyennes. Un cas particulier de cette technique où $k = n$ est appelé "*Leave-one-out cross-validation*".

1.2.2 Diffusion de l'information dans les Réseaux Sociaux

La diffusion de l'information est l'un des rôles les plus importants des Réseaux Sociaux. En conséquence, de nombreuses études ont été proposées pour la modéliser en tenant en compte deux caractéristiques qui sont : *(i)* la structure topologique du réseau et *(ii)* la dynamique temporelle [Nguyen and Jung, 2015].
Deux types de modèles sont distingués [Guille et al., 2013] :

1. **Les modèles descriptifs** : visent à découvrir les cascades de diffusion cachées une fois les séquences d'activation sont collectées. Ces modèles sont utilisés pour chercher un chemin entre les noeuds qui permet de comprendre comment l'information est diffusée ;

2. **Les modèles prédictifs** : sont utiles pour comprendre la propagation de l'information en retraçant le chemin emprunté par celle-ci. En effet, ils visent à prédire le déroulement du processus de diffusion dans un réseau donné à partir des traces de diffusion passées de point de vue spatial et/ou temporel. En particulier, deux catégories sont distinguées : *(i)* les modèles basés sur la structure du réseau et *(ii)* les modèles basés sur l'analyse du contenu. En ce qui concerne la première catégorie, il existe deux modèles phares à savoir Independent Cascades (*IC*) (Goldenberg et al. 2001) et Linear Threshold (*LT*) (Granovetter 1978). Leur principe est de considérer les structures du processus de diffusion et du graphe qui la sous-tend. Chaque noeud du graphe peut être activé ou non en se basant sur l'hypothèse de monotonie. Cette dernière suppose qu'un noeud activé ne peut pas être désactivé. Dans (Kempe et al. 2003), des modèles généralisés de *IC* et *LT* sont proposés.

Particulièrement, l'étude de la diffusion de l'influence à travers un Réseau Social intéresse de plus en plus des travaux dans les sciences sociales. Effectivement, lorsque les

acteurs sociaux voient leurs “amis” effectuer une action, ils peuvent effectuer la même action eux-mêmes. Réellement, une action effectuée par un utilisateur peut avoir l'une des raisons suivantes : ils en ont entendu parler en dehors du réseau mais l'action peut devenir populaire ; ou ils peuvent être influencés par leurs “amis” ayant effectué cette action. Par exemple, l'idée derrière le marketing viral est qu'en ciblant les utilisateurs les plus influents du réseau, une réaction en chaîne de l'influence pourrait être activée de sorte qu'une grande partie du réseau est atteinte avec un faible coût de marketing.

1.2.3 Agrégation des données pour le Big Data Social

L'agrégation des Réseaux Sociaux permet de collecter et représenter l'ensemble des données sociales d'un utilisateur de divers sites en une représentation unifiée. D'abord l'utilisateur en question est identifié de manière unique en repérant ses différents comptes dans les plateformes sociales. Les données utiles sont identifiées parmi celles collectées. Elles alimentent des modèles communs de représentation de données sociales hétérogènes et souvent non structurées. Dans le but de collecter les données sociales des utilisateurs, plusieurs méthodes ont été proposées dont deux techniques sont automatisées : 1) le parcours des pages des profils des utilisateurs en fouillant les pages avec un script automatisé qui scanne et extrait les informations recherchées à partir de codes *HTML* en utilisant les requêtes et les réponses *HTTP* ; et 2) l'utilisation des *API* de développement fournies par les fournisseurs des sites des Réseaux Sociaux. Cette technique nécessite d'abord de créer une application auprès du site du Réseau Social correspondant. Les utilisateurs doivent ensuite donner des permissions appropriées à l'application afin de pouvoir envoyer des requêtes pertinentes à l'*API* associée au réseau pour collecter des données. Avec cette technique, une application ne se limite pas aux informations publiques, mais peut accéder à beaucoup plus de données des utilisateurs.

L'hétérogénéité sémantique est un enjeu important de la fusion des données. En effet, les Réseaux Sociaux présentent intrinsèquement différentes sémantiques qui incluent non seulement les différences linguistiques mais aussi les différences des structures conceptuelles. Pour pallier ces problèmes, les ontologies sont exploitées et plus important encore, les correspondances sémantiques obtenues des méthodes d'appariement des ontologies. Plus concrètement, les technologies sémantiques utilisent les données ouvertes liées (*LOD*) basées sur le modèle de données *RDF*, comme modèle de données unifié pour combiner, agréger et transformer des données à partir de sources hétérogènes.

1.2.4 Analyse des Sentiments

Savoir ce que pensent les gens à propos d'un sujet d'intérêt donné est fondamental. 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'**opinion** étant l'expression d'un individu à propos du sujet d'intérêt. Cet individu qui s'exprime est qualifié comme le **porteur d'opinion** (*opinion holder*) et le sujet d'intérêt comme la **cible de l'opinion** (*opinion target*). La **fouille d'opinions** se réfère donc au domaine du traitement automatique des opinions. Cependant, le jugement que porte un individu sur un sujet d'intérêt est défini comme un **sentiment** (*sentiment*). Ce jugement est caractérisé par une **polarité** (*polarity*). La polarité d'un sentiment est positive si l'opinion est en faveur de l'action bénéfique ; et négative si elle s'y oppose. La polarité "neutre" peut être une position intermédiaire entre celle positive et celle négative. Ainsi, l'**Analyse des Sentiments** (*Sentiment Analysis*) est le champ du traitement automatique des textes qui étudie les sentiments [Pang and Lee, 2008]. Les Réseaux Sociaux permettent d'aller plus loin en identifiant par exemple les leaders d'opinions. Pour mieux appréhender ces concepts, considérons l'exemple suivant [Liu, 2010].

"(1) I bought an iPhone a few days ago. (2) It was such a nice phone. (3) The touch screen was really cool. (4) The voice quality was clear too. (5) However, my mother was mad with me as I did not tell her before I bought it. (6) She also thought the phone was too expensive, and wanted me to return it to the shop ...".

Dans cet exemple, il faut noter qu'il y a plusieurs opinions. Les cibles de l'opinion sont : "iPhone", "écran tactile", "qualité vocale", "moi" (l'auteur de la réplique) et "prix" ; respectivement dans les phrases (2), (3), (4), (5) et (6). Les trois premières phrases expriment des opinions positives et les deux dernières expriment plutôt des sentiments négatifs.

La classification des sentiments permet de donner une impression sur le ton du texte. Ainsi, le sentiment d'un utilisateur peut être déduit à partir de l'ensemble des textes qu'il a publié.

1.3 Les applications sociales

Afin de tirer parti du succès des plateformes des Réseaux Sociaux, beaucoup de gens essaient de mettre en oeuvre des applications sociales. Elles sont des applications reposant sur des interactions avec les Réseaux Sociaux et exploitent leurs infrastructures comme moyen de dialogue avec les utilisateurs finaux. En effet, ces plateformes sont conçues

pour prendre en charge un nombre élevé d'utilisateurs et de contenus. Par conséquent, les applications sociales peuvent utiliser l'infrastructure des Réseaux Sociaux pour atteindre plus d'utilisateurs. Donc, il s'agit bien d'une interface front-end robuste. En outre, divers services et applications peuvent réutiliser les données sociales et les informations disponibles dans les réseaux. L'échange de contenu entre les services est facilité grâce à une telle portabilité. Cette dernière nécessite des mécanismes de représentation qui sont fournis grâce aux formalismes de web sémantique, permettant aux personnes et aux ordinateurs de travailler en coopération. En particulier, *FOAF* fournit une solution de portabilité en termes de profils d'utilisateurs et de connexions réseau.

Cependant, développer des applications sociales a des aspects particuliers qui sont différents des applications web génériques, par exemple :

- Généralement dans les Réseaux Sociaux seule une partie des données est publique. Les applications sociales ont donc besoin d'obtenir des autorisations spécifiques des utilisateurs afin d'accéder au reste de données privées demandées.
- L'accès aux ressources des Réseaux Sociaux nécessite l'utilisation des *API* de développement correspondantes. L'inconvénient que les développeurs ont besoin de gérer des concepts de programmation complexes en manipulant ces interfaces.
- Les applications sociales doivent s'authentifier auprès des Réseaux Sociaux avec lesquels elles vont interagir. Chaque Réseau Social possède son propre processus d'authentification pour les applications.

1.3.1 Qui a besoin d'implémenter des applications sociales ?

Les besoins de construction d'applications sociales peuvent être classés dans les catégories suivantes :

- **Applications de Crowdsourcing** : ces applications se basent sur la répartition des tâches entre un grand nombre de participants. Elles peuvent être construites à travers les Réseaux Sociaux pour atteindre plus d'utilisateurs.
- **Diffusion ciblée** : ces applications sont utilisées par des utilisateurs voulant envoyer une requête à un groupe d'amis.
- **Interfaces utilisateur pour les systèmes d'entreprises** : certaines entreprises utilisent des Réseaux Sociaux comme moyen d'accès à leurs systèmes d'informations. L'avantage est que les utilisateurs n'ont pas besoin d'installer un nouveau logiciel pour interagir avec ces systèmes. Ils utilisent leurs comptes des sites des Réseaux Sociaux pour l'interaction.

- **Création de contenus guidés** : ces applications permettent de guider les utilisateurs à travers les processus de création de contenus proposés.

- **Applications pour des situations inattendues** : ces applications permettent de diffuser des informations lors des catastrophes naturelles ou la réaffectation des vols en cas d'annulations ou grèves, etc. Les Réseaux Sociaux sont populaires et supportent de fortes charges donc les utilisateurs n'auraient pas besoins d'installer de nouvelles applications.

- **Applications Ad-hoc collaboratives** : des groupes d'utilisateurs finaux ont besoin d'utiliser certains Réseaux Sociaux pour collecter des données structurées sous forme d'informations qui peuvent être par la suite traitées automatiquement. Ces applications comprennent celles d'organisation événementielle ou de votes. Les premières permettent aux utilisateurs de définir des messages d'invitation et de les programmer pour une diffusion périodique via les Réseaux Sociaux pour des événements sportifs, de jeux éducatifs ou d'autres. Les secondes présentent aux utilisateurs des listes d'objets et leurs demandent de sélectionner un ou plusieurs d'entre eux.

Dans la plupart du temps, la construction des applications utilisant les Réseaux Sociaux se fait manuellement sans support automatisé. Dans ce contexte, nous observons un besoin croissant d'automatiser cette construction.

Dans le cadre de notre thèse, nous visons à générer des applications interactives à contenus personnalisés basées sur les Réseaux Sociaux en suivant une approche de type *MDA*[12] (*Model Driven Architecture*). Ce type d'approche permet la spécification des applications à partir de modèles conceptuels de manière abstraite. Une suite de transformations automatiques ou semi-automatiques est par la suite appliquée à travers des spécifications afin de générer l'interface finale. Ainsi l'automatisation de certaines tâches du processus de développement permet d'assurer une qualité plus fiable des interfaces et de réduire l'intervention des concepteurs.
MDA, standard lancé par *OMG*[13], se base sur le paradigme de l'*Ingénierie Dirigée par les Modèles*. Cette discipline fait référence à l'utilisation des modèles, qui sont des abstractions des éléments du monde réel, comme élément principal dans le cycle de vie du logiciel. Ces modèles, qui sont les piliers de cette technologie, sont normalement moins liés à la technologie et plus proches du domaine. Ils permettent de modéliser des systèmes complexes en un temps de développement plus réduit. Selon *OMG* 2003, un modèle est

12. http ://www.omg.org/mda/
13. http ://www.omg.org

une description ou une spécification d'un système et de son environnement dans un but bien déterminé. Un modèle est souvent présenté comme une combinaison de dessins et de textes. Le texte peut être dans un langage de modélisation ou dans une langue naturelle.
La technique utilisée pour la définition des relations entre les éléments d'un modèle et sa syntaxe abstraite est la méta-modélisation. Cette dernière est une abstraction du modèle lui-même en définissant ses propriétés.
Selon *OMG* 2006 un méta-modèle est défini comme un modèle définissant le langage permettant d'exprimer un modèle.
La solution proposée dans notre thèse est de générer des systèmes interactifs personnalisés qui s'adaptent aux besoins des utilisateurs en termes de leurs intérêts et leurs contextes. Dans ce mémoire, nous adoptons la définition de la personnalisation proposée par [Brossard, 2008], [Bacha et al., 2011].

Définition 10 (Personnalisation)**.** *La personnalisation est la capacité de fournir à un utilisateur, à chaque instant, des contenus et des services adaptés à ses besoins et à ses attentes en utilisant des interactions homme-machine appropriées.*

Ce processus peut être appliqué à plusieurs niveaux qui sont les suivants [Abbas, 2008, Kobsa et al., 2001] :

- **Le contenu** : il s'agit d'adapter le contenu aux besoins et intérêts de l'utilisateur ainsi que son contexte ;
- **La navigation** : ce type de personnalisation permet d'orienter l'utilisateur, en fonction de sa connaissance par le système, afin d'éviter les chemins lui menant à des informations non pertinentes ;
- **La présentation** : elle consiste à adapter le format des éléments d'interaction de l'interface de l'application en tenant compte des contextes et besoins des utilisateurs ;
- **Les fonctionnalités** : il s'agit d'adapter le système en fonction de sa connaissance des besoins fonctionnels de l'utilisateur afin de faciliter l'accomplissement des tâches ;
- **La structure** : la personnalisation conduit à la structure d'un site web c'est-à-dire aux liens entre les pages qui le composent.

Dans le cadre de nos travaux, nous nous intéressons à la personnalisation du contenu dans la génération des applications interactives en tenant compte des données de contexte de l'utilisateur. Ces dernières permettent de définir les mécanismes de personnalisation.

Cette dernière est sensible aux différents contextes d'un utilisateur. Il convient donc de spécifier une définition pouvant être utilisée de manière normative dans le domaine de l'informatique contextuelle.

Définition 11 (Contexte)**.** *Le contexte est une information qui peut être utilisée pour caractériser la situation d'une entité (une personne, un lieu ou un objet) considérée comme pertinente pour l'interaction entre un utilisateur et une application, y compris l'utilisateur et les applications eux-mêmes [Abowd et al., 2000].*

Certains facteurs peuvent influencer le processus de personnalisation du contenu, qui sont les suivants :

- **Les facteurs d'information ou de contenu** : sont les facteurs qui représentent les caractéristiques des contenus comme le type et méta-données d'un document. Ces propriétés permettent de spécifier la manière de personnalisation de l'information ;
- **Les facteurs liés aux utilisateurs** : ils représentent l'ensemble des centres des intérêts, des préférences et caractéristiques des utilisateurs permettant de décrire la manière avec laquelle la personnalisation de l'information doit être effectuée ;
- **Les facteurs liés à la méthode de personnalisation** : différentes méthodes de personnalisation peuvent être utilisées ce qui explique la différence dans les résultats obtenus ;
- **Les facteurs de contexte** : ils représentent des facteurs des environnements techniques et physiques de l'utilisateur et du système.

Plusieurs méthodes peuvent être utilisées pour la mise en place de la personnalisation du contenu [Doucet et al., 2004],[Abbas, 2008],[Ioannidis and Koutrika, 2005] :

- **La recommandation** : cette méthode permet de recommander à un utilisateur des informations en se référant aux différentes expériences des autres utilisateurs avec le système ;
- **La recherche personnalisée de l'information** : elle consiste à satisfaire un besoin d'un utilisateur exprimé par une requête sur un ensemble de documents appelés corpus ;
- **Le remplissage automatique des formulaires** : il s'agit de remplir automatiquement les formulaires qui sont traditionnellement remplis manuellement par les utilisateurs. Ce remplissage tient compte des profils de ces derniers ainsi que leurs contextes ;

- **Le filtrage de l'information** : il s'agit d'un processus permettant d'extraire des informations pertinentes et d'éliminer celles non pertinentes parmi l'ensemble des informations sollicitées par l'utilisateur [Hanani et al., 2001]. Il s'avère une solution efficace au problème de surcharge d'information.

1.4 Conclusion

Dans ce premier chapitre, nous avons présenté les définitions et la terminologie utilisées dans cette thèse. Nous avons aussi introduit les concepts de Réseaux Sociaux, communautés d'intérêts, l'Analyse des Sentiments comme application du Big Data social et la génération des applications interactives. En outre, nous avons donné les définitions des communautés d'intérêts et indiqué les apports et l'importance de ces dernières dans plusieurs applications. Dans notre travail, nous nous basons sur ces communautés afin de générer des applications sociales interactives et personnalisées.

A la différence de la représentation basée sur les graphes, la représentation sémantique se concentre sur les connaissances encodées dans les données sociales analysées. Dans notre travail, nous optons donc pour l'ontologie *FOAF* pour représenter les Réseaux Sociaux. En effet, les ontologies offrent plusieurs avantages.

Pour achever notre étude bibliographique, nous allons présenter, dans le chapitre qui suit, un état de l'art sur les approches de détection des communautés, d'Analyse des Sentiments et de génération des applications sociales.

CHAPITRE 2

Analyse des Réseaux Sociaux et Applications sociales : Etat de l'art

Sommaire

Dans ce chapitre, nous présentons l'état de l'art se rapportant à la détection des communautés dans les Réseaux Sociaux, l'Analyse des Sentiments dans ces derniers et la génération des applications sociales interactives. D'abord, les problèmes de détection de communautés ont été étudiés dans plusieurs travaux de recherche et apparaissent sous différentes formulations qui ne posent pas nécessairement les mêmes méthodes de résolution. Dans la première section, nous présentons une revue de littérature sur la problématique de détection des communautés d'intérêt dans les Réseaux Sociaux. Cette étude bibliographique est effectuée à travers une catégorisation des méthodes existantes distinguant trois classes ; à savoir celles basées sur la structure topologique du Réseau Social, les approches exploitant les attributs des nœuds en plus de la topologie et les approches considérant les phénomènes sociaux comme l'influence et la confiance avec la structure du réseau. En décrivant les principaux travaux existants, nous allons examiner ces catégories en précisant à chaque fois les avantages et les limites rencontrées. La section 2 sera consacrée à l'Analyse des Sentiments dans les Réseaux Sociaux et ses utilisations. Si ces deux problèmes (exposés dans les sections 1 et 2) ont connu des évolutions relativement indépendantes, nous expliquerons aussi à quels points ils peuvent se croiser dans la réso-

lution de notre problématique. La section 3 est destinée à étaler les travaux existants dans la génération des applications sociales et son automatisation.

L'objectif de cette étude est de bien définir le cadre de recherche afin de mettre en valeur la contribution de notre travail par rapport à l'existant.

2.1 Détection des communautés d'intérêt

La recherche des partitions en communautés est l'un des principaux problèmes liés à l'analyse des Réseaux Sociaux. Les méthodes de sa résolution ont fait l'objet de nombreux travaux, depuis l'article fondateur de Girvan et Newman [Girvan and Newman, 2002]. Cependant, il reste encore plusieurs questions en suspens. En effet, les Réseaux Sociaux, qui sont des réseaux complexes, présentent en général des propriétés topologiques non triviales, contrairement aux graphes aléatoires. Ces propriétés caractérisent la connectivité du réseau et impactent la dynamique des processus qui y sont appliqués. De plus, les Réseaux Sociaux sont parmi les plus gros producteurs de données. Les utilisateurs y postent, consultent et échangent des données de toutes sortes. Donc, l'identification des communautés dans les Réseaux Sociaux s'avère un problème complexe qui a été abordé sous différentes perspectives. En général, les techniques de détection de communautés d'intérêt se réfèrent à une classification des noeuds du réseau plus densément connectés que d'autres, pour construire des classes connexes d'utilisateurs ayant les mêmes caractéristiques au regard d'une mesure de similarité se référant à des intérêts communs.

2.1.1 Classifications des méthodes de détection de communautés

Plusieurs classifications des méthodes proposées pour la détection des communautés d'intérêt dans les Réseaux Sociaux ont été publiées. La plupart de ces classifications sont basées sur les types des algorithmes de détection des communautés et leurs principes méthodologiques. Notons que, aujourd'hui, détection de communautés, partitionnement de graphe et clustering en anglais sont souvent utilisés indifféremment.

Fortunato [Fortunato, 2010] a mené une étude exhaustive et classé les méthodes en huit catégories, qui sont les suivantes :

1. **Méthodes traditionnelles** : Consistent au partitionnement optimal en k partitions ou "clusters", k étant donné, des graphes représentant les Réseaux Sociaux. La méthode la plus connue est l'algorithme de Kernighan-Lin [Kernighan and Lin, 1970]. La solution proposée est de chercher des partitions souvent de même taille. Cette contrainte étant trop stricte et difficile à atteindre dans des cas réels, elle a été relâchée de manière à chercher des communautés mais sans avoir à préciser la taille

exacte [Hastie et al., 2001]. Le partitionnement hiérarchique se base alors sur une fonction de similarité de telle sorte que les "clusters" contiennent des noeuds avec des similarités fortes ;

2. **Algorithmes divisifs** : Ils sont basés sur la recherche d'une propriété des liens inter-communautaires pour pouvoir les identifier afin de les éliminer [Girvan and Newman, 2002]. Ces éliminations déconnectent le graphe pour former des composantes connexes qui représentent les communautés. L'algorithme le plus populaire est celui proposé par Girvan et Newman [Newman and Girvan, 2004] ;

3. **Méthodes basées sur la modularité** : La modularité, proposée par Girvan et Newman, est utilisée par plusieurs algorithmes comme fonction de qualité en l'optimisant ou la modifiant [Newman, 2004] telles que les techniques gloutonnes [Clauset et al., 2004] et d'autres fonctions d'optimisation [Agarwal and Kempe, 2008, Berry et al., 2009, Chen et al., 2008, Xu et al., 2007]. Cependant, [Brandes et al., 2006] ont démontré que trouver la partition optimale en modularité est un problème NP-complet. Donc d'autres alternatives, comme l'algorithme de Louvain [Blondel et al., 2008], sont basées sur des techniques gloutonnes pour apporter des solutions acceptables en temps de calcul ;

4. **Algorithmes de partionnement spectral** : Ces algorithmes sont basés sur la notion du spectre définissant la proximité entre les noeuds. Les vecteurs propres, agissant comme des propagateurs de temps pour le processus de marche aléatoire (Random Walk) dans le graphe du Réseau Social, associés aux valeurs propres les plus faibles décrivent les groupes à similarité interne forte [Alves, 2007, Simonsen, 2005, Yang and Liu, 2008]. Généralement la matrice Laplacienne est utilisée comme matrice de similarité ;

5. **Algorithmes dynamiques** : Simulent des processus dynamiques où les particules s'influencent entre elles. Ainsi, les particules proches les unes des autres ont tendance à partager le même état. Parmi les processus appliqués aux graphes des Réseaux Sociaux nous citons Spin-Spin [Middleton and Fisher, 2002, Reichardt and Bornholdt, 2004], Random-Walk [Hu et al., 2008, Zhou, 2003] et la synchronisation où le système unifie progressivement tous ses éléments au même état [Boccaletti et al., 2007, Li et al., 2008] ;

6. **Méthodes basées sur l'inférence statistique** : Comme l'inférence Bayésienne, y compris les modèles génératifs [Hastings, 2006, Newman and Leicht, 2007] et la modélisation de blocs [Reichardt and White, 2007, Rosvall and Bergstrom, 2008].

Ces méthodes supposent que le graphe a été généré suivant un modèle admettant l'appartenance des noeuds aux communautés comme des paramètres. Le but est alors d'inférer les paramètres qui auraient généré les observations trouvées avec la probabilité la plus élevée ;

7. **Méthodes pour extraire des communautés recouvrantes** : Il est évident qu'un noeud peut appartenir à plusieurs groupes ou communautés ; c'est la caractéristique de recouvrement de communautés. La première méthode prenant efficacement le recouvrement a été proposée en 2005 par [Palla et al., 2005]. Par la suite, d'autres approches ont été proposées [Nepusz et al., 2007] ;

8. **Méthodes multi-résolution** : L'application du paradigme multi-résolution à la détection de communautés cherche à intégrer un facteur d'échelle permettant de déterminer directement l'échelle de détection et indirectement la taille caractéristique des communautés [Tapio et al., 2008].

De même, [Yang et al., 2010] ont proposé de classer les méthodes de détection communautaire existantes en trois familles :

1. **Algorithmes d'optimisation** : Où la classification des noeuds du graphe du réseau est considérée sous l'angle d'un problème d'optimisation d'une fonction objective prédéfinie ; y compris les méthodes spectrales [Kernighan and Lin, 1970, Newman, 2004] et la recherche locale ;

2. **Algorithmes heuristiques** : Ces algorithmes, contrairement aux algorithmes d'optimisation, n'impliquent pas l'optimisation de fonctions objectives. Ils sont plutôt basés sur des règles heuristiques [Palla et al., 2005], comme la règle utilisée dans l'algorithme de Girvan et Newman stipulant que le nombre de liens inter-communautés devrait être supérieur à celui intra-communautaire [Girvan and Newman, 2002] ;

3. **Algorithmes basés sur la similarité et méthodes hybrides** : Le reste des méthodes comprend des approches ascendantes qui essaient de joindre de façon répétitive des paires de groupes de noeuds en fonction de leur similarité. En effet, le problème de détection de communautés est transformé en un problème de clustering dans un espace vectoriel où chaque noeud est représenté par des points avec des coordonnées k-dimensionnelles. Ces points spatiaux sont ensuite regroupés en utilisant des algorithmes de clustering spatial tel que k-means [Donetti and Muñoz, 2004].

[Papadopoulos et al., 2012], dans leur étude bibliographique, ont classifié les méthodes de détection de communautés en cinq catégories :

1. **Détection de sous-graphes cohésifs** : La philosophie de ces approches est de considérer une communauté comme un sous-graphe du réseau satisfaisant une certaine spécification des propriétés structurelles. Une fois que ces dernières sont spécifiées, les méthodes impliquent l'énumération des structures de sous-graphes dans le réseau. Les cliques et n-cliques sont des exemples des structures cohésives [Palla et al., 2005] ;

2. **Clustering des sommets** : Inspirées de la recherche traditionnelle sur la classification des données, ces techniques reposent sur l'intégration des sommets du graphe dans un espace vectoriel. Cette intégration est exploitée afin de pouvoir utiliser des méthodes de clustering traditionnelles [Pons and Latapy, 2005] ;

3. **Optimisation de la qualité des communautés** : Plusieurs travaux ont été basés sur l'optimisation de la qualité des communautés détectées dans le graphe du réseau. En plus de la technique d'optimisation gloutonne de la modularité de Newman [Newman, 2004], d'autres méthodes ont visé à optimiser des mesures locales de la qualité des communautés telle que la modularité locale Clauset [2005] ;

4. **Méthodes divisives** : Ces méthodes permettent d'identifier les éléments inter- communautés quelques soient des arêtes ou bien des sommets. [Girvan and Newman, 2002] a proposé un algorithme pour supprimer progressivement les arêtes du réseau en fonction d'une mesure d'écart entre les liens jusqu'à obtenir des composantes déconnectées du graphe qui sont les communautés ;

5. **Méthodes à base de modèle** : Ces méthodes se basent sur des modèles soit en considérant un processus dynamique qui règne dans le réseau révélant ainsi ses communautés, soit un modèle sous-jacent de nature statistique expliquant la composition du réseau en communautés [Reichardt and Bornholdt, 2006]. En outre, le problème de détection de communautés a été considéré comme un problème d'inférence statistique en supposant le modèle probabiliste sous-jacent ;

[Ding, 2011] considère qu'il existe deux types de connexions dans les Réseaux Sociaux : les connexions sociales qui sont souvent des relations réelles dans les réseaux comme les relations d'amitié, de communication ou de collaboration et les connexions à base de similarité qui sont des connections dérivées et physiquement n'existent pas. En se basant sur ces types, il a distingué les méthodes basées sur la topologie structurelle [Clauset et al., 2004, Girvan and Newman, 2002] et les méthodes basées sur les centres ou sujets d'intérêt [Blei et al., 2008, Tang et al., 2008]. La notion de sujet d'intérêt peut être différenciée en tant que basée sur un évènement défini par un ensemble d'histoires déclenchées par des

évènements du monde réel ; ou sur un sujet qui découle de la notion plus large de sujets comme ceux concernant un document. Ding a prouvé que, dans la première catégorie, les communautés détectées consistent en des noeuds densément connectés mais avec des centres d'intérêt qui peuvent être différents. Et dans la seconde catégorie, il peut résulter des communautés avec des centres d'intérêt cohérents, mais des noeuds assez isolés.

[Zhou et al., 2009] ont intégré le clustering dans les graphes à vecteurs d'attributs où les entités du réseau sont décrites par des vecteurs numériques. Le partitionnement des graphes est en fonction de la similarité des attributs de sorte que les noeuds avec les mêmes valeurs d'attributs sont regroupés en une seule partition. Il a classé les méthodes correspondantes en deux classes : celles basées sur la distance et celles sur les modèles. Dans la première classe, une mesure de distance est définie pour comparer les informations structurelles et les valeurs des attributs de deux noeuds du graphe. Deux fonctions de mesures permettent de mesurer la distance structurelle et la distance entre les attributs en attribuant des poids aux deux types d'information. Cependant, il a été prouvé que l'apprentissage de ces poids est en général couteux en temps [Xu et al., 2012]. Cependant, dans les modèles probabilistes, le problème de clustering est un problème d'inférence probabiliste. Cette approche évite la définition de deux mesures de similarité et fusionne les informations structurelles et des attributs. En outre, le clustering dans les graphes attribués distingue entre les attributs qui sont assignés aux noeuds et ceux qui sont assignés aux arcs ou arêtes [Bothorel et al., 2015].

[Bothorel et al., 2015] ont mené une étude sur les méthodes de clustering dans les graphes attribués et les ont classé en trois familles :

1. **Les méthodes exploitant les attributs puis les relations** : ces méthodes se basent sur l'enrichissement du graphe du Réseau Social soit en ajoutant des sommets et des arêtes basés sur les attributs [Zhou et al., 2009] soit en valuant directement les arêtes à l'aide des attributs [Yang et al., 2013] ;

2. **Les méthodes exploitant les relations puis les attributs** : dans ce contexte, un regroupement des communautés est effectué en se basant sur les valeurs des attributs [Li et al., 2008] ;

3. **Les méthodes exploitant les attributs et les relations conjointement** : parmi les méthodes proposées dans cette perspective, des extensions de la méthode de Louvain sont identifiées [Combe et al., 2012].

Cependant, la richesse de l'information disponible n'est pas limitée aux liens existant entre les entités. Dans un Réseau Social, un individu interagit avec d'autres, et ces interactions obéissent aux différents phénomènes sociaux réels comme l'influence et la

confiance. Le large éventail des données sur ces phénomènes suscitent un intérêt certain dans l'analyse des communautés. [Barbieri et al., 2013] ont proposé un modèle pour la détection de communautés basé sur la propagation des informations dans le Réseau Social ainsi que les relations entre ses acteurs.

2.1.2 Comparaison de quelques méthodes

Dans ce qui suit, nous présentons un ensemble de critères pour comparer quelques approches de détection de communautés dans les Réseaux Sociaux. Nous avons déterminé trois groupes de critères : établissement du modèle, processus de mise en oeuvre et performance.

En ce qui concerne les critères sur la mise en oeuvre du modèle, nous tenons compte des cinq composants suivants :

- **Caractéristiques structurelles (A)** : Elles sont impliquées par diverses interactions entre les utilisateurs. Elles sont définies sur la base de divers liens explicites et implicites. Les liens explicites sont définis comme les relations qui peuvent être captées par l'interaction directe des utilisateurs à travers les fonctions proposées dans les sites des Réseaux Sociaux. Outre ces liens, il existe des relations implicites qui correspondent à des relations tirées des intérêts ou des activités des utilisateurs ne pouvant être tracées par les enregistrements des interactions sur les sites de ces réseaux. L'extraction de telles relations est relativement plus difficile. Diverses mesures structurelles, basées sur les liens explicites et implicites, telles que la connectivité et le coefficient de clustering sont utilisées pour la détection des communautés.

- **Activités sociales (B)** : Les utilisateurs effectuent différentes activités dans Réseaux Sociaux comme commenter une publication ou s'abonner à un membre ou partager un contenu,etc. En particulier, la compréhension de ces comportements à partir de l'agrégation des flux des activités permet de déterminer les relations implicites en termes d'intérêts communs ainsi qu'expliquer la formation et l'évolution de la structure communautaire.

- **Attributs (C)** : Les noeuds du réseau peuvent être étiquetés avec des attributs contenant des informations sur les propriétés des sommets. En effet, les utilisateurs créent des profils mentionnant des informations comme le genre, la localisation, l'âge, les intérêts, etc. Ces attributs peuvent être exploités pour mesurer la similarité entre les utilisateurs en étudiant le phénomène de l'homophilie. Cependant, l'extraction de certains attributs pourrait être difficile.

- **Contenu (D)** : Dans les Réseaux Sociaux, les utilisateurs sont devenus des producteurs d'information en postulant des publications, textuelles ou multimédias, et en échangeant des informations. Les réseaux enrichis sont ainsi des sources d'information contextuelle qui reflètent les préférences et intérêts des utilisateurs.

- **Influence sociale (E)** : Les utilisateurs s'influencent mutuellement à travers les interactions et la communication. Ainsi une propagation de l'information axée sur l'influence est menée dans les Réseaux Sociaux. Le processus de diffusion sous-jacent encode la structure communautaire.

Les critères sur le processus de mise en oeuvre des approches sont :

- **Orientation du graphe** (orienté **(F)** ou non orienté **(G)**) : Les liens dans les Réseaux Sociaux sont souvent orientés. Analyser de tels graphes est souvent plus difficile que les graphes non orientés. L'orientation révèle le flux d'information au sein du réseau.

- **Communautés détectées** (couvrantes **(H)** ou non **(I)**) : Les noeuds du réseau peuvent appartenir à plusieurs communautés à la fois car les utilisateurs correspondants peuvent avoir différents centres d'intérêt. Par conséquent, les communautés peuvent se chevaucher. La plupart des méthodes traditionnelles de détection de communautés sont incapables de retrouver des communautés couvrantes.

- **Fonction de qualité (J)** : Pour formaliser la notion de communauté de réseau, les méthodes de détection communautaire peuvent viser à optimiser une fonction objective qui mesure la qualité de communauté.

- **Démarrage à froid (Cold Start) (K)** : Parfois, les utilisateurs ne précisent pas tous leurs centres d'intérêt dans leurs profils. Ainsi, peu d'informations sont disponibles à leur sujet. Face au manque d'information, des techniques permettent de conclure avec précision les intérêts des utilisateurs.

Les critères sur les performances liées aux méthodes étudiées sont :

- **Passage à l'échelle (L)** : Les Réseaux Sociaux sont des graphes larges avec des milliards d'arcs et de noeuds. Dans ce contexte, les algorithmes de détection de communautés sont mis à l'échelle pour faire face à la taille du réseau.

- **Qualité des communautés détectées (M)** : Les algorithmes de détection de communautés sont censés identifier les "bonnes" partitions satisfaisant les propriétés de

base de communautés. Plusieurs mesures de qualité, telles que la modularité, la mesure de Mancoridis *MQ* et la qualité de compression, ont été proposées pour évaluer la qualité des communautés détectées.

- **Complexité (N)** : Etant donné que les Réseaux Sociaux sont d'énormes sources de données, les méthodes de détection des communautés sont très exigeantes en terme de calcul. La complexité est l'estimation des ressources requises pour effectuer la tâche de leur détection.

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**
Hierarchical agglomeration [Clauset et al., 2004]	+	+	-	+	-	+	-	-	+	-	+	+	+	+
Link-Content model [Natarajan et al., 2013]	+	-	+	-	-	-	+	+	-	+	-	+	+	+
SA-Cluster [Zhou et al., 2009]	+	-	+	-	+	-	+	+	-	+	-	+	+	+
CESNA [Yang et al., 2013]	+	-	+	-	+	-	+	-	+	-	+	+	+	+
SemTagP [Erétéo et al., 2011]	+	+	+	+	-	+	-	+	-	+	+	-	+	+
CCN model [Barbieri et al., 2013]	+	+	-	-	+	+	-	-	+	-	+	+	-	-

TABLE 2.1 – Comparaison de quelques méthodes de détection de communautés dans les Réseaux Sociaux

Dans notre travail, nous comparons certaines approches existantes selon les trois critères définis. La comparaison est résumée dans le *tableau* 2.1 où + (respectivement −)

signifie que l'approche a pris en compte (n'a pas pris en compte) les critères correspondants aux colonnes.

Méthode	**Principe**	**Avantages**	**Désavantages**
[Clauset et al., 2004]	L'algorithme d'agglomération hiérarchique est basé sur l'optimisation de la modularité. Il utilise une optimisation gloutonne. En commençant par chaque vertex en tant que membre unique d'une communauté, deux communautés se rejoignent dont la fusion produit la plus grande augmentation de la modularité.	- S'exécuter en temps quasi-linéaire $O(nlog^2n)$ où n est le nombre de noeuds (N^+) ; - Révèler les modèles à grande échelle (L^+).	Ne considère pas la sémantique (D^-), les attributs (C^-) et les phénomènes sociaux (E^-).
[Natarajan et al., 2013]	Cette approche considère que la structure et le contenu du réseau sont corrélés. Elle propose un modèle probabiliste qui exploite le graphe social et le contenu publié. La structure topologique et le contenu sont donc combinés pour la détection communautaire. Une communauté est ainsi une distribution par rapport aux utilisateurs agrégés par leurs centres d'intérêt.	- Extraire des communautés cohérentes en exploitant la structure topologique (D^+) et le contenu (A^+) ; - Inférer les communautés implicites (cachées).	- En utilisant le modèle Latent Dirichlet Allocation (LDA) pour trouver des centres d'intérêts latents, l'approche suppose uniquement le contenu textuel ; - Ce modèle génératif exploite les liens et les textes mais ignore les attributs contenus dans les profils des utilisateurs (C^-).

[Zhou et al., 2009]	Cette approche propose un modèle qui définit une mesure de distance unifiée basée sur le modèle de Neighborhood Random Walk. Les auteurs ont intégré les similitudes structurelles et des attributs grâce à l'augmentation du graphe du réseau en insérant des sommets des attributs.	- Proposer une mesure de distance unifiée pour les similitudes structurelles et des attributs ; - L'algorithme converge rapidement lors du traitement de larges graphes (L^+).	- Le nombre de communautés détectées doit être spécifié à l'avance ; - Les communautés détectées sont disjointes (H^-).
[Yang et al., 2013]	Ce modèle détecte des communautés chevauchantes en fonction de la structure du réseau et des attributs des noeuds. Il suppose que les noeuds appartenant à la même communauté sont susceptibles d'être connectés les uns aux autres. Donc, plus des noeuds partagent des communautés en commun plus c'est probable qu'ils soient connectés.	Extraire des communautés chevauchantes ou non ainsi que des communautés hiérarchiques (H^+).	Supposer des attributs ayant des valeurs binaires.
[Erétéo et al., 2011]	Cette méthode prend comme entrée un graphe typé et orienté formé par la description *RDF* des Réseaux Sociaux et des folkonomies. Le graphe formé par *RDF* est une représentation de modèle de métadonnées. En utilisant les ontologies pour représenter le Réseau Social, cette méthode est basée sur la propagation des étiquettes (tags).	Exploiter les données sémantiques pour détecter et étiqueter les communautés (D^+).	Ne pas intégrer assez de sémantique pour propager plusieurs étiquettes à travers différents types de relations (L^-).

[Barbieri et al., 2013]	Dans cette approche les communautés expliquent la propagation de l'information et les utilisateurs influents. L'algorithme considère comme entrée un graphe orienté et un historique des cascades d'information. La sortie est des communautés chevauchantes qui expliquent les cascades d'information. Un modèle génératif stochastique pour le graphe social et l'ensemble des cascades est proposé afin de déterminer la probabilité d'observer une action dans une communauté, le niveau d'intérêt actif ou passif de chaque utilisateur dans chaque communauté et de déduire la participation d'un utilisateur dans une communauté. Ce modèle permet de distinguer l'influence sociale de l'homophilie.	Détecter des communautés chevauchantes et, pour chaque noeud, spécifier le niveau d'intérêt actif ou passif (H^+).	Un ensemble de cascades d'informations, c'est-à-dire des informations propagées basées sur l'influence, est nécessaire pour exécuter l'algorithme ; cependant, de telles informations sont parfois manquantes et non précises.

TABLE 2.2 – Avantages et inconvénients des méthodes de détection de communautés comparées

En effet, il est intéressant d'étudier la façon dont les méthodes sont établies pour déterminer les avantages et les inconvénients de chacune d'elles ; résumés dans le *tableau* 2.2. Soit $Critère^+$ (respectivement $Critère^-$) signifiant que la valeur $+$ (respectivement $-$) a été affectée au critère *{A-N}*.

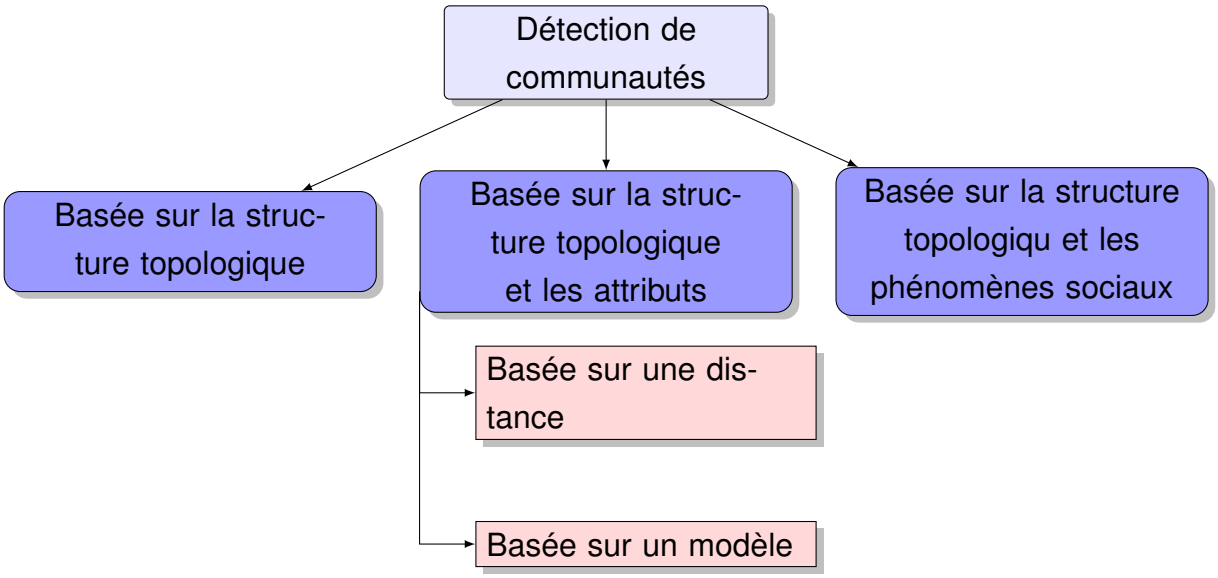


FIGURE 2.1 – Catégories des méthodes de détection de communautés

2.1.3 Nouvelle classification

La présente étude bibliographique nous a mené à une catégorisation des méthodes de détection de communautés distinguant trois familles de méthodes (voir figure 2.1) ; à savoir, des méthodes basées sur des caractéristiques structurelles, des méthodes tenant en compte des attributs en plus de la topologie et d'autres utilisant la structure du réseau avec les phénomènes sociaux comme l'influence.

2.1.3.1 Méthodes basées sur la structure topologique

Les méthodes basées sur la topologie utilisent l'analyse structurelle des réseaux pour trouver des communautés. L'un des algorithmes les plus populaires est celui proposé par Clauset et al. [Clauset et al., 2004]. Dans ce contexte, les communautés sont détectées sur la base d'une approche de clustering dans les graphes, de sorte que la densité des liens est plus élevée au sein des partitions qu'entre elles. Plusieurs mesures de positions dans les réseaux sont proposées et sont liées aux caractéristiques structurelles afin de définir des groupes cohésifs qui forment les communautés.
La détection de communautés purement basée sur la topologie s'avère problématique vu qu'il est difficile d'expliquer la sémantique de leur formation [Zhou et al., 2006]. En effet, cette dernière résulte de la similitude entre les acteurs sociaux et la simple prise en compte de la structure topologique pour identifier les communautés semble être insuffisante. Les chercheurs ont donc conclu que la détection de communautés doit prendre en compte les caractéristiques topologiques et sémantiques conjointement. Dans des travaux récents, il

y a eu des efforts vers une telle direction.

2.1.3.2 Méthodes basées sur la structure topologique et les attributs

Ces approches ont développé les deux sources d'information. En fait, le clustering combine les similarités de structure et des attributs. Les noeuds dans les mêmes communautés devraient être fortement connectées et avoir des attributs similaires. Dans ce contexte, les communautés sont définies par des structures cohésives et des valeurs d'attributs homogènes.

2.1.3.3 Méthodes basées sur la structure topologique et les phénomènes sociaux

Les Réseaux Sociaux présentent des caractéristiques importantes de la communication humaine. En effet, un utilisateur est considéré comme une entité sociale. Les comportements des utilisateurs au sein d'un réseau reflètent la logique de la propagation de l'information et son échange. Ces faits sociaux sont traduits par des phénomènes tels que l'influence et la confiance. Plusieurs approches supposent ces caractéristiques et proposent des modèles d'autocorrélation pour la détection communautaire en tenant compte des caractéristiques structurelles et des phénomènes sociaux en modélisant le graphe social et les cascades d'information.

En particulier, un modèle d'influence sera détaillé dans le chapitre 3.

2.1.4 Discussion

La présente revue de littérature couvre le domaine de la détection communautaire dans les Réseaux Sociaux. En effet, la détection de communauté semble être un outil important pour l'analyse de réseaux complexes et l'étude de structures mésoscopiques. Cependant, cela s'avère être un problème difficile. En particulier, la taille de ces réseaux ne cesse de croître avec une quantité de contenu explosive, ce qui rend l'efficacité du clustering un véritable défi. De plus, le comportement du réseau dépend des interactions humaines. Les utilisateurs publient des profils faisant référence à des attributs contenant des informations telles que la localisation géographique, les intérêts, l'âge, etc. Cependant, tous les utilisateurs ne fournissent pas ces informations. De plus, les nouveaux utilisateurs ne disposent d'aucune information historique ; connu sous le nom de phénomène de démarrage à froid. Ainsi, un autre défi est la capacité de prédire les attributs des nouveaux utilisateurs. En outre, les communautés se chevauchent car les utilisateurs peuvent appartenir à plusieurs groupes d'intérêts en même temps. Par conséquent, il est nécessaire de disposer d'une méthode de détection de communautés qui se chevauchent, car les méthodes disjointes ne

permettent pas de trouver la véritable structure communautaire des réseaux réels.

La description de plusieurs approches existantes et leur observation incluant les avantages et les inconvénients ont été présentées. En outre, nous avons examiné de manière approfondie les classifications existantes des méthodes proposées pour la détection de communautés. À première vue, il existe des paradigmes méthodologiques largement utilisés, notamment la topologie du réseau et les caractéristiques de performance, mais moins de technologies sémantiques. La structure topologique des Réseaux Sociaux est déjà complexe et a attiré l'attention de plusieurs chercheurs. Ceci explique l'abondance des travaux classifiant les algorithmes sur la base de caractéristiques structurelles. Plus récemment, et avec l'évolution des technologies du Web sémantique, de nouvelles approches sont proposées pour tirer parti de la diversité des données issues de l'utilisation des Réseaux Sociaux en ligne. Ces approches prennent en compte, outre la topologie du réseau, les données sociales hétérogènes qui peuvent être traitées avec les technologies sémantiques.

Sur la base de cette évolution et de l'évaluation des travaux existants, nous avons proposé une nouvelle classification des méthodes de détection des communautés [Chouchani and Abed, 2018a]. Cette classification prend en compte les données sociales enrichies avec la sémantique ; ce qui, à notre connaissance, n'a pas été pris en compte dans les autres travaux existants.

2.1.5 Evaluation de la qualité d'une communauté

[Combe, 2013] a distingué entre des critères internes et d'autres externes. Les premiers servent surtout à évaluer relativement à d'autres la qualité des communautés détectées. Parmi ces critères, ceux dont l'utilisation est spécifique à une mesure de distance donnée. D'autres sont utilisés avec des méthodes de classification spécifiques. Aussi, il existe des critères qui ne sont pas spécifiques à une topologie des données.
La seconde catégorie des critères se sert d'une partition comme référence de vérité de terrain (exemple le Taux d'éléments bien classés). Dans certaines solutions, les catégories réelles de communautés et les classes obtenues sont appariées pour évaluer les résultats. De plus, des indices se basant sur différentes approches combinatoires (indice de Rand), probabiliste (Information Mutuelle) ou de la théorie de l'information (Entropie) sont utilisés.

2.2 Analyse des sentiments

L'analyse du ton émotif avec lequel les gens s'expriment aide à mieux comprendre leurs attitudes et réagir en conséquence. Ce processus est connu sous le nom de l'Analyse

des Sentiments. Il s'agit du traitement des opinions, des sentiments et de la subjectivité dans les textes [Pang and Lee, 2008]. En particulier, les utilisations de l'Analyse des Sentiments sont larges et puissantes. Ce domaine est de plus en plus utilisé pour obtenir des opinions et des émotions publiques sur certains sujets d'intérêt dans les Réseaux Sociaux étant des ressources riches en opinions. En effet, la pratique consistant à extraire des connaissances des données du web social est largement adoptée par les entreprises. Il a été démontré que les changements des tonalités sur les Réseaux Sociaux sont corrélés avec les variations du marché boursier [Bollen et al., 2011, Leitch and Sherif, 2017]. En fait, la disponibilité croissante et la popularité des plateformes des Réseaux Sociaux en font les médias de communication principaux en raison de l'énorme quantité de données générées par les utilisateurs. Par conséquent, ils représentent un secteur nouveau et émergent dans le domaine de l'Analyse des Sentiments, traitant de la classification des polarités. Cette classification est une tâche visant à extraire les sentiments "positif" et "négatif", également appelés polarités, sur un sujet d'intérêt spécifique.

Comment les Réseaux Sociaux, qui sont complexes et si omniprésents, peuvent-ils affecter les polarités des sentiments ? Considérons une entreprise qui utilise les Réseaux Sociaux pour vendre ses produits. Ses décideurs savent que les clients qui sont amis dans le réseau achètent des produits similaires. La question qui se pose est, quelles sont les raisons de cette similitude ? Ont-ils des goûts et opinions similaires parce qu'ils sont reliés par une relation personnelle d'amitié ? Ou sont-ils influencés l'un par l'autre puisqu'ils communiquent fréquemment ? C'est sur la base de la réponse à cette question que les décideurs sauront comment interpréter les données commerciales et mettre en évidence leurs plans stratégiques.

L'Analyse des Sentiments dans les Réseaux Sociaux est un nouveau domaine de recherche qui applique les techniques traditionnelles héritées de l'Analyse des Sentiments à l'analyse des Réseaux Sociaux. L'Analyse des Sentiments, en général, suppose que les contenus générés par les utilisateurs sont indépendants et répartis de manière identique. Néanmoins, et compte tenu des informations disponibles dans le réseau de l'utilisateur, l'analyse des Réseaux Sociaux peut améliorer l'Analyse des Sentiments.

2.2.1 Analyse des Sentiments au niveau publication

Afin de déterminer les sentiments exprimés dans les Réseaux Sociaux, plusieurs approches ont été proposées. Traditionnellement, il y a eu des efforts principalement basés sur la classification de la polarité des sentiments des contenus textuels individuels sans tenir compte de l'information sur les sentiments globaux des utilisateurs qui les ont publiés. Au niveau de la publication (**post-level**), la plupart des approches proposées peuvent être

regroupées en deux catégories : celles fondées sur le lexique et celles sur l'apprentissage automatique.

2.2.1.1 Approches basées sur le lexique

D'une part, les approches basées sur le lexique ont généralement tendance à comparer le nombre de mots positifs et négatifs en utilisant des ressources externes prédéfinies et des dictionnaires, ou d'appliquer une propagation d'étiquette sur un graphe d'allongement des mots (lengthening words). Elles ont utilisé des dictionnaires de polarité ou des lexiques tels que *SentiWordNet*[14], *ANEW* [Margaret and Peter, 1999] ou *MPQA*[15] comme ressources externes pour détecter les polarités de sentiment des mots. Par exemple, *MPQA* a été utilisé pour déterminer les mots positifs et négatifs contenus dans les Tweets afin de détecter leurs sentiments. De plus, dans leur approche, Bollen et al. ont prouvé que les mots d'allongement emphatiques, tels que "*cooooool*" , sont fortement associés à la subjectivité et au sentiment [Bollen et al., 2011]. Donc, ils peuvent être considérés comme des mots d'opinion supplémentaires au lexique *MPQA*. Cependant, les termes qui ne sont pas inclus dans les lexiques préconstruits et dans les dictionnaires sont généralement ignorés, ce qui peut fausser les résultats. Ainsi, l'inconvénient des méthodes basées sur le lexique est qu'elles en dépendent fortement. C'est-à-dire leur performance se dégrade considérablement avec la croissance exponentielle de la taille des lexiques. Par exemple, SentiStrength est un système de détection des sentiments basé sur le lexique dans les sites de microblogging [Thelwall et al., 2010]. Les auteurs ont construit leur propre lexique de sentiment composé d'abord de 298 termes positifs et 465 termes négatifs, puis de 2310 mots ainsi que des listes d'émoticônes. L'inconvénient de cette approche est qu'elle dépend fortement du lexique prédéfini.

2.2.1.2 Approches basées sur l'Apprentissage automatique

D'autre part, les algorithmes d'apprentissage automatique supervisé, tels que *Naive Bayes*, *Maximum Entropy* et *Support Vector Machines*, sont utilisés dans les approches basées sur l'apprentissage. Ces approches comportent deux phases : une phase d'entraînement et une phase de prédiction. Dans la première phase, les données d'apprentissage qui sont généralement libellées manuellement sont utilisées pour extraire un ensemble de caractéristiques pour générer un modèle de classification. Les sentiments correspondants aux données non libellées parmi les données de test sont prédites via le modèle de classification précédemment construit. Parmi les caractéristiques qui peuvent être utilisées, il

14. http ://sentiwordnet.isti.cnr.it
15. http ://mpqa.cs.pitt.edu

y a les n-grammes, les bag-of-words, la syntaxe et les fonctionnalités propres à certains Réseaux Sociaux comme le hashtag et les émoticônes. L'inconvénient de ces approches est qu'elles nécessitent beaucoup de données étiquetées mais ceci est obtenu manuellement. Afin de surmonter ce problème, il y avait des tentatives de collecte automatique des données d'apprentissage, appelée surveillance à distance. Un travail pionnier de [Go et al., 2009] a utilisé les émoticônes telles que " :)" et " :(" pour construire un corpus de tweets positifs et négatifs. Les auteurs ont prouvé que les méthodes *SVM* atteignent la meilleure performance soit 82.9%. Cependant, les émoticônes sont parfois rares pour préparer une grande quantité de données pour certains mots clés cibles. Alors d'autres approches ont été proposées telles que l'utilisation des hashtags comme indicateurs de sentiments ou bien des résultats de certains sites tiers d'Analyse des Sentiments tels que Twendz[16], TweetFeel[17] et Sentiment140[18]. De plus, un classifieur construit pour un domaine donné, pourrait ne pas bien fonctionner pour un autre domaine.

2.2.1.3 Approches hybrides

Récemment, certains travaux ont combiné ces deux approches. Ils ont obtenu de meilleurs résultats en termes de prédiction de polarité. Dans cette perspective, nous distinguons deux catégories de méthodes. D'abord il y a eu des efforts pour construire un système qui intègre deux classifieurs développés séparément basés sur les deux approches déjà évoquées. Dans [Akshi and Teeja, 2012], une méthode basée sur l'apprentissage automatique a été utilisée pour détecter l'orientation sémantique des adjectifs et une autre méthode basée sur le lexique pour celle des verbes et des adverbes. Le sentiment global est ensuite calculé en utilisant une interpolation linéaire des deux méthodes. Dans un second temps, certaines méthodes ont proposé d'incorporer les informations de lexique dans un modèle de classification basé sur l'apprentissage automatique.
La plupart de ces travaux sont indépendants d'un centre d'intérêt (**target**), c'est-à-dire que classer la polarité des messages est général et non conforme à un sujet d'intérêt cible spécifique. Ils utilisent des classifieurs basés sur l'apprentissage automatique ou des lexiques où toutes les caractéristiques utilisées sont indépendantes de la cible. Cependant, les utilisateurs peuvent se référer à plusieurs sujets cibles dans une seule publication, donc il n'est pas raisonnable d'utiliser des approches indépendantes de la cible. [Long et al., 2011] ont été les premiers à proposer des analyses de sentiment dépendantes des cibles dans le réseau social de Twitter.

16. http ://twendz.waggeneredstrom.com
17. http ://www.tweetfeel.com
18. http ://www.sentiment140.com/

2.2.2 Analyse des Sentiments au niveau utilisateur

De plus, au-delà de la polarité des textes individuels, il est important de reconnaître le sentiment de chaque utilisateur. Cela a été abordé dans des études plus récentes. Dans [Kim et al., 2013], les auteurs ont proposé un modèle de prédiction de l'opinion des utilisateurs qui s'appuie sur des techniques de filtrage collaboratif où prédire un sentiment repose sur les caractéristiques des contenus des messages Twitter. Des approches ont essentiellement supposé que les sentiments globaux des utilisateurs sont estimés en agrégeant les sentiments de leurs messages dans leurs historiques. Cependant, l'agrégation des sentiments des messages est susceptible de générer des résultats insatisfaisants en raison de leur ambiguïté (écrits en langage naturel) et du format de données non structuré, ce qui peut induire des erreurs et du bruit.

Afin d'améliorer l'analyse des sentiments au niveau de l'utilisateur (**user-level**), plusieurs chercheurs ont exploité les relations sociales et la structure du réseau. Leur but était de déduire les sentiments des utilisateurs à partir de leurs contenus publiés en intégrant leurs relations d'amitié. Leur motivation, en référence au phénomène de l'homophilie, est que les utilisateurs connectés peuvent être plus susceptibles d'avoir des opinions similaires. Dans [Tan et al., 2011], un sentiment général d'un utilisateur a été déterminé en regardant ses tweets et à qui il est connecté. Cependant, il a été démontré que, compte tenu des connexions d'amitié, c'est une hypothèse faible pour la modélisation de l'homophilie. [Pozzi et al., 2013] a proposé une solution pour la classification de polarité des sentiments des utilisateurs qui intègre les contenus des publications avec les relations d'approbation. Des études récentes ont défini des modèles pour inférer simultanément les sentiments au niveau des publications et au niveau des utilisateurs. Les sentiments des publications sont influencés par ceux des utilisateurs, et de la même manière, les sentiments des publications peuvent influencer les sentiments des utilisateurs [Nozza et al., 2014]. Dans *la table* 2.3, nous détaillons certaines des approches mentionnées.

2.2.3 Discussion

Dans le *tableau* 2.3, nous décrivons certaines approches d'Analyse des Sentiments. Outre les approches aux niveaux publication et utilisateur, certaines approches combinent les deux niveaux.

Nous résumons les différentes catégories dans la *figure* 2.2.

Méthode	**Niveau**	**Technique**	**Homophilie**	**Analyse de texte**	**Dépendant du centre d'intérêt**
[Alec et al., 2009]	Publication	Apprentissage automatique	Non	Requêtes à termes et émoticones	Non
[Luciano and Junlan, 2010]	Publication	Modèle de classification pour la prédiction des polarités	Non	Méta-données des mots et syntaxe	Non
[Michael et al., 2011]	Publication	Classifieur de maximisation de l'entropie	Relations d'abonnement de Twitter	Types des mots dans des lexiques	Non
[Tan et al., 2011]	Utilisateur	Apprentissage semi-supervisé	Raltions d'amitié	Non	Oui
[Pozzi et al., 2013]	Utilisateur	Modèle de classification des polarités des utilisateurs	relations d'approbation	Non	Oui
[Long et al., 2011]	Publication	Classifieur *SVM* binaire	Non	Etiquetage morpho-syntaxique	Oui
[Laura et al., 2013]	Publication et utilisateur	Classifieur *SVM-light*	Relations d'abonnement	caractéristiques, vecteurs de mots TFIDF	Oui

TABLE 2.3 – Comparaison de quelques méthodes d'Analyse des sentiments dans les Réseaux Sociaux

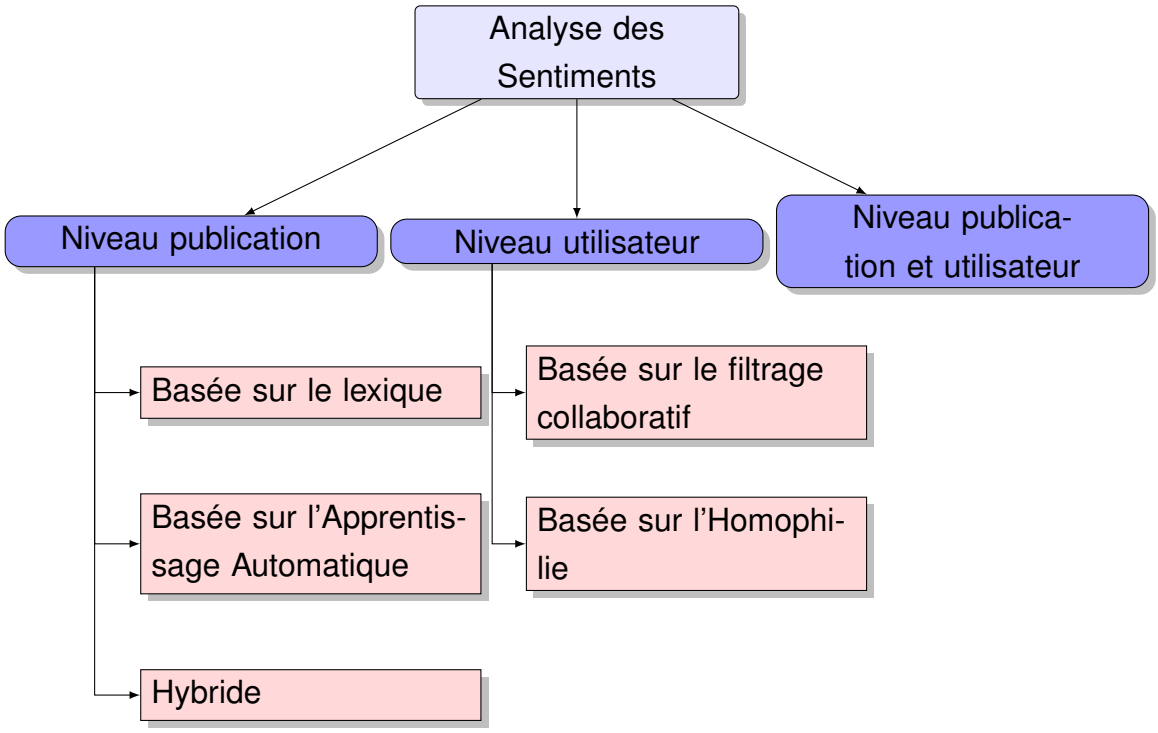


FIGURE 2.2 – Catégories des méthodes d'Analyse des Sentiments

2.3 Applications sociales interactives

Les Réseaux Sociaux sont adaptés comme des interfaces front-end pour les applications interactives. Par conséquent, le contexte social est considéré comme une nouvelle dimension dans la conception de tels logiciels. En effet, un Réseau Social comme Twitter a été utilisé pour détecter les tremblements de terre en Chine [Li, 2010]. Ce travail a exploité la nature en temps réel des Tweets pour enquêter sur les interactions de certains événements, comme les catastrophes naturelles, et pour les détecter. En outre, ce réseau était l'intermédiaire pour l'État japonais pour diffuser des informations lors de la catastrophe nucléaire de Fukushima [Li et al., 2014]. De plus, certaines plateformes telles que HyperTwitter [Hepp, 2010] sont utilisées pour la construction de représentations de connaissances structurées collaboratives basées sur le Réseau Social *Twitter* .

Les applications sociales sont des applications interactives qui sont étendues avec de nouvelles exigences et fonctionnalités permettant leur interaction avec les plateformes des Réseaux Sociaux. Dans cette perspective, des outils et des approches ont été proposés pour l'automatisation de la construction de ces applications. Ces approches sont destinées aux utilisateurs qui n'ont pas de compétences de programmation élevées, en particulier dans le domaine des Réseaux Sociaux, et visent à faciliter la tâche de développement.

Dans ce qui suit, nous présentons certaines approches.

[Scott, 2000] ont développé une application d'assistance sociale en utilisant un Réseau Social implicite construit suite à une analyse contextuelle des interactions des utilisateurs collectées de leurs Réseaux Sociaux comme les courriels, les SMS et les appels téléphoniques. Cette application est destinée à aider les utilisateurs finaux dans leurs besoins de communication. Elle est basée sur l'analyse des interactions sociales en calculant une mesure de proximité sociale entre deux personnes. Une approche pour la construction d'un système d'analyse des interactions sociales a été proposée. Son architecture comprend trois étapes, à savoir, la collecte des données des différents dispositifs et identités des utilisateurs, l'analyse continue du réseau et une base de données des profils sémantiques des utilisateurs. Ce système tente de construire et qualifier des réseaux implicites contextuels des utilisateurs au delà de leurs relations explicites. Ces réseaux sont utilisés pour trouver automatiquement le contact avec la personne avec la plus grande probabilité de proximité. Cette vision contextuelle des interactions peut servir de back-end à de nombreuses applications.

Social *BPM* (Business Process Modeling), l'extension sociale du Business Process, permet de fusionner la gestion des processus d'affaires avec les plateformes de Réseaux Sociaux. *BPMN4Poeple* est une approche basée sur les modèles pour la mise en oeuvre sociale des processus d'affaires [Brambilla et al., 2011]. Elle fournit une notation spécifique pour décrire les comportements du *BPM* Social, une méthodologie et un cadre technique permettant aux entreprises de mettre en oeuvre des processus intégrés aux Réseaux Sociaux privés ou publics. L'architecture proposée pour la conception, l'implémentation, le déploiement et le suivi des applications de *BPM* Social repose sur deux phases. Une première phase de conception comprend un environnement de développement intégré dirigé par les modèles. La seconde phase est celle d'exécution où les applications générées automatiquement sont exécutées sur des plateformes connectées à un ou plusieurs Réseaux Sociaux via leurs *API*s de services web correspondants. *BPMN4Poeple* étend *BPMN* (Business Process Modeling Notation) avec de nouveaux types de tâches pour représenter un flux de données complexes. Ces dernières proviennent du domaine social où les tâches représentent les interactions entre les acteurs du processus (vote, invitation à événement, publication sociale...). Cette extension fournit la possibilité de capturer des activités sociales et des événements à partir d'un Réseau Social, d'utiliser des profils d'utilisateurs sociaux pour accéder à la plateforme *BPM*, d'effectuer des activités sociales et de modéliser explicitement les données et contenus sociaux. Cependant, l'utilisation de *BPMN* pour la définition de processus nécessite beaucoup de temps d'apprentissage pour les utilisateurs finaux.

[Brambilla and Mauri, 2012] ont défini une extension de la notation WebML (un

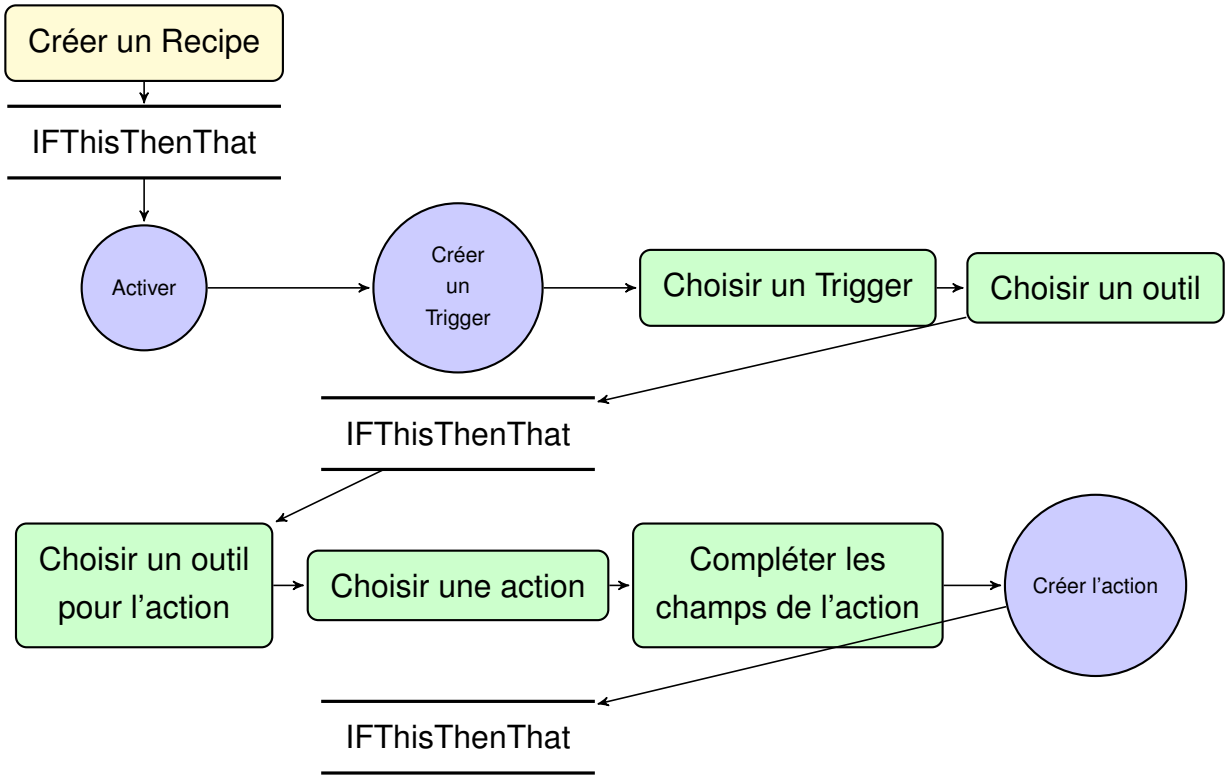


FIGURE 2.3 – Principe de la méthode *IFTTT*

langage spécifique au domaine conçu pour modéliser des applications web) [Ceri et al., 2002]. Cette extension comprend un ensemble de concepts de modélisation encapsulant la logique de l'interaction avec les plateformes des Réseaux Sociaux. Ceci est utilisé dans la définition de modèles de conception pour répondre aux besoins standards des entreprises.

IFTTT (*IF This Then That*) est une approche pour l'automatisation des tâches qui se base sur la création de processus en une seule étape [ift]. Ce sont les utilisateurs qui créent des règles simples de type *ECA* (Event-Condition-Action) sur différents Réseaux Sociaux. Cette création se fait en combinant des déclencheurs et des actions avec les réseaux associés. En effet, les processus sont exécutés une fois les déclencheurs sont déclenchés (voir *figure* 2.3).

SimpleFlow, inspiré par l'*IFTTT*, est un outil proposé pour la conception et l'exécution des applications sociales [Laconich et al., 2013]. Il est basé sur de simples processus qui s'exécutent sur les Réseaux Sociaux. Cependant, ces processus peuvent avoir plus d'une seule étape. Cette approche cible les utilisateurs finaux n'ayant aucune ou peu d'expérience en programmation pour les Réseaux Sociaux. Cet outil est basé sur deux phases : une phase de conception et une phase d'exécution. Les utilisateurs ont juste à concevoir des processus en concaténant des actions des Réseaux Sociaux. Ils peuvent définir plu-

sieurs déclencheurs et actions pour respectivement la condition "If this" et la partie "Then that" [Laconich et al., 2013]. A l'exécution de ces processus, SimpleFlow définit interconnecte les pages web du modèle en fonction de la conception définie précédemment.

Plus récemment, une approche pour automatiser la construction rapide d'applications interactives a été proposée [Segura et al., 2014]. Ce travail a introduit la notion d'applications "post-based", c'est-à-dire basées sur les publications textuelles dans les Réseaux Sociaux. Ils ont proposé une approche *MDE* pour leur génération automatique et rapide. Une première tâche consistait à extraire les informations pertinentes des Tweets adressés à l'application en question. Pour ce faire, la solution comprend un *DSL* pour l'expression des concepts à chercher dans les Tweets. Ce *DSL* est connecté à WordNet, une base de donnée lexicale pour la langue anglaise. En effet, les concepts de l'application sont définis par le concepteur en utilisant le DSL proposé. Ce dernier est destiné spécifiquement au réseau Twitter. Son modèle est composé de concepts. La seconde tâche est destinée à exécuter des requêtes sur les tweets retenus suite à l'application du premier *DSL*. Les données ainsi extraites des requêtes avec celles fournies par le Système d'Information sont utilisées pour synthétiser des tweets ou messages privés destinés vers les utilisateurs. Donc un deuxième *DSL* était défini pour décrire le traitement logique de l'application.

Dans la *figure* 2.4, nous comparons quelques méthodes de génération d'applications sociales.

Méthode	**Processus**	**Mushups**	**Ingénierie Web**	**Spécifique**	**Générique**
IFTTT	Oui	Oui	Non	Non	Oui
BPM4Poeple	Oui	Non	-	Oui	-
Simple Flow	Non	Oui	-	-	-
Applications Post-based	-	Non	-	Oui	Non

TABLE 2.4 – Comparaison de quelques méthodes de génération d'applications sociales

2.4 Conclusion

De nos jours, les Réseaux Sociaux ont émergé en supportant un large éventail d'utilisateurs et d'intérêts dans le monde entier et en incitant les chercheurs à exploiter les

données sociales numériques. En effet, de nombreux problèmes de recherche complexes liés à ce domaine sont apparus. Plus précisément, la détection de communautés d'intérêts revêt une grande importance et peut avoir de nombreuses applications concrètes. Une conclusion remarquable est que la plupart des méthodes existantes se concentrent sur la structure topologique des réseaux. Ainsi, la plupart d'entre elles appartiennent à la catégorie des méthodes basées sur la topologie. Cependant, les travaux récents s'intéressent de plus en plus à la sémantique des Réseaux Sociaux.

Dans notre travail, nous avons proposé une nouvelle classification des approches existantes pour résoudre ce problème. En fait, nous avons étudié de manière approfondie les éléments pris en compte pour identifier des communautés homogènes et précises. Nous avons classé ces approches en trois catégories. Sur la base de cette nouvelle classification, nous concluons qu'aucune approche ne prend en compte à la fois la structure topologique du réseau, les attributs des utilisateurs et la sémantique. Ainsi, il sera intéressant de tirer parti de toutes ces informations pour détecter des communautés de plus en plus cohérentes. Une fois les communautés d'intérêts détectées, il convient d'étudier les sentiments des membres au sein de ces communautés. Pour ce qui est de l'Analyse des Sentiments, nous avons présentés des approches qui s'intéressent aux niveaux publication et utilisateur ou bien les deux ensembles.

Nous remarquons que plusieurs travaux ont profité du phénomène d'homophilie dans la résolution des problèmes de détection communautaire ainsi que d'Analyse des Sentiments. Cependant, aucun travail n'a exploité le phénomène d'influence sociale qui est aussi crucial dans les Réseaux Sociaux.

De plus, les applications sociales présentées manquent de capacités de personnalisation. Par conséquent, en exploitant les informations disponibles dans les Réseaux Sociaux, l'originalité de notre travail réside dans la personnalisation des systèmes interactifs sociaux en utilisant les communautés d'intérêts.

C'est sur ces points que porte notre travail dans les prochains chapitres.

CHAPITRE 3

Approche de détection des communautés d'intérêt

Sommaire

Dans ce chapitre, nous présentons notre approche pour répondre à la problématique de détection des communautés d'intérêt basée sur les Réseaux Sociaux. Ayant présenté l'état de l'art dans le chapitre précédent, nous allons pouvoir décrire en détails notre solution et nos contributions. Cette dernière vise à s'appuyer sur les travaux existants dans l'Analyse des Réseaux Sociaux pour répondre à la problématique énoncée. Pour ce faire, ce chapitre est structuré comme suit. Dans un premier temps, nous présentons un modèle de profil "social" des utilisateurs. Ensuite, nous introduisons notre approche générale de détection des communautés en se basant sur des algorithmes de dérivation de connaissances explicite et implicite à partir des profils ainsi construits. Puis nous détaillons la procédure d'implémentation algorithmique de cette proposition. Enfin, nous terminons le chapitre par une conclusion qui récapitule les principaux éléments de notre contribution.

3.1 Modèle de données : 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 :

Définition 12 (Données sociales)**.** *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 modélisation 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 social de profil utilisateur 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 en l'exploitant 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.

3.1.1 Sélection des données

Cette phase fait intervenir deux éléments qui sont : les producteurs de données 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 utilisateur, 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 ; il s'agit ainsi d'un réseau appelé "boule de neige". 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 cette thèse, 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 (jusqu'à un niveau maximal spécifié). Quant aux données sociales, elles comprennent les

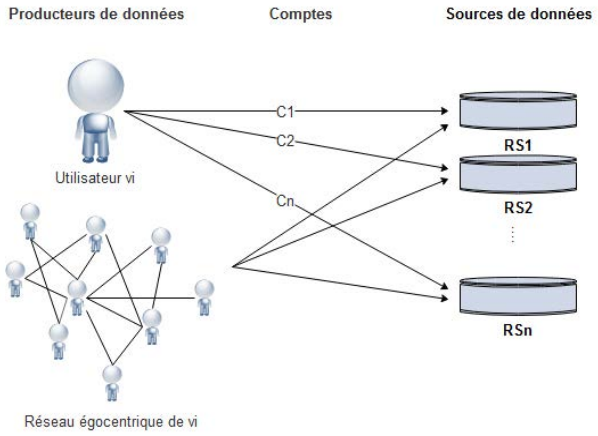


FIGURE 3.1 – Producteurs et sources de données sociales

attributs des utilisateurs, leurs relations, leurs publications ainsi que leurs intérêts. Cependant, nous sommes intéressés à extraire seulement les données pertinentes. Concernant les publications, il s'agit de celles qui sont en rapport avec un sujet d'intérêt bien spécifié. Une technique de filtrage d'information est donc nécessaire pour extraire les "bonnes" publications. Ce modèle est détaillé dans le chapitre 4.

3.1.2 Prétraitement 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. En particulier, 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 pourrait être disponible dans un autre. Pour ce faire, une identification unique de chaque utilisateur et une collection et récupération des données sont à effectuer.

3.1.3 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 et les interactions 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 se 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.

3.1.4 Fouille de données

Afin d'analyser les données et extraire des connaissances, des algorithmes du domaine de fouille de données (Data Mining) sont appliqués. Il existe différentes techniques utilisées suivant les modèles de profils utilisateurs, qui sont les suivantes :

3.1.4.1 Modélisation comportementale

Les comportements des utilisateurs sont enregistrés dans des structures de données historiques. Différentes méthodes existent pour la modélisation du comportement comme les règles d'association, les modèles de Markov et les arbres de décision.

3.1.4.2 Modélisation des centres d'intérêt

Une fonction de préférence est définie afin de représenter le degré d'intérêt ou désintérêt d'un utilisateur à un concept ou sujet donné. L'extraction des préférences se fait à l'aide d'approches directes en demandant explicitement aux utilisateurs de dire ce qu'ils préfèrent ; ou approches semi-directes en leur demandant de noter les centres d'intérêt ou utiliser des approches indirectes en capturant les préférences à partir des données disponibles.

Pour une meilleure compréhension des contenus publiés par les utilisateurs, notre but est d'identifier automatiquement les sujets qui les intéressent en fonction de leurs publications. En effet, généralement, les utilisateurs n'expriment pas explicitement leurs centres 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 centre 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., 2003b] 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.

TABLE 3.1 – Matrice "documents-mots"

	mot 1	mot 2	...	mot p
doc 1				
...				
doc n				

TABLE 3.2 – Matrice "sujets-mots"

	mot 1	mot 2	...	mot p
sujet 1				
...				
sujet k				

TABLE 3.3 – Matrice "documents-sujets"

	topic 1	topic 2	...	topic k
doc 1				
...				
doc n				

L'ensemble des données peut être initialement décrit sous forme de trois matrices : qui sont décrites dans les *tableaux* 3.1, 3.2 et 3.3.

Soient k le nombre des sujets d'intérêt, n le nombre des documents et p le nombre des mots. Les valeurs dans les matrices caractérisent l'association pouvant s'agir de coefficients de combinaison linéaire ou des probabilités.
Etant un modèle probabiliste génératif, *LDA* suppose un processus pour générer chaque document à l'aide de facteurs latents, comme suit :

- Sélection d'un sujet k pour le mot j ;
- Distribution des mots pour chaque sujet suivant la distribution de Dirichlet [Blei and Jordan, 2006] ;
- Sélection d'un sujet pour chaque couple "mot-document" ;
- Distribution des sujets pour chaque document.

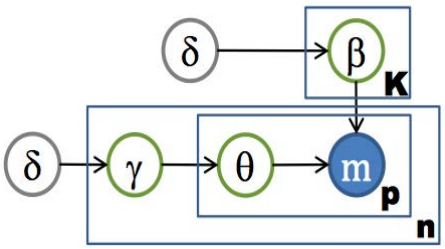


FIGURE 3.2 – Représentation graphique du modèle *LDA*.

Une représentation graphique du modèle *LDA* est décrite dans la *figure* 3.2. 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.

3.1.4.3 Modélisation des intentions

La modélisation des intentions repose sur des systèmes de classification comportant des catégories prédéfinies. Cette modélisation se présente à un niveau plus élevé que les deux autres types de modélisation. En effet, elle est plus approfondie, et tient compte du contexte comme étant une information utilisée pour caractériser la participation de l'utilisateur lors de l'interaction.

Parmi ces techniques, nous allons utiliser des algorithmes de classification semi supervisée (pour la classification des sentiments des utilisateurs) et d'autres de prédiction (une classification supervisée pour la prédiction des intérêts) de la catégorie de la modélisation des intentions (intention modeling).

3.1.5 Représentation des profils des utilisateurs

Les profils des utilisateurs peuvent être représentés suivant différents modes qui sont proposés : représentation ensembliste, représentation multi-dimentionnelle ou représentation sémantique.

Dans notre travail, nous optons pour une représentation sémantique conceptuelle des profils des utilisateurs 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 standard 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 (agrégation des données sociales). Plus particulièrement, *FOAF* comprend, parmi d'autres, les propriétés suivantes :

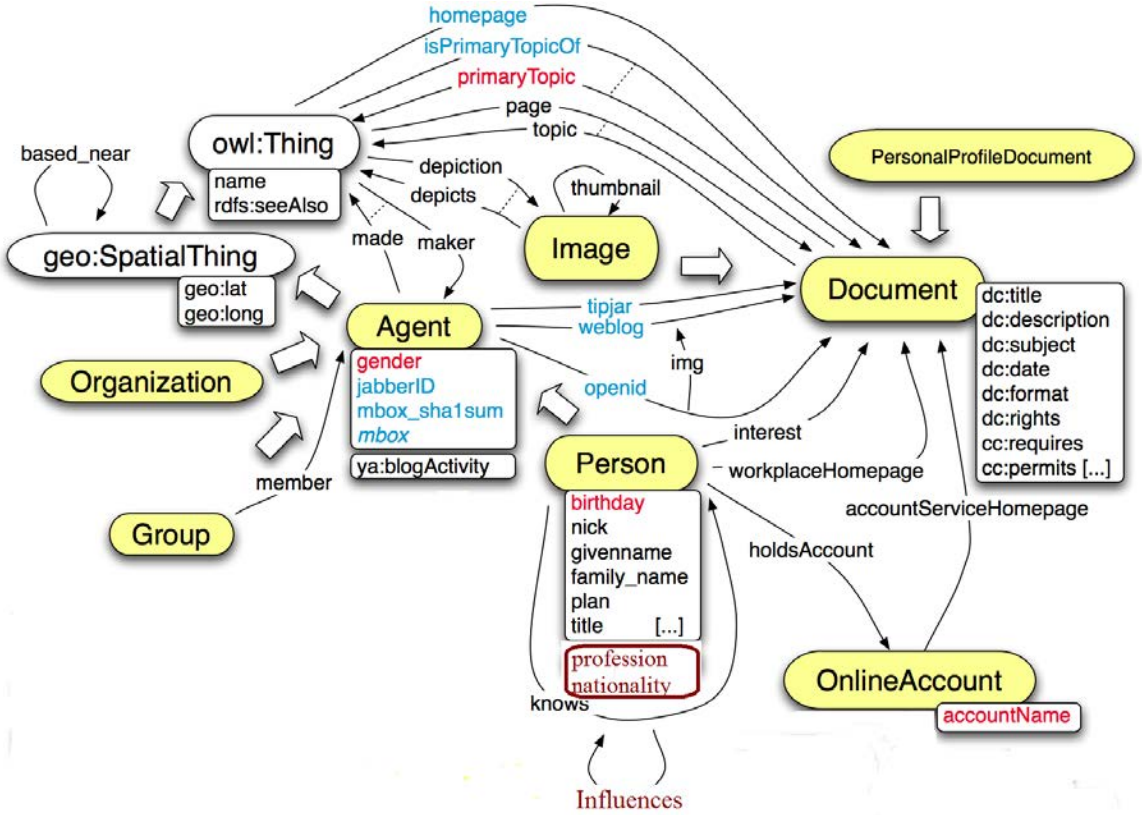


FIGURE 3.3 – Ontologie FOAF étendue

- ***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 sexe 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 l'occupation (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* 3.3).

- ***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 :occupation*** : c'est une chaîne qui représente la profession d'une personne.

3.2 Approche générale

Soient q un sujet d'intérêt et $C(q) = \{v_i \in V\}$ la communauté d'intérêt des utilisateurs v_i intéressés en q parmi tout l'ensemble V des utilisateurs du Réseau Social. 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 sociales comme les relations et les publications.

3.2.1 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 [Chouchani and Abed, 2018b]. En partant de la définition des communautés d'intérêt retenue dans le premier chapitre, nous considérons la topologie ainsi que la sémantique conjointement. A partir de la modélisation des utilisateurs et leurs sujets d'intérêt, nous construisons les regroupements autour ce 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 :

- 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 noeuds actifs repérés ainsi que les liens entre eux forment la communauté d'intérêt ;

- Pour chaque noeud dans les réseaux égocentriques des noeuds de la communauté qui sont ordonnés selon leurs degrés, intégrer les noeuds actifs par rapport à q à $C(q)$ et inférer parmi les noeuds passifs ceux qui pourraient être intéressés en q et les ajouter à $C(q)$;

- Pour chaque noeud 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.

3.2.2 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 noeud candidat et une communauté, et retourne soit vrai si le noeud 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 noeuds qui sont intéressés positivement et ceux qui sont intéressés négativement au sujet q.

3.3 Procédure d'implémentation

3.3.1 Détails de l'étape 1 : Formation

Un algorithme de parcours du réseau égocentrique d'un noeud donné visite séquentiellement tous les noeuds 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 noeud 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.

Mesure	**Définition formelle *SPARQL***

$int_{<topic>}$	$SELECT\ ?name\ ?interest$
	$WHERE$
	{
	$?x\ foaf:name\ ?name$
	$?x\ foaf:topic_interest\ ?interest$
	$FILTER\ (?interest, "topic")$
	}

$int_{<person,topic>}$	$SELECT\ ?name\ ?interest$
	$WHERE$
	{
	$?x\ foaf:name\ ?name$
	$?x\ foaf:topic_interest\ ?interest$
	$FILTER\ (?interest, "topic")$
	$FILTER\ (?name, "person")$
	}

$deg_{<type,length>}(y)$	$SELECT\ ?y\ count(?x)$
	$WHERE$
	{
	$?x\ \$path\ ?y$
	$FILTER(match\ (\$path, star(param[type])))$
	$FILTER(pathLength\ (\$path) <= param[length])$
	}
	$UNION$
	{
	$?x\ \$path\ ?y$
	$FILTER(match\ (\$path, star(param[type])))$
	$FILTER(pathLength\ (\$path) <= param[length])$
	}
	$GROUP\ BY\ ?y$

TABLE 3.4 – Définition formelle dans SPARQL des mesures paramétrées sémantiquement

Le niveau maximal à considérer est passé comme paramètre. Pour chaque noeud, ***ENES*** va procéder à une acquisition explicite des connaissances.

L'acquisition explicite des connaissances nécessite la participation active des utilisateurs en créant leurs profils eux-mêmes. En effet, ils seraient très probablement responsables de l'exposition de leurs données personnelles.

Nous définissons ***EAK*** (**E**xplicit **A**cquisition of **K**nowledge) qui est une fonction d'acquisition explicite des connaissances à partir de l'ontologie *FOAF* qui représente les profils sociaux des utilisateurs.

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

1: $C(q) = \emptyset$; ▷ Initialisation de la communauté d'intérêt $C(q)$
2: $level = 0$;
3: $file = (v_i)$;
4: **while** $file \neq ()$ **do**
5: $v = défile(file)$;
6: **for** $v_j \in voisins(v)$ **do**
7: **if** v_j n'est pas marqué enfilé **then**
8: Marquer v_j comme enfilé;
9: $Enfile(file, v_j)$;
10: **if** $v_j = EAK(v_j, q)\ and\ v_j \notin C(q)$ **then**
11: $C(q) = C(q) \cup \{v_j\}$;
12: **end if**
13: **end if**
14: **end for**
15: **end while**
16:
17: **for** $v_j \in C(q)$ **do** ▷ Calcul de la centralité de degré locale
18: $CDL(v_j, C(q))$;
19: **end for**
20:
21: **for** $v_j \in C(q)$ **do** ▷ Contraction
22: **if** $CDL(v_j, C(q)) < k$ **then**
23: $C(q) = C(q) \backslash \{v_j\}$;
24: **end if**
25: **end for**
26: **return** $C(q)$;

Elle permet d'en extraire ceux qui sont actifs par rapport à l'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* 3.4). La plupart des propriétés dans l'analyse des Réseaux Sociaux peuvent être calculées directement en utilisant des requêtes *SPARQL*. En effet, ces dernières peuvent être utilisées pour interroger les données sociales contenues dans *FOAF*. Elles sont bien adaptées pour extraire des informations riches et pertinentes à partir des graphes *RDF* et préparer les étapes requises des algorithmes suivants. En particulier, il s'agit d'interroger le graphe social pour identifier les intérêts explicitement énoncés par les utilisateurs.

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 :

$$CD(v_i) = \frac{d(v_i)}{n-1} \tag{3.1}$$

Où $d(v_i)$ est le degré du noeud v_i dans le réseau, et $n-1$ le nombre total des connexions directes.

Nous définissons la **Centralité de Degré Locale** (*CDL*) 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.

3.3.2 Détails de l'étape 2 : Evolution

Sur le plan pratique, la détection des communautés doit s'appuyer sur une structure incrémentale 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 consiste 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 disponibles. La question qui se pose est de savoir quelles caractéristiques sociales dans ces profils corrèlent à la similarité d'intérêt. En effet, notre modèle (voir la *figure* 3.4) 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 sujets d'intérêt et aux informations contextuelles qui peuvent être calculées à partir des données sociales représentées dans les profils utilisateurs.

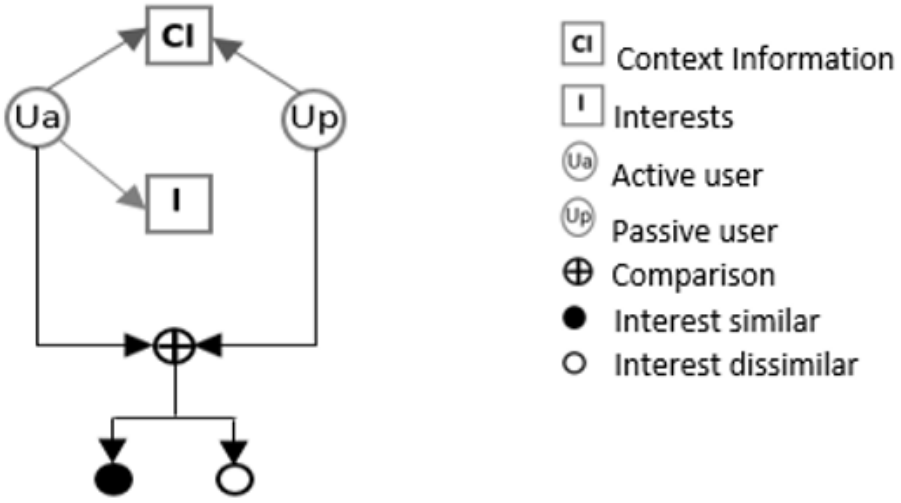


FIGURE 3.4 – Modèle de prédiction des intérêts

Cette étape repose sur un parcours en profondeur du réseau égocentrique d'un noeud afin d'effectuer une acquisition implicite de connaissances. ***ENIS*** (**E**gocentric **N**etwork **I**mplicit **S**earch) visite un noeud v_i puis choisit comme noeud 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 noeud précédent et ainsi de suite. A chaque fois qu'un noeud visité est passif (***EAK*** retourne "Faux"), un algorithme de prédiction est appliqué afin de déterminer l'intérêt potentiel de ce noeud en q.

Algorithm 2 $ENIS(v_i,)$

```
Marquer v_i comme visité ;                          ▷ Parcours du réseau égocentrique
for v_j ∈ voisins(v_i) do
    if v_j n'est pas marqué visité and v_j = EAK(v_j, q) and v_j ∉ C(q) then
        if EAK(v_j, q) = Faux and v_j ∉ C(q) then
            Prédiction(v_j, q)
        end if
        ENIS(v_j) ;
    end if
end for
```

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

3.3.2.1 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 :

$$sim_t(i,j) = 1 - |DT'_{it} - DT'_{jt}| \tag{3.2}$$

- 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 :

$$s = \frac{\sum_{(u,v)\in C} sim_t(u,v)}{\|C\|} \tag{3.3}$$

3.3.2.2 Homophilie

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 :

$$h_t = \frac{N_t}{(\sum_{t=1}^{n} N_t) \div n} \tag{3.4}$$

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

$$w_t = \frac{h_t}{n} \tag{3.5}$$

Et l'entropie d'intérêt est donc :

$$H(I_u) = -\sum_{x_i \in I_u} w(x_i) \log w(x_i) \tag{3.6}$$

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.

3.3.2.3 Influence sociale

Nous introduisons d'abord quelques notations que nous utiliserons dans le reste du manuscrit (voir *tableau* 3.5).

Symbole	Description
V_q	Ensemble des utilisateurs intéressés en q
P_q	Ensemble des publications à propos de q
P	Ensemble de toutes les publications
$p_{i,j}$	Publication j de l'utilisateur i
$L(p_{i,j})$	Ensemble des utilisateurs ayant effectué un "like" sur $p_{i,j}$
$D(p_{i,j})$	Ensemble des utilisateurs ayant effectué un "commentaire" sur $p_{i,j}$
$S(p_{i,j})$	Ensemble des utilisateurs qui ont effectué un "partage" sur $p_{i,j}$
$N(v_i)$	Ensemble des voisins de v_i
$F(v_i)$	Ensemble des abonnés (followers) de v_i
$P(v_i)$	Ensemble des publications de v_i à propos de q

TABLE 3.5 – Notations

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.

En mettant l'accent sur le potentiel d'un individu à amener son "ami" à s'engager dans un certain sujet d'intérêt, nous définissons des mesures d'influence qui reposent sur les métriques définies ci-dessus. Pour ce faire, nous introduisons les fonctions suivantes :

Définition 13 (LCS)**.** ***L**ike **C**omment **S**hare est une fonction qui détermine si un utilisateur $v_j \in V$ a effectué une ou plusieurs activités interpersonnelles ("likes", commentaire, partage) sur une publication $p_{i,k}$ publiée par un utilisateur v_i dans un Réseau Social, (sachant que les deux utilisateurs sont déjà connectés dans le réseau).*

$$LCS(v_i, p_{i,k}, v_j) = \begin{cases} 1 & if : v_j \in (L(p_{i,k}) \cup D(p_{i,k}) \cup S(p_{i,k})) \\ 0 & otherwise \end{cases} \tag{3.7}$$

Définition 14 (ROI)**.** ***R**atio **O**f **I**nfluence représente la mesure de l'influence d'un utilisateur particulier v_i sur un autre utilisateur v_j par rapport à toutes les publications de v_i à propos d'un sujet d'intérêt q. Ce ratio est proportionnel au nombre d'activités interpersonnelles de v_j sur les publications de v_i, et aux relations d'abonnement (follower/followed) entre eux.*

$$ROI(v_i, v_j) = \frac{\sum_{p \in P(v_i)}(LCS(v_i, p, v_j))}{|P(v_i)|} + Fol(v_j, v_i) \tag{3.8}$$

Où Fol(v_j, v_i) est égale à 1 si v_j est abonné à v_i ($v_j \in F(v_i)$) et 0 sinon.

Définition 15 (MOI)**.** ***M**agnitude **O**f **I**nfluence est la valeur moyenne quadratique de **ROI** pour tous les amis d'un utilisateur v_i. Elle est généralement utilisée pour mesurer l'amplitude d'une quantité variable. Dans notre cas, elle indique l'amplitude du rapport d'influence pour différents amis d'un utilisateur particulier dans le Réseau Social. MOI reflète l'influence totale d'un utilisateur sur son réseau.*

$$MOI(v_i) = \sqrt{\frac{\sum_{v_j \in N(v_i)}(ROI(v_i, v_j)^2)}{|P(v_i)|}} \tag{3.9}$$

Afin de donner une structure formelle du Réseau Social couplée avec les informations d'influence, nous introduisons et définissons la notion de réseau d'influence (**influence newtork**). Ce dernier est représenté par un graphe hétérogène où les nœuds et les arêtes peuvent être de types différents. Ce graphe comprend les deux types d'informations : publications et relations d'influence.

Définition 16 (Directed Influence Graph)**.** *Étant donné un sujet d'intérêt q, le réseau d'influence* ***Directed Influence Graph*** *est un quadruple* DIG_q *= {*V_q*,* E_q*,* $X_{q,v}$*,* $X_{q,e}$*}, où* V_q *= {*v_1*, .. ,* v_n*} l'ensemble des utilisateurs intéressés par q ;* E_q *={(*v_i*,*v_j*) |* v_i*,*$v_j \in V_q$*} est l'ensemble des liens d'influence dirigés (arcs) (ce qui signifie que* v_i *influence* v_j *à propos de q, sachant que* v_i *et* v_j *sont déjà connectés dans le réseau par une relation d'amitié);* $X_{q,v}$ *= {*$MOI(v_i)|v_i \in V_q$*} est l'ensemble des poids assignés aux nœuds ;* $X_{q,e}$ *= {*ROI *(*v_i*,* v_j*) |* v_i*,* $v_j \in V_q$ *} est l'ensemble des poids assignés aux arcs.*

Définition 17 (Normalized Directed Influence Graph)**.** *Etant donné un graphe* DIG_q*, un graphe normalisé* ***Normalized Directed Influence Graph*** *est un triplet* $NDIG_q$ *= {*V_q*,* E_q*,* $C_{q,e}$*}, où* $C_{q,e}$ *= {*$w(v_i, v_j) = \frac{ROI(v_i, v_j)}{MOI(v_i)} | v_i, v_j \in V_q$*} est l'ensemble des poids normalisés des arcs d'influence.*

Définition 18 (Heterogeneous Normalized Directed Influence Graph)**.** *Soit un graphe normalisé* $NDIG_q$*, un graphe hétérogène normalisé* ***Heterogeneous Normalized Directed Influence Graph*** *est un cinquplet* $HNDIG_q$*={*$V_q, E_q, C_{q,e}, P_q, X_{q,v}$*}.*

Un exemple d'un $HNDIG_q$ est présenté dans la *figure* 3.5.

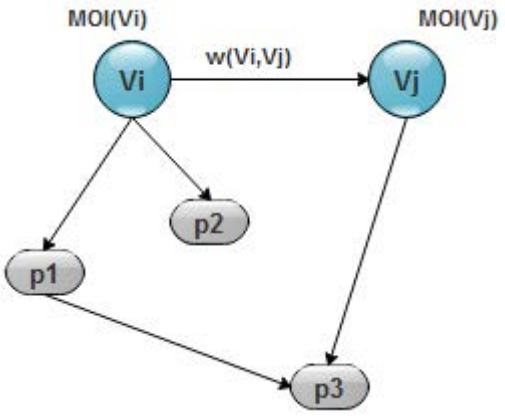


FIGURE 3.5 – Un exemple de $HNDIG_q$

Définition 19 (Normalisation du degré entrant)**.** *La* ***fonction de normalisation du degré entrant*** $\delta(v_i)$ *permet la normalisation du nombre des utilisateurs abonnés (degré entrant) d'un utilisateur donné* v_i *dans l'intervalle [0,1].*

$$\delta(v_i) = \frac{\ln(|F(v_i)| - \min_{v_j \in V}|F(v_j)|)}{\ln(\max_{v_j \in V}|F(v_j)| - \min_{v_j \in V}|F(v_j)|)} \tag{3.10}$$

Définition 20 (Influence rank)**.** *Le* ***Rang d'influence (Influence rank)*** *IR permet d'ordonner les utilisateurs influents. Les leaders sont ceux avec les valeurs de IR les plus élevées. Cette mesure est proportionnelle à l'amplitude d'influence des abonnés.*

$$IR(v_i) = (1 - \delta(v_i)\frac{\sum_{v_j \in F(v_i)} IR(v_j)}{\mid F(v_i) \mid} + \delta(v_i)MOI(v_i)) \tag{3.11}$$

Afin de calculer les valeurs *IR*, nous utilisons l'algorithme *LRA* (Recursive Algorithm). De manière récursive, il explore le voisinage du nœud pour lequel le rang d'influence est estimé.

Algorithm 3 $LRA - IR(v_i, level, max)$

```
sum = 0;
if level = max then
    return δ(v_i)MOI(v_i);
else
    for v_j ∈ N(v_i) do
        sum := sum + LRA − IR(v_j, level + 1, max);
    end for
    AvgIR := sum/|F(v_i)|;
    return 1 − δ(v_i))AvgIR + δ(v_i)MOI(v_i);
end if
```

3.3.2.4 Méthode SVM

Support Vector Machines (*SVM*) est une méthode qui 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 :

$$\begin{aligned} \min L(w) &= \frac{1}{2} * \|w\| + \mu \sum_{i=0}^{i=l} \beta_i \\ \textit{Subject to} &\begin{cases} \beta_i \geq 0 \\ h_i < w, x_i > \geq 1 - \beta_i \end{cases} \end{aligned} \tag{3.12}$$

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 :

$$SV_k(v_i, v_j) =< GC(v_i, v_j), PC(v_i, v_j), NC(v_i, v_j), CD(v_i, v_j), CI(v_i, v_j), ROI(v_i, v_j) > \tag{3.13}$$

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

3.3.3 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.

$$\begin{aligned}\log P(Y_v) = (\sum_{v_i \in V} [& \sum_{p \in P(v_i), k, l} \mu_{k,l} f_{k,l}(y_{v_i}, y_p) \\ & + \sum_{v_j \in N(v_i), k, l} \lambda_{k,l} h_{k,l}(y_{v_i}, y_{v_j})]) \\ & - \log Z\end{aligned} \tag{3.14}$$

Où :

- Les indices *k*,*l* ∈ {-1, + 1} sont 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 évalue le facteur Utilisateur-Utilisateur ;
- y_p l'étiquette du sentiment de la publication p ;
- Z un facteur de normalisation.

3.3.3.1 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.

$$f_{k,l}(y_{v_i}, y_p) = \begin{cases} \frac{\tau_{labeled}}{\mid P(v_i) \mid} & y_{v_i} = k, y_p = l, v_i : labeled \\ \frac{\tau_{unlabeled}}{\mid P(v_i) \mid} & y_{v_i} = k, y_p = l, v_i : unlabeled \\ 0 & otherwise \end{cases} \tag{3.15}$$

3.3.3.2 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 relations sociales d'amitié et d'influence.

$$h_{k,l}(y_{v_i}, y_{v_j}) = \begin{cases} \frac{\tau_{relation}}{\mid N(v_i) \mid} + \frac{\tau_{influence}}{\mid N(v_i) \mid} \times \frac{1}{1 - IR(v_j)} & y_{v_i} = k, y_{v_j} = l \\ 0 & otherwise \end{cases} \tag{3.16}$$

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

3.3.3.3 Apprentissage des paramètres

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].

Pour simplifier, nous nous référons par ϕ au vecteur des paramètres $\mu_{k,l}$ et $\lambda_{k,l}$. Nous visons à apprendre ces paramètres en maximisant $logP(Y)$ (selon ϕ). Pour ce faire, nous utilisons l'algorithme *SampleRank*.

Algorithm 4 $SampleRank(HNDIG_q, \eta)$

Initilize ϕ = (μ,λ) ;
Randomly initilize Y_v ;
for $step \in \{1, MaxSteps\}$ **do**
    $y^{new} := Sampling(Y)$;
    $\nabla := \log \frac{P(Y^{new})}{P(Y)}$;
    if $(w(Y^{new}, Y) \geq 0$ and $\nabla \leq 0)$ or $(w(Y^{new}, Y) \leq 0$ and $\nabla \geq 0)$ **then**
        $\phi := \phi - \eta \nabla_{\phi}$;
    end if
    if $Convergence$ **then**
        $break$;
    end if
    if $w(Y^{new}, Y) \geq 0$ **then**
        $Y := Y^{new}$;
    end if
end for

Où :

- La fonction "Sampling" est utilisée pour échantillonner à partir d'une distribution uniforme qui retourne la polarité d'un élément de Y_v choisi aléatoirement.
- La fonction "Initialiser" initialise les valeurs des paramètres en utilisant simplement les comptes des sous-ensembles d'utilisateurs et de publications étiquetés.

$$\mu_{k,l} = \frac{\sum_{(v_i,p) \in E_{labeled}} I(Y_{v_i} = k, Y_p = 1)}{\sum_{(v_i,p) \in E_{labeled}} (I(Y_{v_i} = k, Y_p = 1) + I(Y_{v_i} = k, Y_p = -1))} \tag{3.17}$$

et :

$$\lambda_{k,l} = \frac{\sum_{(v_i,v_j)\in E_{labeled}} I(Y_{v_i} = k, Y_{v_j} = 1)}{\sum_{(v_i,v_j)\in E_{labeled}} (I(Y_{v_i} = k, Y_{v_j} = 1) + I(Y_{v_i} = k, Y_{v_j} = -1))} \quad (3.18)$$

I(.) Est la fonction d'indicateur. Ainsi, $\mu_{k,l}$ est mis à 1 si k = l et 0 sinon. Dans nos expériences, nous définissons η à 0.001.

- La fonction de performance "w" mesure la différence de précision entre Y_{new} et Y, uniquement sur les données étiquetées. Cette fonction est détaillée dans la section (5.4).

- "Convergence" : la solution converge lorsque la fonction objectif n'augmente pas pour un nombre d'étapes donné. Nous définissons le nombre maximal d'étapes à 10000.

3.4 Conclusion

Dans ce chapitre, 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'intérêt 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.

CHAPITRE 4

Génération d'applications interactives personnalisées basée sur les communautés d'intérêt

Sommaire

Dans ce chapitre, nous présentons une méthode de personnalisation d'information contextuelle basée sur les centres d'intérêt et voisinages sociaux des utilisateurs. En outre, nous mettons en oeuvre cette méthode en proposant une approche pour générer automatiquement des applications interactives personnalisées basées sur les plateformes des Réseaux Sociaux. En effet, ces plateformes ont augmenté la participation des utilisateurs en les transformant en producteurs de contenu. Grâce aux formalismes du Web Sémantique, les contenus produits que nous désignons par les documents, les contextes de ces utilisateurs et leurs profils sont portables et exploités dans le but de la personnalisation de l'information générée. Ainsi, des fonctionnalités étendues sont nécessaires pour exploiter ces données et créer des applications sociales à contenus personnalisés basées sur les Réseaux Sociaux. Une approche automatisée d'ingénierie basée sur les modèles, utilisant les Langages Spécifiques au Domaine, est proposée pour la génération de ce que nous appelons les applications personnalisées basées sur les documents (Personalized Document-based Applications) en exploitant le pouvoir d'interaction des Réseaux Sociaux.

D'abord, nous commençons par décrire l'approche générale proposée. Ensuite, nous

détaillons chaque modèle et la manière avec laquelle il est intégré avec les Réseaux Sociaux.

4.1 Scénario de motivation

Un objectif commun pour toutes les entreprises est d'atteindre le plus d'individus possible, donc être plus reconnaissable parmi les nombreuses autres entreprises disponibles. En effet, elles doivent être connues pour fonctionner et générer des profits. C'est aussi un élément clé d'une stratégie réussie de recrutement efficace, car les gens ont tendance à s'adresser d'abord aux entreprises qu'ils reconnaissent [Greengard, 2012][Galanaki, 2002].

Plusieurs raisons expliquent la tendance récente de recrutement électronique [Natasha and Lisa, 2013]. En particulier, les Réseaux Sociaux permettent aux entreprises de gérer leurs stratégies de recrutement à moindre coût et de consacrer moins de temps au traitement de l'information rendant ainsi le processus plus efficace [Richard, 2010]. Trouver le bon employé au coût le plus bas possible contribue à une main-d'œuvre efficace et efficiente et à un fort avantage concurrentiel. Par conséquent, les Réseaux Sociaux sont actuellement établis comme un outil important pour attirer et sélectionner les candidats. Les employeurs semblent comprendre ces observations et montrent une préférence particulière pour l'utilisation de la technologie et des médias sociaux comme moyen de rechercher et de recruter de nouveaux candidats.

Supposons, à titre d'exemple, que des entreprises utilisent le Réseau Social Twitter durant leurs processus de recrutement et comme moyen d'accès à leurs systèmes d'information. Ainsi, les candidats n'ont pas besoin d'installer de nouveaux logiciels ; ils utilisent leurs comptes Twitter pour interagir avec ces entreprises.

La *figure* 4.1 montre l'intégration d'une interface utilisateur basée sur Twitter avec un système d'information d'une application de e-recrutement. Le but d'une telle application est de fournir aux utilisateurs des offres d'emploi adaptées à leurs intérêts et profils. Ces offres sont communiquées à travers des Tweets. En effet, outre la socialisation, les Réseaux Sociaux en ligne sont utilisés pour diffuser des informations, communiquer, collaborer ou coordonner des activités ciblées, de manière répartie. Ils sont donc adaptés en tant que des interfaces Front-end pour les applications interactives. Donc, les données sociales doivent être considérées comme une nouvelle dimension dans la conception de tels logiciels. Par conséquent, ces applications sociales doivent être étendues avec de nouvelles exigences et fonctionnalités permettant leur interaction avec l'infrastructure des Réseaux Sociaux afin d'atteindre plus d'utilisateurs. Le mécanisme de dialogue entre ces

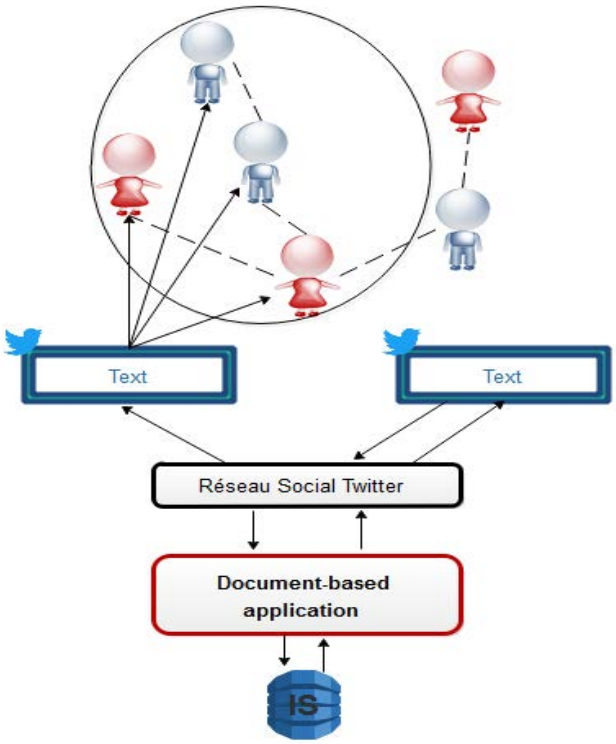


FIGURE 4.1 – Exemple de mise en oeuvre d'application sociale basée sur les Réseaux Sociaux

derniers et les applications sont les contenus partagés dans ces réseaux. Nous considérons, ces contenus comme des documents au format textuel à savoir des publications ou des messages, d'où la proposition de la notion d'applications personnalisées basées sur les documents.

Dans cette vue, nous proposons la notion d'application sociale personnalisée basée sur les documents dans les Réseaux Sociaux, que nous définissons comme suit :

Définition 21 (Personalized Document-based Application)**.** ***PDBA*** *est une application sociale interactive personnalisée dont les entrées et les sorties sont des documents extraits et produits à partir des Réseaux Sociaux en ligne.*

Dans la suite de ce chapitre, nous présentons une approche pour la conception et la génération semi-automatique de ces applications à contenus personnalisés en nous basant sur une approche de type *MDA*. L'objectif de la personnalisation de l'information est d'intégrer l'utilisateur dans le processus global d'accès à l'information en adaptant les résultats de recherche à leurs intérêts et préférences.

4.2 Solution générale

D'une part, l'adaptation des applications interactives aux préférences et intérêts des utilisateurs est un facteur clé de leur succès. L'idée est donc de conquérir les utilisateurs en leur fournissant des informations pertinentes adaptées à leurs besoins à l'aide de systèmes personnalisés. D'autre part, la méthodologie *MDA* est un paradigme important pour le développement des applications interactives. Etant donné sa capacité à accélérer le processus de développement et réduire sa complexité, elle peut être appliquée à plusieurs domaines.

Dans cette optique, nous proposons une approche automatisée basée sur les modèles pour la génération semi-automatique des applications sociales interactives à contenus personnalisés en mettant en oeuvre une méthode de personnalisation basée sur les communautés d'intérêt. Afin de répondre à cet objectif, nous avons identifié les exigences suivantes :

- La mise en œuvre de la personnalisation du contenu ;
- La prise en compte de cette personnalisation dès les premières phases de conception des applications de type *PDBA* ;
- La génération de l'application finale, d'une manière semi-automatique.

En se basant sur ces exigences, nous allons détailler les différents constituants de notre approche.

4.2.1 Méthode de contextualisation collaborative

Les applications cibles permettent de conquérir des utilisateurs en leur fournissant des systèmes personnalisés qui s'adaptent à leurs besoins et intérêts. Rappelons que la personnalisation est définie comme la capacité à fournir à tout moment à un utilisateur des contenus et des services adaptés à ses besoins et à ses attentes en utilisant des interactions Homme-Machine appropriées. Dans notre travail, nous procédons par une personnalisation de contenu. Elle consiste à sélectionner et adapter les informations contextuelles pertinentes. Le contexte s'étale sur différentes dimensions, en particulier la dimension utilisateur. Cette dernière comprend le contexte social et le contexte personnel. Le premier correspond au voisinage social de l'utilisateur comme ses groupes et communautés d'intérêt. Alors que le deuxième contient ses informations démographiques, psychologiques et cognitives. Ainsi, l'opérationnalisation de la personnalisation est fonction des centres d'intérêt de l'utilisateur en plus de son voisinage social.

Donc, pour la mise en oeuvre de la personnalisation de contenu basée sur les communautés d'intérêt, nous allons répondre aux questions suivantes :

1. Comment construire le contexte ?

2. Comment exploiter le contexte pour sélectionner les contenus ?

Les profils utilisateur sociaux construits et représentés dans le chapitre 3 sont perçus comme de la connaissance que nous exploitons pour renvoyer à l'utilisateur des contenus correspondant à ses besoins. Ce modèle contient les deux dimensions du contexte. En effet, il est construit non seulement sur la base des intérêts et activités des utilisateurs mais aussi en tenant compte des autres utilisateurs qui leurs sont similaires. Plus particulièrement, et afin de bénéficier des données des autres pour une meilleure personnalisation des contenus, certaines recherches ont exploré l'utilité de l'exploitation des informations d'appartenance à un groupe qui pourraient être suffisamment similaires les uns aux autres. Cette similarité est bien garantie dans les communautés d'intérêt comme étant des groupes d'utilisateurs ayants des intérêts communs. Ainsi, nous proposons d'exploiter les données dans les communautés d'intérêt en tirant parti des points communs de ses membres pour réaliser la personnalisation.

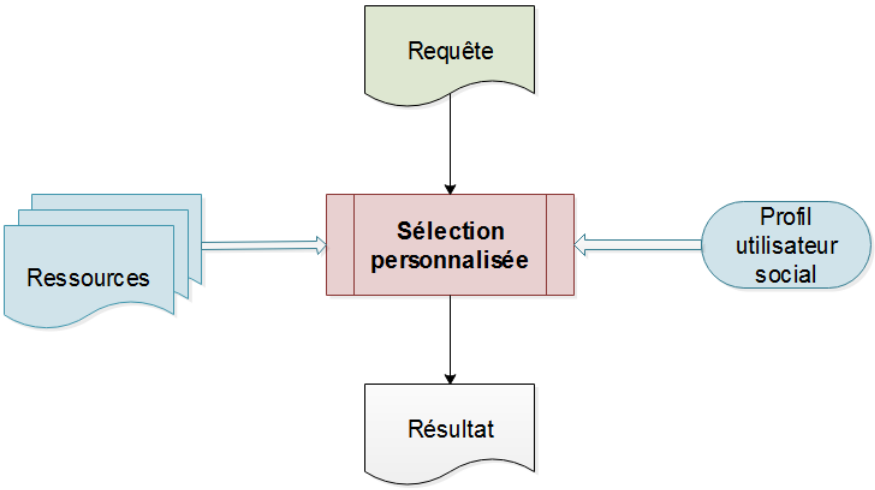


FIGURE 4.2 – Intégration du profil utilisateur social dans la personnalisation des informations dans les *PDBA*

A la différence de la personnalisation individuelle où l'utilisateur est considéré individuellement, la personnalisation collaborative considère l'utilisateur comme membre d'un groupe d'utilisateurs. Nous proposons ce que nous appelons une personnalisation

basée sur les communautés (community-based personalization) où un utilisateur n'est pas considéré individuellement mais comme membre d'une communauté d'intérêt.

Définition 22 (Community-based Personalization)**.** *La personnalisation basée sur les communautés est une méthode de personnalisation de contenu collaborative dont le principe est de considérer l'utilisateur comme membre d'une communauté d'intérêt afin de lui fournir des informations adaptées à ses intérêts.*

Pour ce faire, nous nous basons sur la technique de personnalisation appelée "**Groupization**" [Morris et al., 2008]. Il a été prouvé que cette technique augmente le processus de personnalisation en donnant des poids plus élevés aux résultats qui intéressent davantage de membres d'un groupe d'utilisateurs.

Cette technique est un moyen de reclasser les résultats dans l'ordre le plus pertinent pour les membres du groupe. Elle vise à amplifier les similitudes entre ces membres afin de produire une vue partagée des résultats correctement ordonnée.

Cette technique tire parti des points communs d'un groupe pour améliorer la personnalisation. En effet, elle considère les informations sur l'appartenance à un groupe en tant qu'indicateur supplémentaire implicite des utilisateurs qui pourraient être assez semblables les uns aux autres. Le processus de la personnalisation est augmenté en attribuant des poids plus élevés aux documents qui intéressent davantage de membres du groupe, en fonction de l'appariement de l'historique de chaque membre et des fréquences locales des documents. D'abord un score de personnalisation est calculé pour chaque résultat de recherche pour chaque membre du groupe.

Les étapes de l'application de la technique de "Groupization" sur un ensemble de résultats de recherche sont les suivantes :

1. Calculer un score de personnalisation pour chaque résultat de recherche pour chaque membre du groupe ;

2. Pour chaque ensemble de résultats, le score est calculé comme la somme des scores de personnalisation de chaque membre du groupe.

Quant au calcul du score de personnalisation d'un résultat de recherche, nous avons recours à la méthode définie par [Teevan et al., 2005]. Cette méthode incorpore les informations sur les intérêts et les activités des utilisateurs dans la personnalisation de la recherche en modifiant *BM25* [Jones et al., 1998] qui est un processus de pondération probabiliste. Il classe les documents en fonction de leur probabilité de pertinence selon la requête. Plus particulièrement, il les classe en additionnant les termes d'intérêt du produit du poids w_i du terme i et de la fréquence avec laquelle ce terme apparaît dans le

document.

$$w_i = \log \frac{(r_i + 0.5)(N - n_i + 0.5)}{(n_i + 0.5)(R - r_i + 0.5)} \tag{4.1}$$

Où :

- N est le nombre total de documents
- n_i est le nombre de documents qui contiennent le terme i
- R est le nombre des documents correspondant à l'utilisateur en question
- r_i est le nombre de documents correspondant à l'utilisateur en question contenant le terme i

Dans leur travail, [Teevan et al., 2005] considèrent les documents d'un utilisateur comme les pages Web, les messages électroniques (Email), les éléments du calendrier, etc. Dans notre travail, nous considérons les documents comme étant les contenus que l'utilisateur publie et partage dans les Réseaux Sociaux.

4.2.2 Pile d'abstraction pour la génération des applications *PDBA*

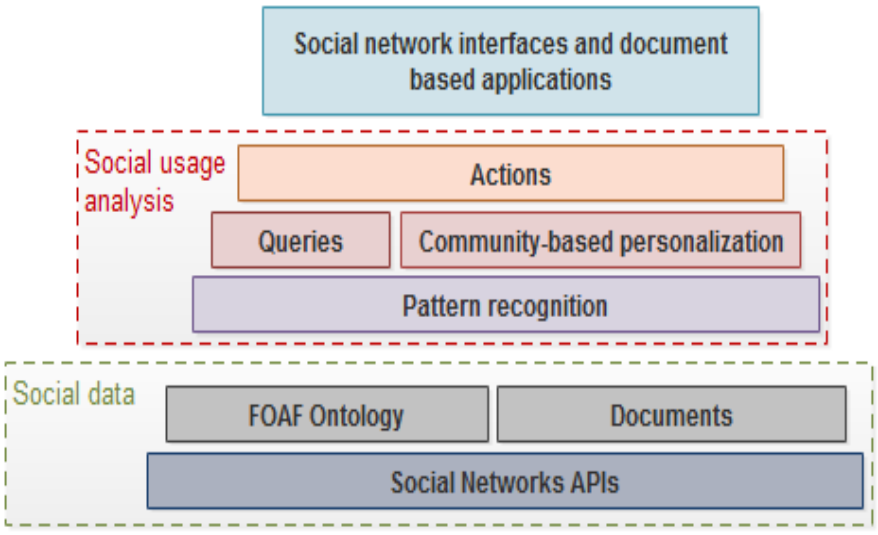


FIGURE 4.3 – Pile d'abstraction pour la génération des applications interactives basées sur les Réseaux Sociaux

La *figure* 4.3 présente la pile d'outils que nous avons conçu pour effectuer une génération d'applications interactives personnalisées basée sur les Réseaux Sociaux. L'objectif

de cette pile est de fournir un framework permettant d'intégrer la méthode de personnalisation basée sur les communautés d'intérêt dans la conception et par la suite la génération des applications interactives *PDBA*. Cette pile est composée de :

1. Outils de représentation des utilisateurs et d'extraction des documents destinés à l'application à travers les plateformes des Réseaux Sociaux ;

2. Outils de requêtage et de personnalisation basée sur les communautés d'intérêt des utilisateurs et l'exécution des actions sur les documents.

Nous représentons les Réseaux Sociaux par l'ontologie *FOAF*, qui fournit la structure du réseau. Nous avons montré dans le chapitre 3 que cette ontologie est aussi adéquate pour représenter les profils utilisateurs sociaux, surtout que ce format facilite le partage et l'interopérabilité des données sociales entre diverses applications. Nous utilisons *FOAF* pour décrire les utilisateurs, leurs relations ainsi que leurs contenus. Les données sociales sont donc disponibles dans un format sémantique et peuvent être exploitées directement ou interrogées en utilisant le langage *SPARQL*. La plupart de ces données sont accessibles à partir des interfaces de programmation correspondantes aux briques de fonctionnalités fournies par les plateformes des Réseaux Sociaux. Cependant, nous avons intérêt à analyser les données disponibles afin d'en extraire seulement les informations pertinentes (Pattern Recognition). Pour ce faire, nous utilisons le modèle proposé dans le chapitre 3. En utilisant les documents pertinents et les requêtes des utilisateurs destinées à l'application, nous appliquons la méthode de personnalisation basée sur les communautés pour sélectionner les résultats pertinents adaptés aux intérêts des utilisateurs. La composition des résultats est effectuée en appliquant des actions sur les documents permettant aussi de les renvoyer aux utilisateurs à travers des interfaces des Réseaux Sociaux.

4.3 Approche MDA proposée

4.3.1 Principe et méthodologie de développement

L'Ingénierie Dirigée par les Modèles (***IDM***) est un paradigme qui réfère à l'utilisation des modèles comme éléments principaux dans le cycle de vie du logiciel. Se basant sur une abstraction des problèmes réels et complexes, ce paradigme permet de réduire le temps de développement des systèmes. Ces derniers sont donc plus faciles à spécifier et maintenir. La Méthodologie Dirigée par les Modèles (***MDA***) est un standard basé sur ***IDM***, lancé par l'***OMG***. Elle permet l'abstraction des systèmes complexes à travers l'élaboration de standards pour la définition des modèles ainsi que leurs relations et transformations.

MDA vise à modéliser les spécifications fonctionnelles d'une application, indépendamment des spécifications techniques liées à la plateforme d'exécution. Utilisant des standards de l'*OMG*, ces spécifications sont définies initialement dans un modèle indépendant de l'informatisation (*CIM*). Ce dernier est utilisé pour créer un modèle indépendant de toute plateforme d'exécution (*PIM*), qui avec un modèle de description de plateforme permet la génération (semi-) automatique par transformation des modèles. Ainsi, *MDA* tient compte de tous les aspects de développement logiciel grâce à la modélisation. Le formalisme de modélisation joue donc un rôle important dans la pérennité des savoir-faire métiers. Il existe quatre niveaux de modélisation, décrits sous une forme pyramidale, qui sont :

- **Niveau M0** : le système réel à modéliser ;
- **Niveau M1** : le modèle qui représente une abstraction du monde réel ;
- **Niveau M2** : le méta-modèle est un langage utilisé pour décrire les modèles ;
- **Niveau M3** : le méta-méta-modèle permet de décrire les langages des méta-modèles et se décrit lui-même.

Les modèles véhiculent les informations nécessaires pour la génération du code source de l'application à travers des transformations successives des modèles. Dans sa majorité, le processus de transformation est automatisé. En particulier, les *DSL*s (**D**omain **S**pecific **L**anguage) rendent l'ensemble du processus et de la génération plus souple et rapide. Ils sont des langages à partir desquels sont décrites les spécifications fonctionnelles de l'application à générer. Ces descriptions sont désormais indépendantes des spécifications techniques liées aux plateformes d'exécution à différents niveaux d'abstraction grâce à un vocabulaire personnalisé et compréhensible aussi bien par les experts du domaine que par les développeurs.. Personnalisables, les *DSL*s sont définis par des vocabulaires et règles syntaxiques déterminés par le concepteur. Donc, les modèles sont des énoncés de ces langages. Ils respectent cette syntaxe et décrivent les fonctionnalités de l'application. En effet, ils permettent la génération du code source depuis les *PIM*. De plus, le générateur de code décrit automatiquement l'architecture de la plateforme d'exécution. Des templates sont aussi créées par ce générateur pour l'expression de l'architecture de la plateforme cible et contiennent les règles de transformation d'un modèle en code source.

La première étape du processus de définition d'un *DSL* consiste à établir son vocabulaire. Ce dernier reflète le domaine métier de l'application et doit être simple et compréhensible. Ensuite, il faut définir le méta-modèle et les règles syntaxiques du *DSL*. Ces règles se basent sur le vocabulaire défini lors de la première étape. Dans notre travail,

nous avons recours aux *DSL* sous forme graphique. Nous allons utiliser le langage de modélisation unifiée *UML* pour la conception des méta-modèles de ces *DSL*. Le type graphique est plus visuel et efficace pour représenter les relations entre éléments. Ces éléments dépendent du modèle d'application et des besoins fonctionnels. En effet, ce modèle comprend les types de données, et les concepts contenant des propriétés permettant d'exprimer au mieux les fonctions de l'application.

Pour le développement des *DSL*s, nous avons suivi la méthodologie proposée dans [Mernik et al., 2005]. Cette méthodologie comprend quatre phases, à savoir : la décision, l'analyse, la conception et la mise en œuvre.

1. **Décision** : Faire le choix d'un langage dédié à un domaine particulier permet d'avoir une liberté dans la description du langage qui assure une parfaite adéquation au domaine. De plus, il permet le développement de logiciels par des utilisateurs ayant moins d'expertise en matière de programmation.

2. **Analyse** : Dans cette phase, nous identifions le domaine du problème et rassemblons les connaissances correspondantes. Les sources de ce domaine représentent les entrées. Fréquemment, son analyse est faite de manière informelle où la capture et la représentation des connaissances sont effectuées. Dans notre cas, nous sommes amenés à analyser le domaine des Réseaux Sociaux afin d'extraire les informations qui servent plus tard dans la personnalisation des documents destinés à l'application. A partir de ces ressources, les profils utilisateur, leurs connexions dans le réseau ainsi que les documents doivent être extraits. Dans notre travail, comme mentionné précédemment, nous utilisons les formalismes du web sémantique pour représenter les données. Ainsi, les données extraites servent d'une part à identifier les communautés d'intérêts des utilisateurs et leurs contextes et d'autre part à identifier les documents pertinents parmi ceux envoyés par les utilisateurs.

3. **Conception** : La méthodologie *MDA* repose sur des modèles, et la méta-modélisation joue un rôle important. Elle est considérée comme une technique courante pour définir la syntaxe abstraite des modèles. Dans cette phase, nous concevons les méta-modèles liés aux *DSL*s proposés en utilisant des techniques de conception *UML*.

4. **Implémentation** : Pour la mise en œuvre des méta-modèles proposés, nous utilisons le langage Kermeta. Ce dernier implémente une extension de l'*EMOF* (*Essential MOF*) 2.0 qui permet de spécifier un comportement impératif aux entités du méta-modèle. Le prototypage et la transformation de modèles est donc possible directement dans cet outil en générant le code correspondant au méta-modèle.

4.3.2 Architecture générale

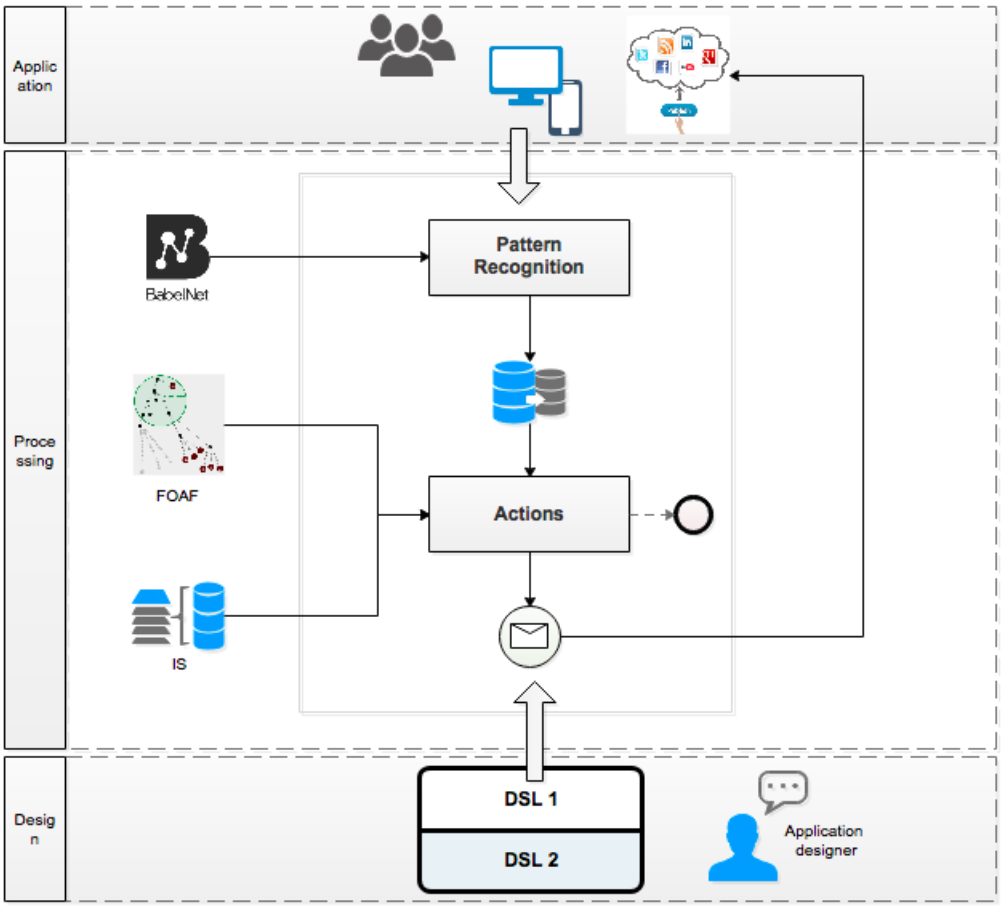


FIGURE 4.4 – Architecture générale

Dans cette section, nous présentons l'architecture de l'approche proposée pour la génération semi-automatique des applications personnalisées basées sur des documents. Cette architecture est organisée en trois couches (*figure* 4.4) : une couche "application" représente l'interface à partir de laquelle l'utilisateur interagit avec l'application et une couche "traitement" exécute les traitements définis par les *DSL*s qui sont construits dans la troisième couche de "conception".

1. En premier lieu, les utilisateurs adressent des requêtes sous forme de documents (publications ou messages privés) à l'application cible via le Réseau Social ;

2. Seulement les documents pertinents mentionnant l'application cible sont pris en considération ;

3. Les patterns qui devraient être trouvés dans les documents pour décider leur pertinence ou pas sont définis par le concepteur de l'application en utilisant le premier *DSL* approprié qui communique avec un dictionnaire externe "BabalNet" ;

4. Les documents pertinents extraits à l'aide du premier *DSL* sont analysés en appliquant différentes actions comme la sélection de concepts, l'agrégation et le calcul des fréquences des termes ;

5. Les actions de traitement sont définies par le concepteur de l'application en utilisant le deuxième *DSL* ;

6. L'exécution de ces actions nécessite éventuellement une communication avec le Système d'Information de l'application pour l'envoi et la réception des informations qui sont utilisées dans la synthèse des réponses adaptées ;

7. Les données extraites suite à l'application des actions ou fournies par le Système d'Information sont exploitées afin d'identifier les documents adéquats aux requêtes des utilisateurs qui sont adaptés à leurs intérêts. La technique de personnalisation est définie au niveau du deuxième *DSL* dont le but est d'appliquer la méthode de personnalisation basée sur les communautés d'intérêt ;

8. La personnalisation utilise des données qui sont stockées dans l'ontologie *FOAF* ;

9. Les données extraites lors de l'exécution de la personnalisation sont utilisées pour synthétiser des réponses personnalisées ;

10. Les réponses sont envoyées automatiquement aux utilisateurs cibles sous forme de documents à travers les Réseaux Sociaux.

Afin de faciliter la construction des applications interactives sociales, la solution proposée dirigée par les modèles est basée sur les *DSL*s. Dans la suite, nous détaillons les *DSL*s proposés.

4.3.2.1 DSL 1 : Reconnaissance des patterns

Afin de manipuler les documents pour extraire les patterns, nous avons construit un *DSL* dont le méta-modèle est présenté dans la *figure* 4.5. Ce *DSL* 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

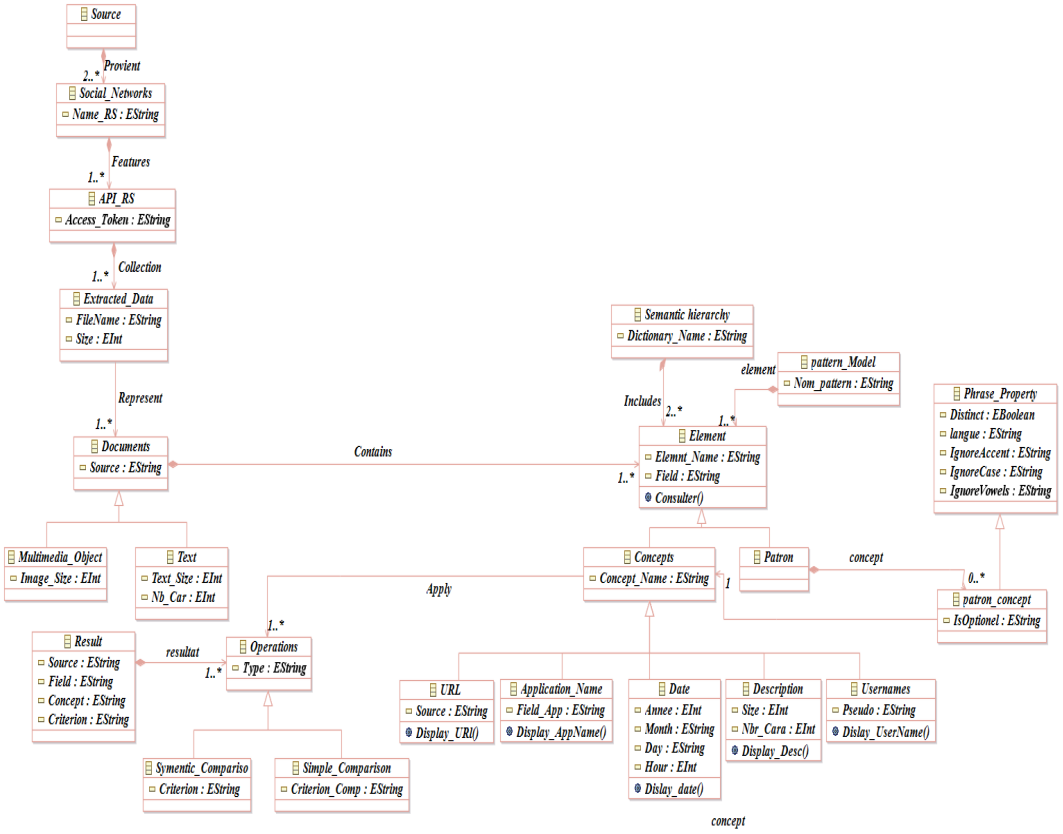


FIGURE 4.5 – Extrait du méta-modèle du *DSL* 1

spécifiques qui sont souvent utilisés dans les Réseaux Sociaux comme les *URL*, les noms des utilisateurs et les hashtags. Ce modèle est lié à un dictionnaire externe pour pouvoir rechercher les relations des hiérarchies sémantiques des concepts. En effet, ils sont souvent liés par des relations sémantiques (synonymie, antonymie, etc.). Les patterns peuvent être définis par le concepteur ou extraits directement du dictionnaire "BabalNet". La classe "Concept" hérite de la classe "Result". Ceci permet d'envoyer les concepts identifiés dans les publications en tant que données externes afin de pouvoir les référencer dans les requêtes. En effet, les concepts sont séparés de leur utilisation dans "*Pattern*" afin de pouvoir les exploiter dans différents contextes. La classe "PatterConcept" est utilisée pour la configuration des concepts, comme par exemple, spécifier un concept comme facultatif dans le modèle. Cependant, les informations de méta-données présentes dans les publications comme la date ou la géolocalisation peuvent être récupérées donc n'ont pas besoin d'être explicitement déclarées dans le modèle.

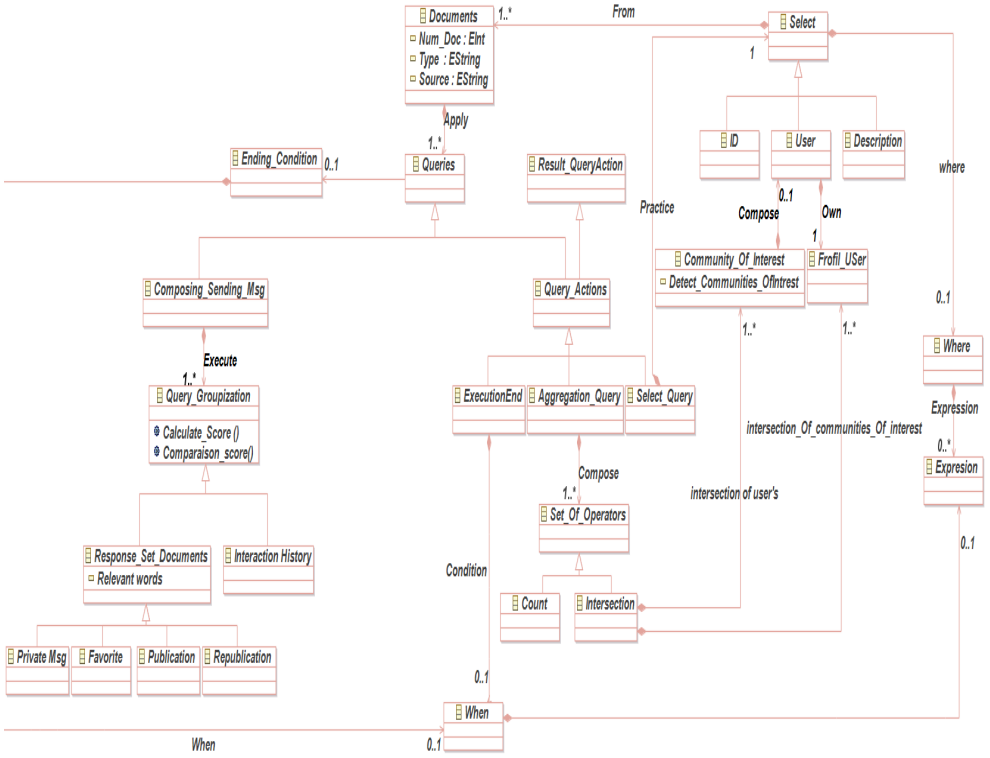


FIGURE 4.6 – Extrait du méta-modèle du DSL 2

4.3.2.2 DSL 2 : Exécution des actions et personnalisation

Un deuxième *DSL* pour la description des actions qui peuvent être effectuées sur les documents est montré dans la figure 4.6. Ces actions sont utilisées pour exécuter des requêtes et sélectionner des concepts à partir des documents pertinents obtenus lors de l'exécution du premier *DSL*. Les requêtes ont une syntaxe similaire aux requêtes *SQL*. Elles comprennent la sélection de concepts qui remplissent certaines conditions spécifiées. De plus, elles permettent d'obtenir certaines métadonnées contenues dans les documents telles que la position géographique. Le *DSL* permet également la communication avec des systèmes d'information externes. Ceci est pris en compte en permettant la définition de dépendances de données externes depuis ces systèmes. Ainsi, les données peuvent être introduites dans ces derniers ou en extraites. Des évènements asynchrones déclenchés par la source externe permettent de fournir des données au modèle.

Une fois les données sont disponibles à partir des requêtes, des documents de réponse sont composés et envoyés aux utilisateurs correspondants. Ceci est reflété par la classe "Action". Ces documents peuvent être sous forme de "messages" publics ou privés. Des informations sur les communautés d'intérêt des utilisateurs sont représentées par la classe "Community_Of_Interest". Elles sont extraites grâce à la communication avec l'ontologie

FOAF. Elles sont exploitées pour l'exécution de la méthode de personnalisation qui repose sur la technique de "Groupization" (classe "Query_Groupization"). Ainsi, des documents à contenus personnalisés sont envoyés aux utilisateurs. Une publication de réponse est envoyée lorqu'un déclencheur (classe "When") prend la valeur "vrai".

La fin de l'exécution de l'application est signalée (classe "EndingCondition") suivant une condition qui peut dépendre de plusieurs facteurs. Ces derniers peuvent être le nombre de publications reçues dans un certain délai de temps.

4.4 De la modélisation à la génération du code

Dans la section précédente, nous avons présenté l'architecture *MDA* de notre approche ainsi que les modèles sous forme de *DSL*s qui supportent ses différents niveaux d'abstraction. Afin de montrer sa mise en oeuvre fonctionnelle, nous abordons dans ce qui suit l'instrumentation des outils qui conviennent le mieux à nos besoins.

4.4.1 Atelier logiciel : l'outil EMF

Intégré au sein de la plateforme Eclipse, *EMF*[19] est un environnement qui permet la modélisation, la méta-modélisation et la génération de code. Dans cet environnement, les modèles peuvent être définis de différentes manières en utilisant les notations *UML*, *XSD* (XML Schema Definition), le *Java* annoté ou bien *Ecore*. Ce dernier est la propre notation de *EMF* définie comme étant un langage canonique pour la description des modèles dont le format de persistance est *XMI* (XML Metadata Interchange). Ce format est utilisé pour permettre l'interopérabilité dans l'échange de modèles entre différents formalismes et outils de modélisation. Un élément appelé "Ressource" est un conteneur pour des instances du modèle. Il s'agit de l'unité de base pour la persistance.

Quelque soit la notation retenue pour les définir, des modèles *Ecore* sont générés. Un *DSL* peut être défini en utilisant Ecore.

Etant facile et intuitive à l'utilisation pour la création et la gestion des modèles, cette plateforme est la plus utilisée dans le domaine de l'*IDM*. Elle fournit la possibilité de non seulement valider les modèles créés mais aussi de les enrichir à l'aide des expressions *OCL* (Object Constraint Language). En outre l'un des principaux avantages de *EMF* est la génération automatique de code. En effet, en utilisant son assistant il est possible d'obtenir une implémentation java à partir d'un modèle *Ecore*.

C'est cet atout que nous avons exploité dans notre travail pour générer une partie du code source des applications sociales interactives basées sur les documents suite à la

19. eclipse.org/modeling/emf

transformation des modèles proposés sous forme de *DSL*s.

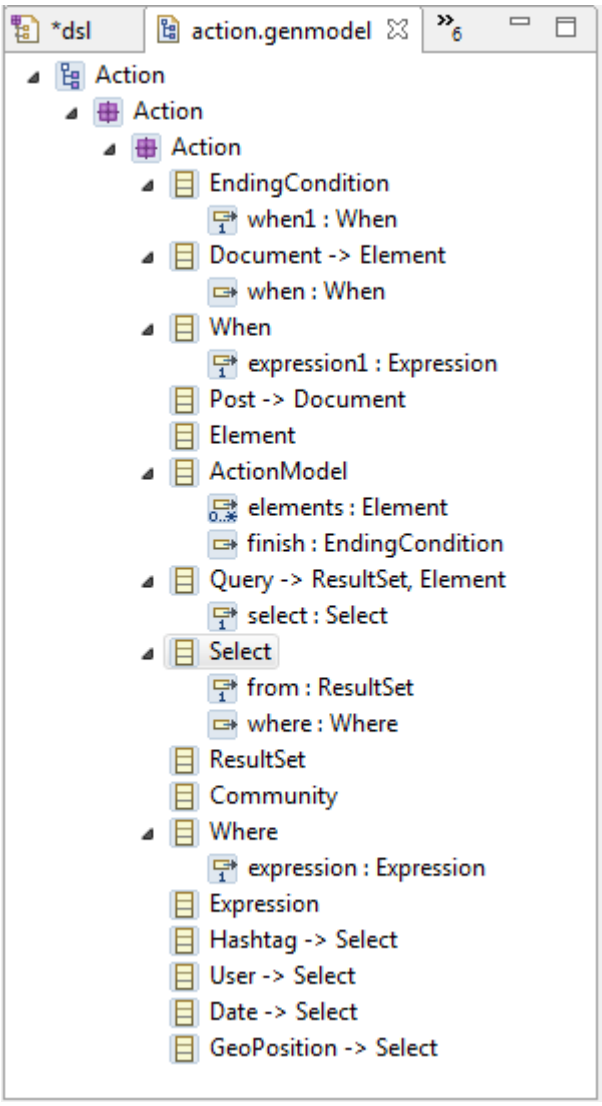


FIGURE 4.7 – Modèle de génération

4.4.2 Mise en oeuvre des modèles *DSL*s

Plusieurs formats sont disponibles pour construire un modèle *EMF*. Dans notre travail, nous utilisons le modèle *Ecore*. Les étapes de modélisation pour la mise en oeuvre des *DSL*s sont les suivantes :

- Création du modèle *EMF* (extension .ecore) : création des classes, des attributs et des associations entre les classes. Ils existent plusieurs représentations pour la visualisation d'un modèle *EMF* : vue de type diagramme, vue de type arbre ou vue de type texte.

- Création du modèle de génération (extension .genmodel) : à partir du modèle défini dans l'étape précédente, il est possible de générer du code Java dédié à la création des instances de ce modèle. La création d'un modèle de génération (appelé *genmodel*) est nécessaire pour la génération du code (voir *figure* 4.7). Il contient des informations sur la génération comme le chemin de génération, package, etc. ; qui ne sont pas intégrées au modèle.

- Paramétrage du modèle de génération : dans cette étape, il convient de préciser quelques paramètres comme le package de génération ($Base_Package) et le modèle *Ecore* ($Package).

- Génération du code *Java* et de l'éditeur graphique : à l'aide de l'utilité "*Generate Model Code*", un ensemble de classes "Java" sont générées automatiquement, comme le montre la *figure* 4.8, ainsi que leurs implémentations. Il est également possible d'apporter des modifications sur le code généré comme l'implémentation du corps des opérations et l'ajout de nouveaux attributs. En outre, l'utilité "*Generate Edit and Editor Code*" permet de générer automatiquement un éditeur pour construire les instances des modèles. Ainsi, les instances des classes peuvent être construites automatiquement via l'éditeur.

```
/**
package eclipse.emf.action.model.action.Action;

import org.eclipse.emf.ecore.EObject;

/**
 * <!-- begin-user-doc -->
 * A representation of the model object '<em><b>Where</b></em>'.
 * <!-- end-user-doc -->
 *
 * <p>
 * The following features are supported:
 * <ul>
 *   <li>{@link eclipse.emf.action.model.action.Action.Where#getExpression <em>Expression</em>}</li>
 * </ul>
 * </p>
 *
 * @see eclipse.emf.action.model.action.Action.ActionPackage#getWhere()
 * @model
 * @generated
 */
public interface Where extends EObject {
     * Returns the value of the '<em><b>Expression</b></em>' containment reference.
    Expression getExpression();

    /**
     * Sets the value of the '{@link eclipse.emf.action.model.action.Action.Where#getExpression <em>Expression</
     * <!-- begin-user-doc -->
     * <!-- end-user-doc -->
     * @param value the new value of the '<em>Expression</em>' containment reference.
     * @see #getExpression()
     * @generated
     */
    void setExpression(Expression value);

} // Where
```

FIGURE 4.8 – Exemple de code *Java* généré automatiquement

- Création des instances : les instances du modèle peuvent être construites en utilisant directement le code généré via l'éditeur (voir *figure* 4.9). La modification de ces instances peut se faire via le méta-modèle sans génération de code.

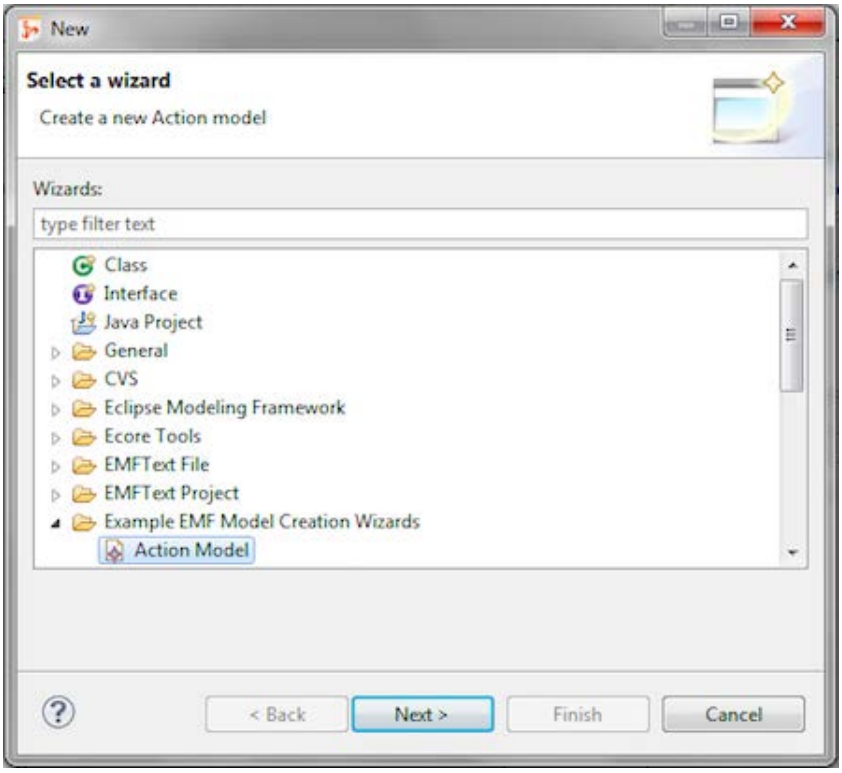


FIGURE 4.9 – Plugin *Eclipse* généré pour la création des instances

4.5 Conclusion

Avec la large utilisation des Réseaux Sociaux, les développeurs utilisent de plus en plus ces réseaux pour développer des applications sociales. Ils utilisent les plateformes de ces réseaux comme front-end pour leurs applications. Afin de les rendre plus attractives, il s'avère important d'adapter ces applications aux intérêts des utilisateurs. Pour pallier ces besoins, nous avons proposé une approche de type *MDA* pour la génération semi-automatique d'applications sociales à contenus personnalisés et basée sur les documents générés dans les Réseaux Sociaux. L'exploitation des communautés d'intérêt détectées dans ces réseaux est le coeur de cette approche. En effet, nous avons proposé une méthode de personnalisation de contenu basée sur les communautés. Cette méthode est prise en compte depuis la phase de conception. L'architecture *MDA* proposée repose sur la définition de deux *DSL*s dont un qui exécute la personnalisation et l'autre qui identifie les

documents pertinents. Tout de même, nous avons abordé les développements réalisés dans le cadre de ce travail, permettant de générer du code automatiquement destiné à utilisation par les développeurs tiers.

CHAPITRE 5

Evaluation sur les Réseaux Sociaux numériques : cas de Twitter et Facebook

Sommaire

Aux deux chapitres précédents, nous avons présenté en détail deux nouvelles approches pour *(i)* la détection des communautés d'intérêt dans les Réseaux Sociaux (chapitre 3) et *(ii)* la génération semi-automatique d'applications interactives à contenus personnalisés basée sur ces communautés (chapitre 4).

En premier lieu, l'évaluation vise à démontrer la pertinence de nos hypothèses de travail qui stipulent que les similarités des intérêts et des sentiments au niveau des utilisateurs sont corrélés au phénomène d'influence sociale en plus de l'homophilie. Ensuite, nous évaluons de manière empirique la performance des modèles et approches proposées de prédiction et de détection des communautés.

5.1 Méthodologie de construction du profil utilisateur social

Afin de bien mener les expérimentations, nous construisons tout d'abord un jeu de données de test. L'acquisition des données sur la structure des Réseaux Sociaux ainsi que les activités des utilisateurs constitue un premier enjeu majeur de notre étude expérimentale. Encore faut-il analyser ces données et les structurer ; c'est à cela que sert le modèle de profil utilisateur qui est un instrument de synthèse de connaissances. Dans le chapitre 3, nous avons défini un modèle de profil utilisateur social. En général, un modèle est un regroupement de connaissances qui concernent quelques faits expérimentaux bien délimités. Notre modèle représente les utilisateurs et l'ensemble des données sociales.

5.1.1 Accès aux données sociales

Dans notre cas particulier, la question qui se pose consiste à analyser l'accessibilité des données de l'utilisateur et de son réseau égocentrique afin d'évaluer les algorithmes que nous avons proposés. Comme présenté au chapitre 1, les plateformes des Réseaux Sociaux fournissent, généralement, des *API* à des développeurs tiers. L'accès aux données nécessite la création d'application dans le site du Réseau Social concerné. Ainsi, les utilisateurs accordent des autorisations appropriées à l'application afin de collecter leurs données. Cette technique permet à l'application de ne pas se limiter aux données publiques mais d'accéder à beaucoup plus de données qui sont privées.
Cependant, l'utilisation des *API* a deux inconvénients majeurs : (i) les fournisseurs de services du réseau local restreignent souvent le nombre d'appels d'*API* qu'une application peut effectuer pendant un certain intervalle de temps ; et (ii) les fonctionnalités fournies varient considérablement d'une *API* à l'autre, donc il faut apprendre à gérer chaque *API* à part.
Parmi les premiers sites de Réseaux Sociaux à proposer des *API* de développement, nous retrouvons le réseau de *Facebook*. Cependant, plusieurs remaniements ont été advenus entre la première version de son *API* en 2006 et la version actuelle. Ces changements concernent en partie les données accessibles par les applications tierces ainsi que la manière d'y accéder. En effet, le Règlement Général sur la Protection des Données (*RGPD*), depuis 2018, a mis en oeuvre de nouvelles règles de protection des données. Ces règles s'appliquent au traitement des données personnelles des individus y compris la manière dont les données sont manipulées. Subséquemment, la politique d'accès aux données définie par *Facebook* a évolué. Des permissions basées sur le protocole *OAuth* ont été mises

en oeuvre permettant aux utilisateurs un contrôle de l'accès à leurs données par des applications tierces. Toutefois, certaines données restent accessibles et d'autres nécessitent l'accord des utilisateurs.
De même pour Twitter, il faut enregistrer des applications pour pouvoir accéder à l'*API*. Ces applications, par défaut, ne peuvent accéder qu'aux données publiques. Les Tweets et réponses publics sont mis à disposition aux développeurs. L'accès aux Tweets se fait en recherchant des mots-clés donnés, ou en demandant un échantillon de Tweets de certains comptes utilisateurs spécifiques.

5.1.2 Méthodologie de la construction des sujets d'intérêt

Nous présentons ici la méthodologie de construction des sujets d'intérêt de la dimension personnelle du profil utilisateur social. Le but est de présenter comment les données issues des contenus partagés par les utilisateurs dans les Réseaux Sociaux de Twitter et Facebook sont exploitées pour dériver le profil utilisateur. Cette méthodologie se décompose en cinq étapes présentées sur la *figure* 5.1.

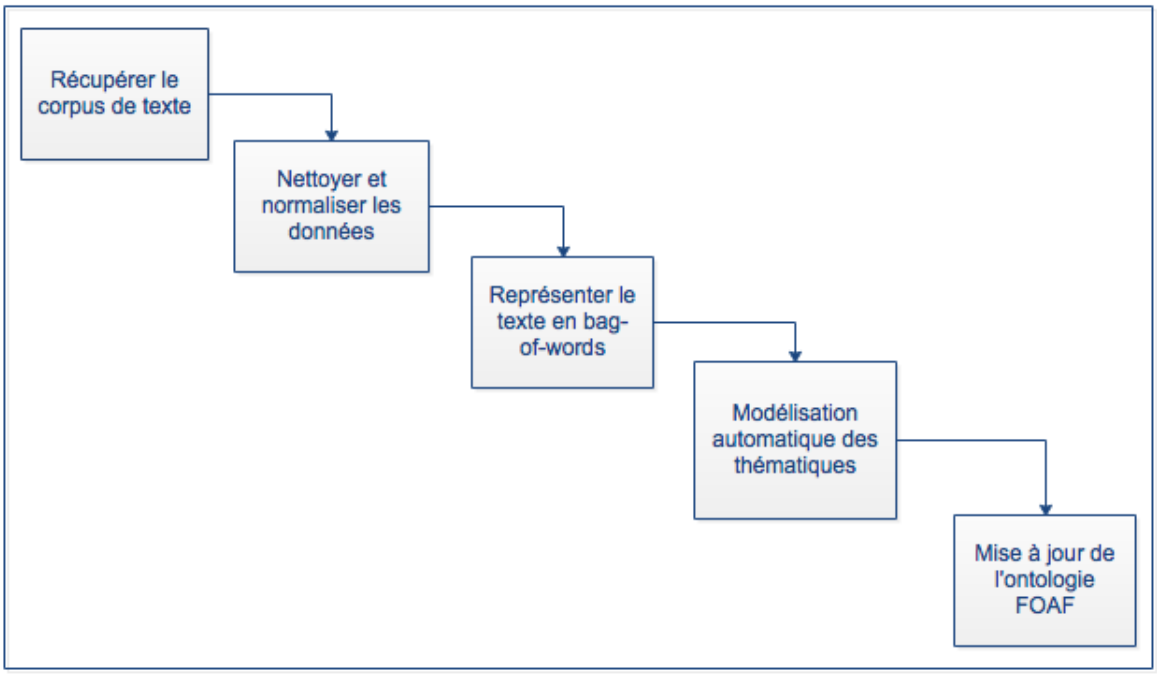


FIGURE 5.1 – Méthodologie de la construction des sujets d'intérêt

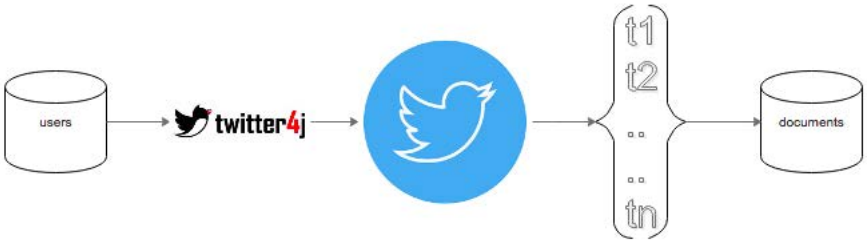


FIGURE 5.2 – Extraction des données en utilisant la librairie *Twitter4j* pour l'*API* de *Twitter*

5.1.2.1 Récupérer le corpus de texte

Dans ce qui suit, nous présentons le cas de Twitter. L'extraction se fait par le biais de la librairie *Twitter4j*[20] qui est une librairie *Java* facilitant l'utilisation de l'*API* de Twitter comme le montre la *figure* 5.2. L'ensemble des textes collectés des Réseaux Sociaux est appelé corpus. Un nombre prédéfini de Tweets est récupéré en exécutant le code *Java* présenté dans la *figure* 5.3 ; soit par défaut 100 Tweets pour chaque utilisateur. Chaque Tweet est représenté comme un document.

```java
public class GetUserTimeLine {
    public static void main(String[] args) {
        // gets Twitter instance with default credentials
        Twitter twitter = new TwitterFactory().getInstance();
        try {
            List<Status> statuses;
            String user;
            int pageno = 1;
            if (args.length == 1) {
                user = args[0];
                Paging page = new Paging(pageno++, 100);
                statuses = twitter.getUserTimeline(user,page);
            } else {
                user = twitter.verifyCredentials().getScreenName();
                statuses = twitter.getUserTimeline();
            }
            System.out.println("Showing @" + user + "'s user timeline.");
            for (Status status : statuses) {
                System.out.println("@" + status.getUser().getScreenName() + " - " + status.getText());
            }
        } catch (TwitterException te) {
            te.printStackTrace();
            System.out.println("Failed to get timeline: " + te.getMessage());
            System.exit(-1);
        }
    }
}
```

FIGURE 5.3 – Extrait du code Java d'extraction des Tweets d'un utilisateur

20. http ://twitter4j.org/en/

Le prétraitement du corpus est sans doute une étape importante. Elle est appliquée systématiquement avant toute tâche de fouille de données. En plus de la préparation des données pour leur analyse, le prétraitement des données consiste à y apporter des transformations. Une fois le texte séparé en unités de mots (tokens), nous comptons la fréquence d'apparition des différents mots pour avoir une idée du champ lexical. La *figure* 5.4 présente un exemple d'analyse fréquentielle. En effet, l'observation du contenu du corpus obtenu après transformation permet de s'assurer que les données obtenues correspondent à celles désirées. Effectivement, leur qualité est déterminante pour l'exécution des traitements ultérieurs.

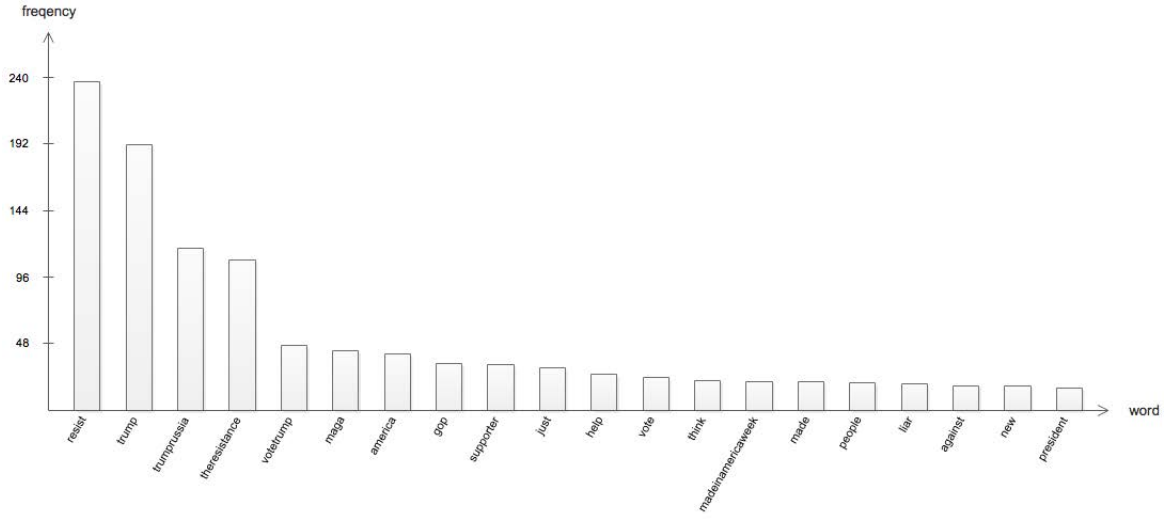


FIGURE 5.4 – Exemple des fréquences des mots dans un extrait du corpus

5.1.2.2 Nettoyer et normaliser les données

Après la tokenization, nous procédons au nettoyage et à la normalisation du corpus afin d'obtenir un dictionnaire représentatif de l'ensemble des documents. Les étapes de nettoyage sont les suivantes :

- Supprimer les mots vides (stopwords) qui sont les mots courants et qui ne représentent pas une valeur informative pour la sémantique des données du corpus ni du pouvoir discriminatif ;
- Supprimer les hashtags pas dans leur ensemble mais seul le caractère "#" est supprimé vu que le reste est souvent une chaîne qui contribue à une meilleure compréhension du Tweet ;

- Supprimer les préfixes, suffixes et autres des mots afin de les représenter sous leur forme canonique : il s'agit du processus de Racinisation (ou stemming en anglais) qui est un traitement purement algorithmique visant à réduire les mots à leurs racines afin de les traiter comme une seule entité.

5.1.2.3 Représenter le corpus en "bag of words"

Dans un modèle vectoriel *VSM* (*Vector Space Model*), chaque document est représenté par un vecteur de valeurs numériques ou catégoriques. Donc, un document est représenté par un ensemble des mots qu'il contient, sans soucis de contexte. Généralement, dans une représentation bag of words (sac de termes) chaque document est représenté par un vecteur de la taille du vocabulaire. L'hypothèse de cette représentation est l'indépendance des termes d'indexation.

5.1.2.4 Modélisation automatique des thématiques

Dans le contexte des Réseaux Sociaux, l'information utile est difficile à extraire étant enfouie dans des énormes volumes de données. Dans notre travail, nous avons affaire à des données textuelles. Pour une exploration efficace de ces données, nous nous reposons sur des outils de fouille de données. Une première tâche de la fouille de textes est la modélisation des thématiques. Elle permet d'extraire et caractériser les sujets ou centres d'intérêt à partir d'un corpus de documents.
LDA, présentée au chapitre 3, est une méthode générative non-supervisée de modélisation automatique des thématiques qui se base sur les hypothèses suivantes :

- Chaque document du corpus est un ensemble de mots représentés en bag of words ;
- Chaque document aborde un certain nombre de thématiques dans différentes proportions qui lui sont propres ;
- Chaque mot possède une distribution associée à chaque thématique qui est représentée par une probabilité sur chaque mot.

LDA, étant un modèle génératif, définit une probabilité de distribution jointe sur les différentes variables latentes identifiées. Les différentes probabilités de distribution associées à ces variables, sont utilisées pour retrouver les distributions latentes par rapport aux variables observées. Dans notre cas, nous cherchons, à partir des différents documents, les différentes proportions de thématiques pour chaque document, les distributions de mots sur les thématiques et les proportions d'apparition d'une thématique sur le corpus. Nous utilisons la distribution de Dirichlet [Blei and Jordan, 2006] sur la proportion globale des

thématiques ainsi que sur chaque distribution de thématique sur les mots. L'hypothèse de cette distribution est que les différentes thématiques sont complètement indépendantes. Ceci permettra de déterminer la thématique d'un document donné.

Techniquement, l'inférence de ce modèle est relativement complexe. Donc, des approximations et des algorithmes simplifiant le modèle sont nécessaires afin de pouvoir l'appliquer. Dans notre travail, nous nous sommes appuyés sur la technique statistique *Gibbs Sampling*.

En outre, nous utilisons des packages existants afin d'effectuer l'inférence sur nos données. Pour exécuter *LDA*, nous utilisons l'outil *Mallet* [McCallum, 2002] basé sur l'environnement d'exécution *Java*.
Les paramètres de la méthode *LDA* sont fixés sur la base des règles généralement utilisées dans la littérature : $\alpha = \frac{50}{T}, \beta = 0.01, \#$ itérations $= 1000$.

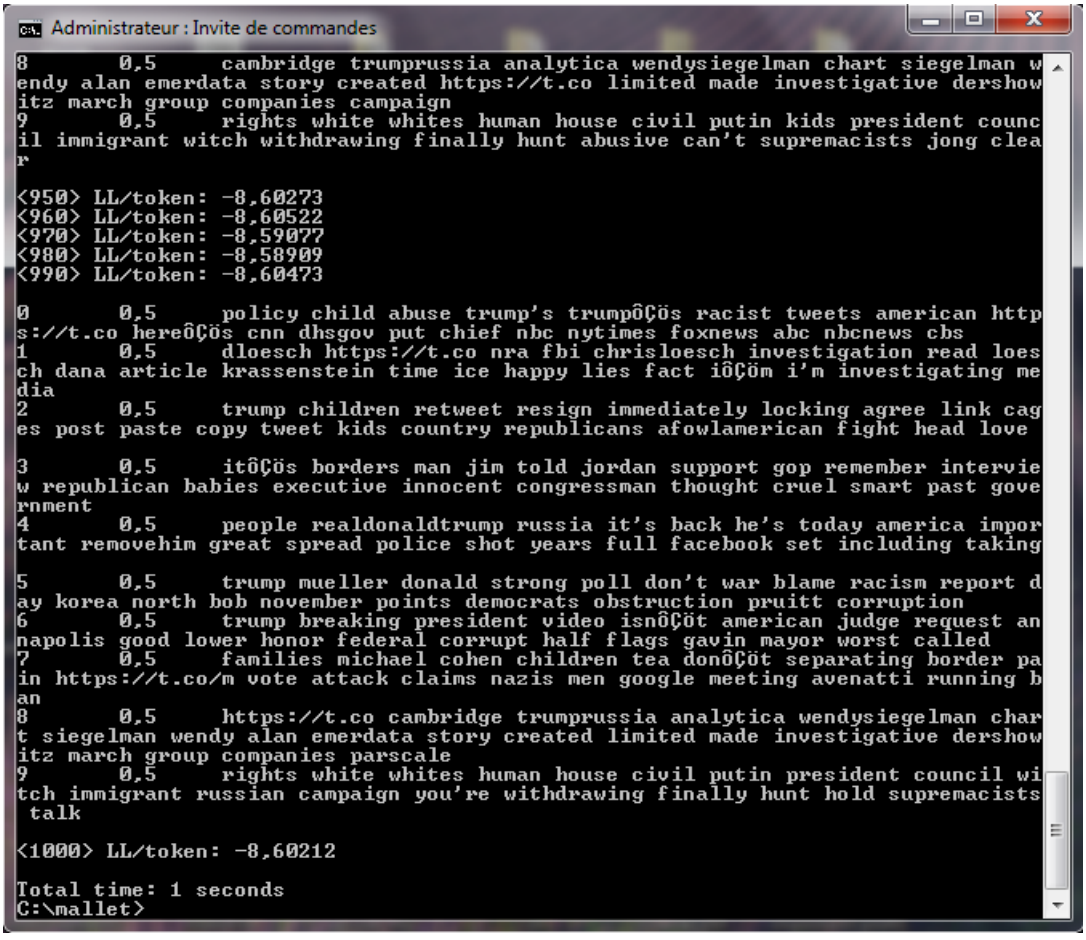
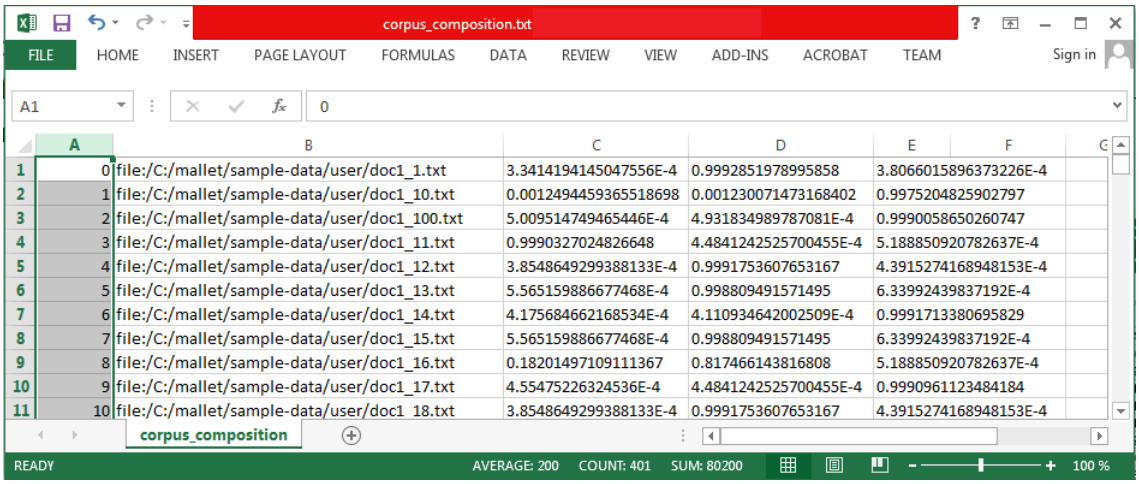


FIGURE 5.5 – Résultat d'application du modèle *LDA* sur une partie du corpus de données

A partir de la ligne de commande, nous exécutons le modèle *LDA*. En utilisant les paramètres par défaut et de manière non supervisée, l'algorithme essaie de trouver la

meilleure division de mots en sujets d'intérêt à partir d'une collection de documents. A l'exécution, les mots clés affichés aident à définir un sujet statistiquement significatif (voir *figure* 5.5).

En modifiant les paramètres, nous recherchons les trois sujets les plus abordés dans le corpus. Afin d'éviter le risque d'obtenir des optima locaux, nous exécutons le test cinq fois et nous retenons la moyenne. Quelles thématiques latentes composent les documents ? La réponse est dans le fichier décrit sur la *figure* 5.6, où la première colonne correspond au numéro du document, la deuxième son chemin d'accès et les autres colonnes représentent la proportion de chaque sujet dans le document correspondant.



	A	B	C	D	E
1	0	file:/C:/mallet/sample-data/user/doc1_1.txt	3.3414194145047556E-4	0.9992851978995858	3.8066015896373226E-4
2	1	file:/C:/mallet/sample-data/user/doc1_10.txt	0.0012494459365518698	0.001230071473168402	0.9975204825902797
3	2	file:/C:/mallet/sample-data/user/doc1_100.txt	5.009514749465446E-4	4.931834989787081E-4	0.9990058650260747
4	3	file:/C:/mallet/sample-data/user/doc1_11.txt	0.9990327024826648	4.4841242525700455E-4	5.188850920782637E-4
5	4	file:/C:/mallet/sample-data/user/doc1_12.txt	3.8548649299388133E-4	0.9991753607653167	4.3915274168948153E-4
6	5	file:/C:/mallet/sample-data/user/doc1_13.txt	5.565159886677468E-4	0.998809491571495	6.33992439837192E-4
7	6	file:/C:/mallet/sample-data/user/doc1_14.txt	4.175684662168534E-4	4.110934642002509E-4	0.9991713380695829
8	7	file:/C:/mallet/sample-data/user/doc1_15.txt	5.565159886677468E-4	0.998809491571495	6.33992439837192E-4
9	8	file:/C:/mallet/sample-data/user/doc1_16.txt	0.18201497109111367	0.817466143816808	5.188850920782637E-4
10	9	file:/C:/mallet/sample-data/user/doc1_17.txt	4.55475226324536E-4	4.4841242525700455E-4	0.9990961123484184
11	10	file:/C:/mallet/sample-data/user/doc1_18.txt	3.8548649299388133E-4	0.9991753607653167	4.3915274168948153E-4

FIGURE 5.6 – Composition des différents documents en thématiques latentes

5.1.3 Mise à jour de l'ontologie FOAF

Une fois les sujets d'intérêts identifiés, l'ontologie *FOAF* représentant le profil utilisateur social est mise à jour avec les sujets d'intérêt des utilisateurs correspondants.

Dans la *figure* 5.7, nous montrons un extrait du Réseau Social dans l'éditeur des ontologies *Protégé* construit en utilisant notre échantillon collecté. Les noeuds représentent les utilisateurs et les liens entre eux représentent les relations d'amitié.

5.2 Observations sur l'échantillon de données étudié

5.2.1 Description des données

Nous rappelons l'hypothèse de notre travail qui consiste en la supposition que les utilisateurs ayant une relation d'influence entre eux ont tendance à être similaires dans

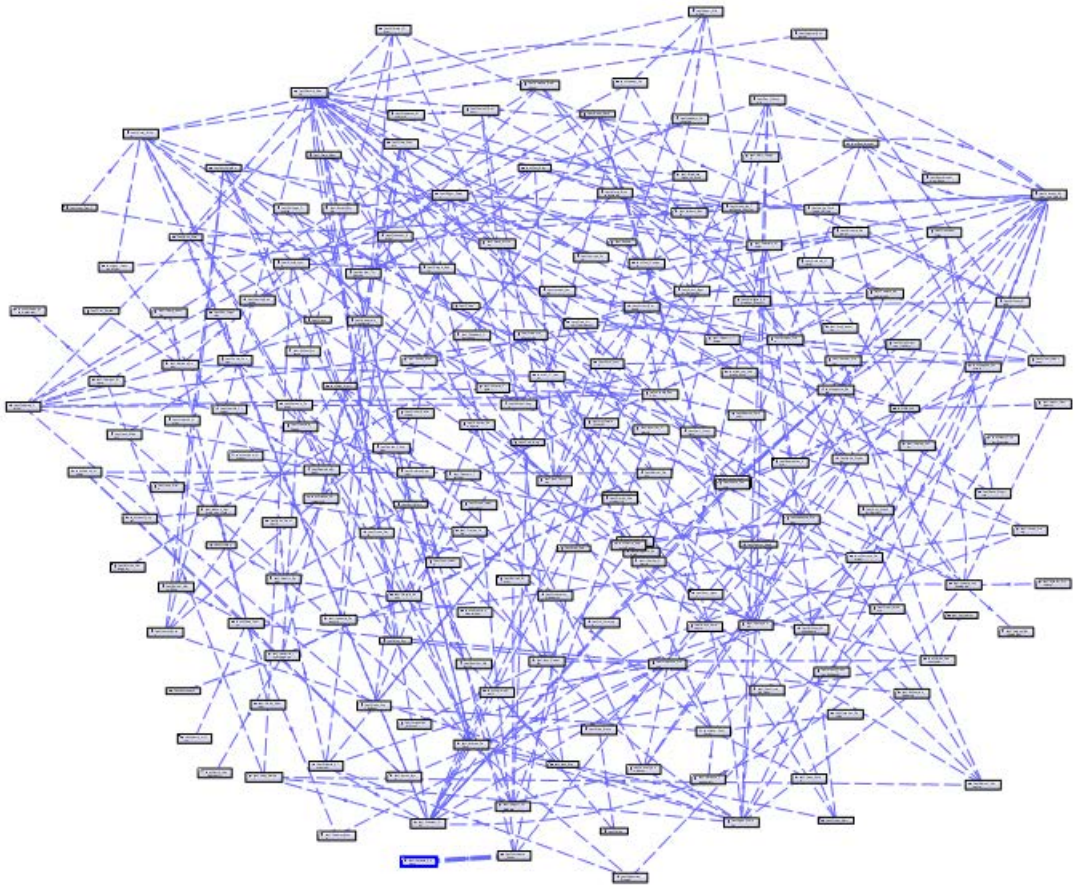


FIGURE 5.7 – Extrait du Réseau Social dans Protégé

leurs intérêts et sentiments. Pour l'étude de cette similarité entre les utilisateurs, nous avons exploré Twitter et Facebook et recueilli des milliers de données.

Pour mener les expériences et étudier les similarités, il nous faut un ensemble de données composé de :

- un ensemble V d'utilisateurs annotés (sentiments positifs ou négatifs) sur un sujet d'intérêt spécifique q,
- ensembles des amis et des abonnés de tous les utilisateurs $\in V$,
- les publications des utilisateurs $\in V$ au sujet de q et qui sont annotées,
- les activités comme "like", "commentaire" et "partage" (Retweet) exercées sur les publications retenues ;
- les données démographiques des utilisateurs (genre, profession et nationalité).

Au meilleur de notre connaissance, il n'existe aucun jeu de données contenant toutes les informations susmentionnées. Ceci nous a donc incité à collecter notre propre échantillon de données.

Le *tableau* 5.1 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

TABLE 5.1 – 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* 5.8, nous décrivons la distribution des publications positives et négatives sur les différents sujets d'intérêt.

5.2.2 Corrélation entre sujets d'intérêt et caractéristiques et relations sociales

Nous présentons une étude empirique sur les corrélations entre d'une part la similarité d'intérêt des utilisateurs et d'autre part diverses caractéristiques et relations d'influence sociales. Cette étude repose sur un ensemble de données capturé à partir des Réseaux Sociaux Twitter et Facebook. Nous identifions les caractéristiques sociales à partir de deux types d'informations : les intérêts des utilisateurs et les informations du contexte.

Aucune étude dans la littérature, au mieux de nos connaissances, n'a été étendue pour discuter de la façon dont la similarité d'intérêt varie avec diverses caractéristiques et relations sociales dont l'influence.

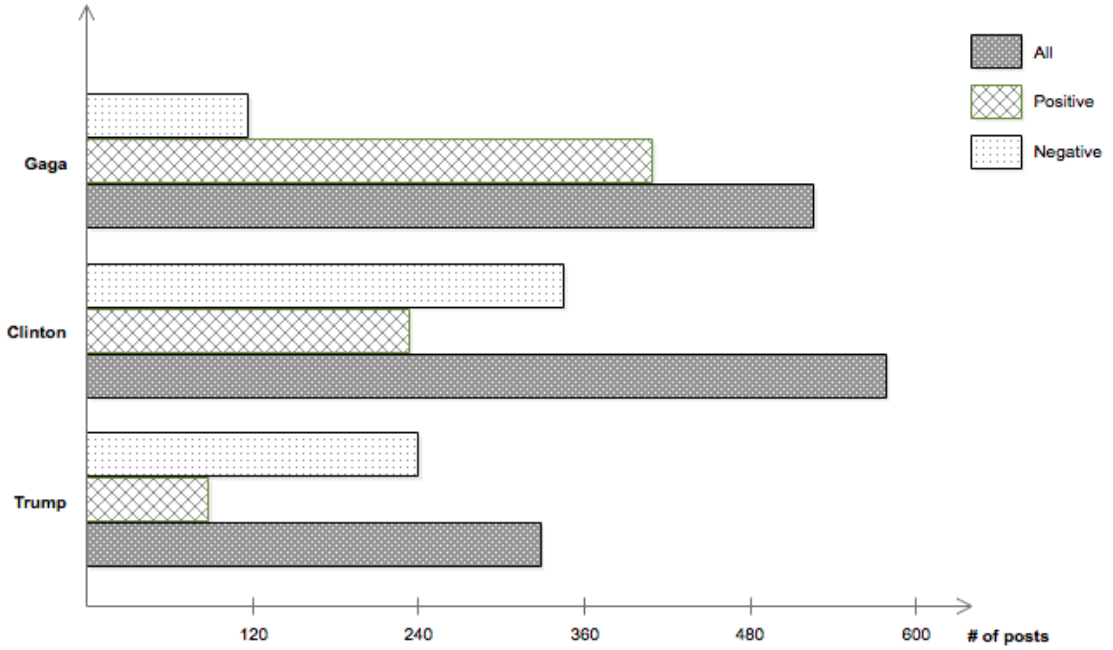


FIGURE 5.8 – Distributions des publications positives et négatives

En particulier, nous quantifions la similarité d'intérêt sur une agrégation de paires d'utilisateurs basée sur une mesure de similarité des sujets d'intérêt pour capturer l'intersection des intérêts entre deux utilisateurs. En outre, nous considérons les informations de contexte sur deux dimensions : personnelle et sociale ; les relations d'influence et les sujets d'intérêt. Notons que nous construisons notre échantillon de données simplement avec les informations publiques des utilisateurs et anonymisons tous les utilisateurs pendant l'analyse.

Nous rappelons les informations contextuelles étudiées et qui sont abordées au chapitre 3 : *GC* (combinaison de genre), *PC* (combinaison professionnelle), *NC* (combinaison de nationalité), et *CD* (distance de connectivité). Chaque information peut avoir deux valeurs possibles 0 si les caractéristiques correspondantes aux utilisateurs de la paire étudiée sont différents (par exemple $GC(v_i, v_j) = 0$ si v_i et v_j sont de genres différents) et 1 si elles sont identiques. En outre, nous menons notre étude par rapport aux intérêts communs (*CI*) pour chaque paire d'utilisateurs ($CI(v_i, v_j) = 1$ si v_i et v_j ont des sujets d'intérêt communs et 0 sinon) ; et la mesure d'influence (*INFLUENCE*) qui est égale à 1 si la valeur de *ROI* (Ratio of Influence) pour la paire correspondante est non nulle , et 0 sinon.

Pour chaque classe (0 ou 1) de l'information étudiée, nous distinguons en rouge les paires d'utilisateurs ayant des sujets d'intérêt similaires et en bleu les paires d'utilisateurs

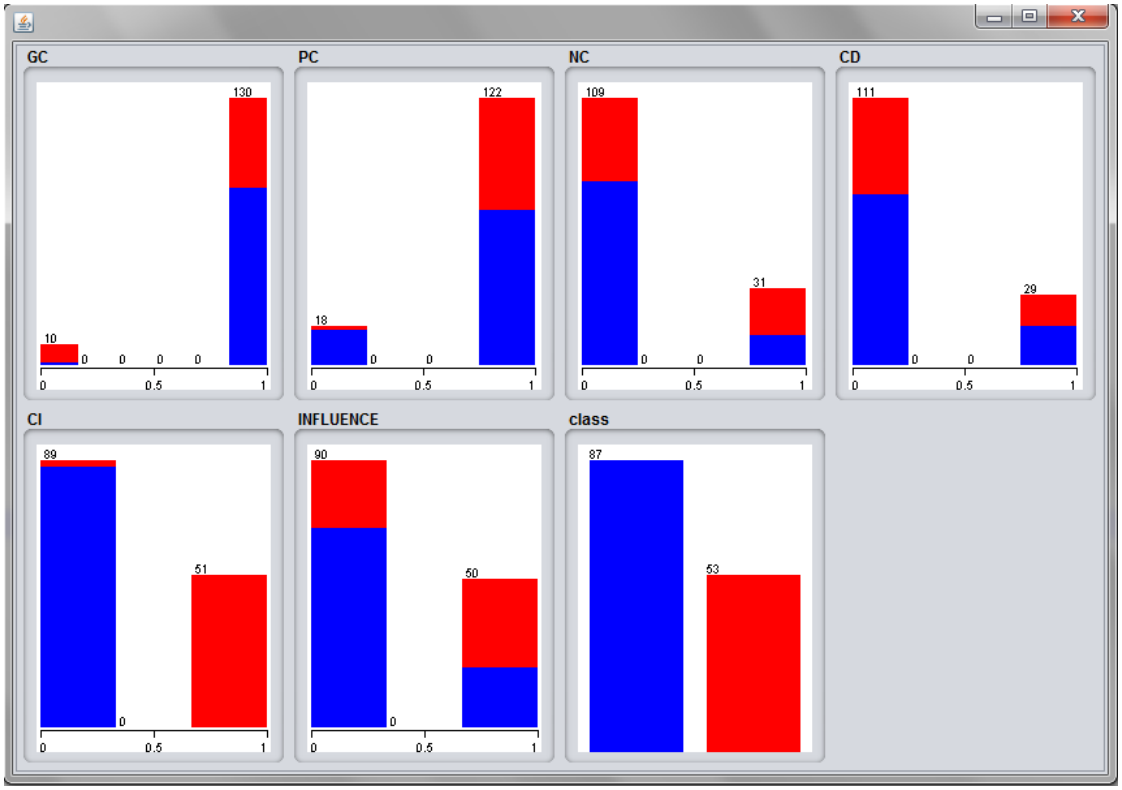


FIGURE 5.9 – Corrélation entre sujets d'intérêt et caractéristiques et relations sociales

qui n'ont pas des intérêts similaires.

Pour mettre en évidence nos principales conclusions, nous révélons l'homophilie et l'influence conjointement concernant la similarité d'intérêt en se basant sur la présente analyse empirique (voir *figure* 5.9). De manière générale, l'homophilie montre une homogénéité dans les Réseaux Sociaux des personnes en ce qui concerne de nombreuses caractéristiques sociodémographiques et comportementale. En plus, l'influence, montre une homogénéité en ce qui concerne les relations d'influence entre les utilisateurs.

Plus précisément :

- L'homophilie révèle que les utilisateurs sont plus susceptibles d'avoir le même intérêt s'ils ont des informations démographiques plus semblables telles que le genre, la profession et la nationalité.
- L'homophilie implique également que les "amis" ont une plus grande similitude d'intérêt que les étrangers.
- En outre, l'homophilie indique que les utilisateurs ayant un intérêt plus populaire sont susceptibles de partager plus d'intérêts entre eux.

- L'influence sociale révèle que les utilisateurs ayant une relation d'influence entre eux sont susceptibles de partager le même intérêt.

5.2.3 Corrélation entre sentiments et relations d'influence sociale

Nous avons défini deux types de statistiques pour étudier l'interaction entre les polarités des sentiments des utilisateurs et les relations d'influence. Ces statistiques sont les suivantes :

- Probabilité que deux utilisateurs influencés l'un par l'autre (ou bien un seul qui influence l'autre) conditionnée par la même polarité de sentiment : cette statistique mesure l'influence conditionnée sur les polarités. La *Figure* 5.10 montre que le sentiment partagé tend à impliquer de l'influence. En fait, dans le graphique résultant, les utilisateurs ont plus de chance d'avoir une relation d'influence s'ils partagent le même sentiment que s'ils sont différents.

- 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* 5.11 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 grande que le hasard.

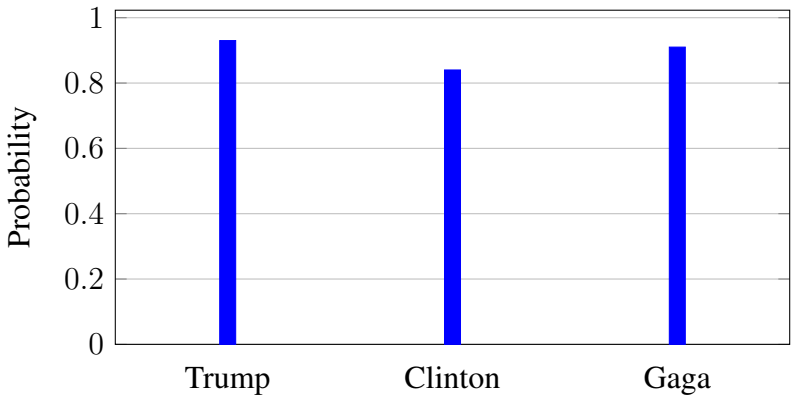


FIGURE 5.10 – Probabilité d'influence conditionnée ou non par la même polarité de sentiment

En résumé, les paires d'utilisateurs dans lesquelles au moins l'un influence l'autre ont tendance à avoir la même polarité de sentiment ; et deux utilisateurs partageant la même polarité de sentiment sont plus susceptibles d'être influencées l'un par l'autre que deux

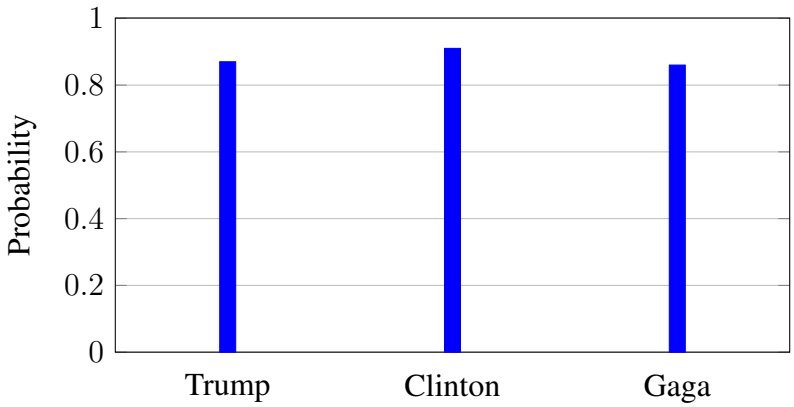


FIGURE 5.11 – Probabilité d'avoir la même polarité de sentiment conditionnée par une relation d'influence

utilisateurs ayant des sentiments différents. Ces observations valident notre hypothèse selon laquelle influence et sentiment partagé sont clairement corrélés.

5.3 Analyse de la performance

5.3.1 Prédiction des sujets d'intérêt

Afin de valider les résultats obtenus, nous évaluons la performance des modèles en utilisant les mesures suivantes : Précision (*P*), Rappel (*R*) et F1-score (*F1*).

$$P = \frac{TP}{TP + FP} \tag{5.1}$$

$$R = \frac{TP}{TP + FN} \tag{5.2}$$

$$F1 = \frac{2PR}{P + R} \tag{5.3}$$

L'exactitude "accuracy" est calculée en utilisant ces mesures. Son équation est la suivante :

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{5.4}$$

Où : *TP* est le nombre des vrais positifs, *FP* des faux positifs, *FN* des faux négatifs et *TN* des vrais négatifs en terme de prédiction.

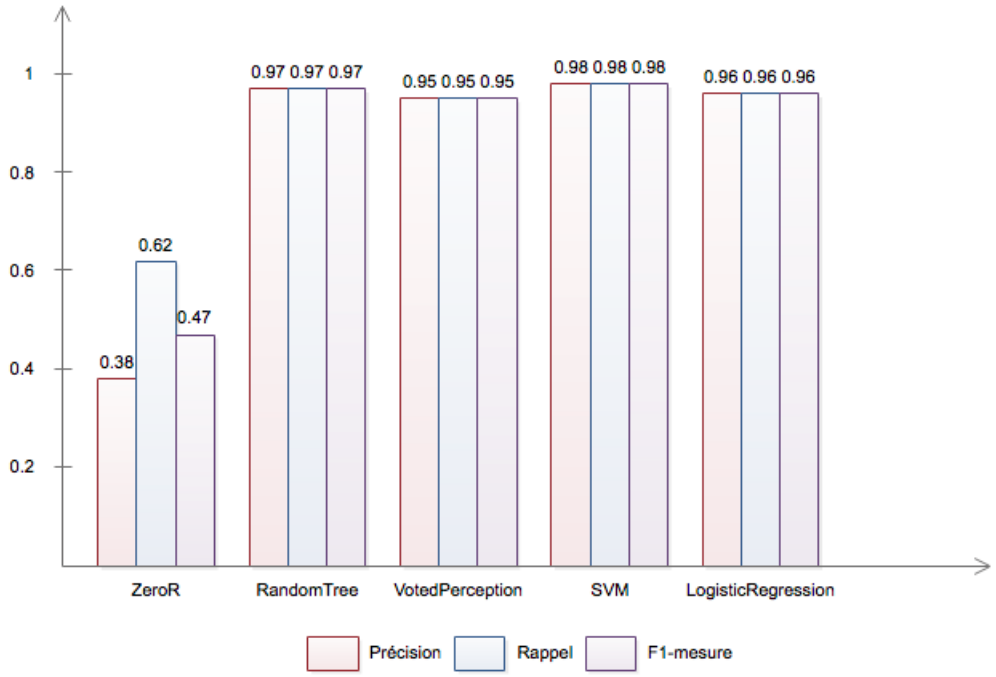


FIGURE 5.12 – Comparaison des mesures de présicion, rappel et *F1*

La première expérience a pour objectif d'évaluer la performance du modèle de prédiction des intérêts. Nous avons effectué la tâche de prédiction en utilisant différentes méthodes d'apprentissage automatique standards. Cinq algorithmes de classification ont été sélectionnés : *ZeroR*, *RandomTree*, *VotedPerception*, *SVM* et *LogisticRegression*. 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 défini dans le chapitre 3. Nous avons mené cette expérience et nous avons obtenu les résultats montrés dans la *figure* 5.12.

Les résultats obtenus indiquent clairement que la méthode *SVM* fonctionne bien avec les mesures de précision.

5.3.2 Prédiction des sentiments

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 parmi les k labels possibles.

Nous menons des expériences 10 fois. À chaque fois, les données avec les labels de

vérité de terrain 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 annotés restants. Le rapport des deux ensembles est différent dans différents sujets.
Nous comparons deux méthodes de classification des utilisateurs afin d'évaluer nos résultats comme le montre la *figure* 5.13.

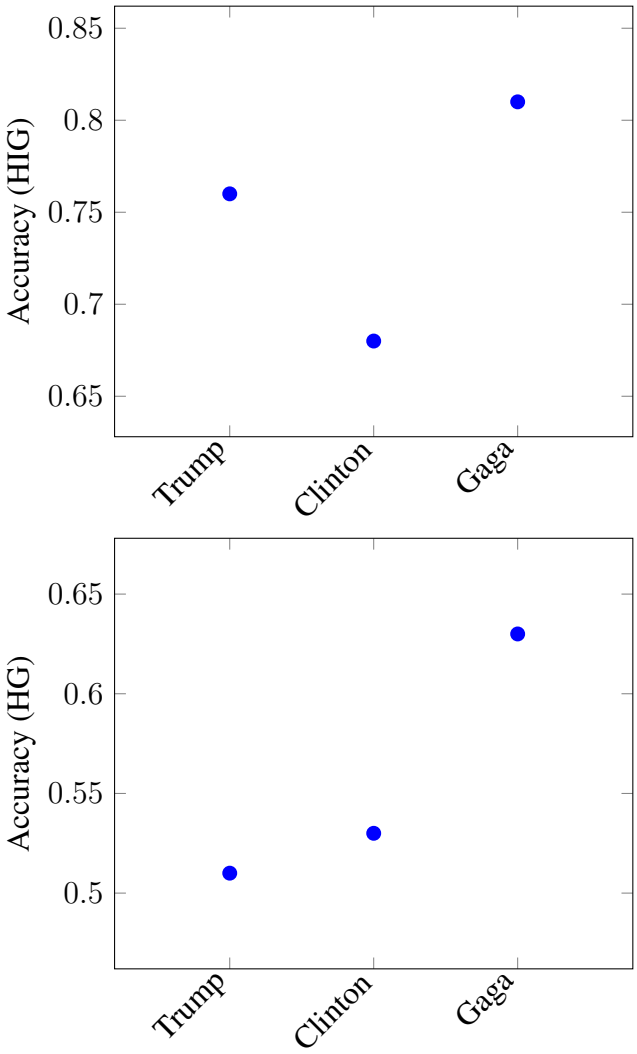


FIGURE 5.13 – Analyse de performance des deux méthodes

- Modèle de graphe hétérogène (HG) [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 pour obtenir des labels de sentiment au niveau utilisateur.

- Modèle de graphe d'influence hétérogène : nous exécutons notre modèle d'appren-

tissage semi-supervisé sur le graphe d'influence hétérogène que nous avons défini pour obtenir la classification des sentiments des utilisateurs.

Les résultats confirment l'amélioration de la performance dans la prédiction du sentiment au niveau utilisateur. La *Figure 5* montre ces résultats pour les différentes méthodes et que notre modèle atteint les plus hautes exactitudes. Enfin, les mesures *F1* sont rapportées dans le *tableau* 5.2.

	Trump	Clinton	Gaga
HIG	0.38	0.35	0.56
HG	0.23	0.31	0.33

TABLE 5.2 – Mesures $F1$

5.3.3 Détection des communautés d'intérêt

Une fois les communautés détectées à l'aide de la méthode proposée, il convient d'évaluer leur qualité. Pour cela, nous utilisons deux critères externes (comme mentionné au chapitre 1). Ces derniers permettent de comparer le résultat obtenu avec un résultat attendu ; dans notre cas, une communauté agissant comme une vérité de terrain. L'évaluation des résultats de la classification est effectuée par rapport à cette communauté. En effet, la comparaison est utilisée pour calculer le taux d'entités bien classées. Ce taux est le rapport entre le nombre d'entités (utilisateurs) bien classées et le nombre total d'entités classées. Il peut être mesuré à partir de la matrice de coïncidence où chaque ligne correspondant à un sujet d'intérêt, chaque colonne à une communauté. Un terme de la matrice contiendra le nombre d'éléments présentant un intérêt de la ligne correspondante qui ont été attribués à la communauté de la colonne correspondante. La matrice n'est pas nécessairement carrée, l'algorithme peut produire un nombre de classes différent de celui de la vérité de terrain. Également à partir de cette matrice, nous pouvons calculer une autre mesure qui est la précision par communauté détectée. C'est le rapport entre le nombre d'éléments bien classés et le nombre total d'éléments affectés à la communauté. Cette mesure évalue la qualité de la communauté.

Afin de comparer notre approche avec d'autres méthodes, nous considérons deux méthodes de référence à comparer à notre algorithme, à savoir les algorithmes de Clauset-Newamn-Moore [Clauset et al., 2004] et de Girvan-Newman [Girvan and Newman, 2002]. Les deux algorithmes sont basés sur la topologie du réseau. Le premier

algorithme est un algorithme d'agglomération hiérarchique basé sur une modularité maximale. Il utilise une optimisation gloutonne en visant à maximiser la modularité. Cependant, l'algorithme de Girvan-Newman [Girvan and Newman, 2002] est basé sur la suppression itérative des arêtes avec des scores élevés de centralité d'intermédiarité.

	[Clauset et al., 2004]	[Girvan and Newman, 2002]	**Notre approche**
Nombre de communautés	16	26	**3**
Taux de bien classés (%)	8.05	5.78	**92.25**

TABLE 5.3 – Comparaison des résultats de classification des communautés

Le *tableau* 5.4 montre les résultats de précision obtenus suite à l'application de notre approche.

	Trump	*Clinton*	*Gaga*
Communauté 0	0.77	0.96	0.97
Communauté 1	0.98	0.55	0.92
Communauté 2	0.92	0.75	0.96

TABLE 5.4 – Comparaison des résultats de précision de classfication des communautés

5.4 Efficacité du filtrage d'information

Pour évaluer l'efficacité des *PDBA* présentées au chapitre 4, nous nous sommes focalisés principalement à évaluer la performance du processus de filtarge des documents qui est défini et implémenté au niveau du *DSL* 1.

Différentes mesures pour l'évaluation de la performance de tels systèmes de filtrage ont été proposées. Les deux importantes mesures les plus communément utilisées sont :

la précision et le rappel. Elles nécessitent la connaissance de tous les documents utilisés pour calculer le nombre de documents pertinents. La précision représente le nombre de documents pertinents retrouvés par rapport au nombre total de documents disponibles. Alors que le rappel est défini par le nombre de documents pertinents rapporté au nombre de documents pertinents disponibles.

$$precision = \frac{\{relevant\,documents\} \cap \{retrieved\,documents\}}{\{retrieved\,documents\}} \tag{5.5}$$

$$recall = \frac{\{relevant\,documents\} \cap \{retrieved\,documents\}}{\{relevant\,documents\}} \tag{5.6}$$

En reprenant l'exemple de motivation présenté au chapitre 4, nous appliquons les plug-ins générés suite à l'exécution des *DSL*s en simulant une application de e-recrutement. Les résultats d'efficaité sont rapportés au *tableau* 5.5.

	precision	recall
Corpus de e-recrutement	0.96	0.98

TABLE 5.5 – Efficacité du filtrage des documents

5.5 Conclusion

Dans ce chapitre, nous avons présenté les expérimentations que nous avons mené afin d'évaluer nos propositions.

D'abord, nous avons collecté un échantillon de données à partir de deux Réseaux Sociaux qui sont *Twitter* et *Facebook*. Ensuite, à partir des publications textuelles des utilisateurs, nous avons appliqué la méthode *LDA* de modélisation automatique des sujets d'intérêt. Nous avons utilisé les publications sous forme de documents dans le but d'explorer les thématiques latentes abordées et qui intéressent les utilisateurs. Ceci nous a permis de construire la dimension des intérêts dans le profil utilisateur socialet la mise à jour de l'ontologie *FOAF* qui le représente. Les sujets d'intérêt ainsi capturé sont utilisés pour alimenter le modèle de prédiction des intérêts qui est basé sur la méthode *SVM* et profitant des deux phénomènes d'homophilie et d'influence dans les Réseaux Sociaux.

En outre, nous avons mené des tests de performance du modèle d'Analyse des Sentiments au niveau des utilisateurs et basé sur deux facteurs qui sont : Utilisateur-Publication et Utilisateur-Utilisateur.

L'étude empirique que nous avons effectué nous a permis de confirmer la corrélation applicative entre d'une part les similarités des intérêts ainsi que des sentiments et d'autre part les relations d'influence. A l'issu de ces vérifications, l'expérimentation a confirmé la perspicacité de nos hypothèses.

De plus, nous avons montré l'efficacité de notre approche de détection des communautés en présentant les résultats d'application des critères d'évaluation. Nous avons confronté ces résultats aux résultats d'application d'autres méthodes connues dans la littérature.

Enfin, nous avons évalué la performance du processus d'extraction des documents pertinents que nous avons défini au niveau du *DSL* utilisé dans la génération semi-automatique des applications sociales personnalisées. Cette évaluation est basée sur deux mesures sui sont la précision et le rappel.

En conclusion des résultats obtenus, nous avons validé empiriquement les modèles et algorithmes définis et implémentés. Le bilan de ces expérimentations semble favorable de nouvelles perspectives ? Ces dernières sont présentées dans le prochain et dernier chapitre.

CONCLUSION ET PERSPECTIVES

Sommaire

I Résumé conclusif

Suite à la démocratisation des technologies numériques, les Réseaux Sociaux qui sont les nouveaux moyens de communication et de collaboration ont évolué en un environnement de production active de contenus. Les chercheurs s'intéressent davantage à cette évolution. Nous avons montré dans cette thèse que les contenus disponibles au sein des Réseaux Sociaux peuvent être sources de connaissances. Nous avons inscrit ces données sémantiques dans le cadre de la détection des communautés d'intérêt en montrant qu'en plus de l'analyse structurelle, l'intégration des données sémantiques permettent d'améliorer la qualité des communautés détectées.

Ce travail est essentiellement consacré à la proposition d'une approche de détection des communautés d'intérêt dans les Réseaux Sociaux et leur exploitation dans la génération d'applications sociales interactives personnalisées. Plus spécifiquement, les principales contributions sont :

1. Définition d'un modèle sémantique des données sociales sous la forme d'un profil utilisateur social générique et extensible ; qui est représenté par l'ontologie *FOAF* que nous avons étendue avec de nouvelles propriétés. Dans le chapitre 3, nous avons présenté les différentes phases du processus de modélisation de ce profil. D'après l'étude menée au chapitre 1, il s'est avéré que différentes ontologies ont été conçues pour décrire sémantiquement les Réseaux Sociaux. *FOAF* est un vocabulaire de base qui semble être assez exhaustif pour la description des individus,

leurs attributs, leurs contenus et leurs relations. Nous avons donc opté pour cette ontologie pour représenter le profil utilisateur social que nous interrogeons grâce à des requêtes *SPARQL* spécifiées. Nous l'avons étendue avec des attributs comme la profession et la nationalité ; et des relations d'influence qui nous semblent être intéressants dans l'inférence des similarités entre les utilisateurs.

2. Proposition d'une nouvelle approche de détection des communautés d'intérêt dans les Réseaux Sociaux en tirant profit de leurs topologies sous-jacentes et la sémantique des contenus conjointement. Dans le chapitre 3, nous avons présenté cette approche qui est une approche sémantique basée sur le contexte utilisateur et orientée donnée. Organisée en trois étapes, elle repose non seulement sur l'extraction explicite des connaissances mais aussi sur l'extraction implicite à l'aide de modèles de prédictions. En outre et afin de garantir l'évolution des communautés détectées, nous avons appliqué les principes de l'équilibre social en évitant les conflits et divisant une communauté en deux sous communautés. Nous avons montré dans le chapitre 5 que l'application de notre approche permet d'avoir des valeurs considérables de précision de classification.

3. Proposition d'un modèle d'influence sociale pour mesurer les valeurs d'influence entre les utilisateurs en se basant sur un graphe hétérogène que nous avons défini afin de profiter des facteurs Utilisateur-Publication et Utilisateur-Utilisateur à la fois. L'influence est un phénomène crucial dont dépendent les comportements et les interactions des utilisateurs dans les Réseaux Sociaux. En particulier, le graphe hétérogène permet la représenter différents types de noeuds et de liens.

4. Proposition d'un modèle pour la prédiction des intérêts des utilisateurs dans les Réseaux Sociaux basé sur les deux phénomènes d'homophilie et d'influence. Notre hypothèse est que la similarité des intérêts entre les utilisateurs dépend de deux phénomènes clés dans les Réseaux Sociaux qui sont l'homophilie et l'influence. L'étude expérimentale que nous avons menée dans le chapitre 5 confirme notre hypothèse.

5. Proposition d'un modèle de prédiction des sentiments des utilisateurs à propos d'un sujet d'intérêt donné. Notre hypothèse est que la similarité des sentiments entre les utilisateurs dépend du phénomène de l'influence dans les Réseaux Sociaux. L'étude expérimentale que nous avons menée dans le chapitre 5 confirme notre hypothèse.

6. Proposition d'une méthode de personnalisation des applications interactives sociales basée sur les communautés d'intérêt.

7. Proposition d'une approche pour la génération semi-automatique des applications sociales personnalisées.

En se basant sur les critères de comparaison présentés dans le chapitre 2, nous reprenons le tableau 2.1 comparatif afin de positionner notre proposition par rapport aux approches déjà présentées.

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**
CCN model [Barbieri et al., 2013]	+	+	-	-	+	+	-	-	+	-	+	+	-	-
SA-Cluster [Zhou et al., 2009]	+	-	+	-	+	-	+	+	-	+	-	+	+	+
CESNA [Yang et al., 2013]	+	-	+	-	+	-	+	-	+	-	+	+	+	+
Hierarchical agglomeration [Clauset et al., 2004]	+	+	-	+	-	+	-	-	+	-	+	+	+	+
Link-Content model [Natarajan et al., 2013]	+	-	+	-	-	-	+	+	-	+	-	+	+	+
SemTagP [Erétéo et al., 2011]	+	+	+	+	-	+	-	+	-	+	+	-	+	+
Notre approche	**+**	**+**	**+**	**+**	**+**	**+**	**-**	**+**	**-**	**-**	**+**	**-**	**+**	**-**

TABLE .1 – Comparaison de quelques méthodes de détection de communautés dans les Réseaux Sociaux

Nous rappelons les critères utilisés qui sont : caractéristiques structurelles (A), activités sociales (B), attributs (C), contenu (D), influence sociale (E), orientation du graphe

(orienté (F) ou non orienté (G)), communautés détectées (couvrantes (H) ou non (I)), fonction de qualité (J), démarrage à froid (K), passage à l'échelle (L), qualité des communautés détectées (M), complexité (N).

Si notre approche présente un certain nombre d'avancées en matière de détection des communautés d'intérêt et leur exploitation dans la génération des applications sociales interactives à contenus personnalisés, elle présente aussi des points d'amélioration qui font partie de nos perspectives de recherche.

II Perspectives

Le travail présenté dans cette thèse a répondu à plusieurs questions concernant l'analyse des Réseaux Sociaux ainsi que les applications sociales interactives. Il a ainsi ouvert de nombreuses pistes prometteuses dans un domaine en plein essor :

- Pour commencer par l'approche de détection des communautés d'intérêt, il nous paraît opportun de prendre en compte la dimension temporelle. En effet, les intérêts des utilisateurs évoluent dans le temps [Sendi et al., 2018]. Donc, les communautés dynamiques reflèteront de plus en plus la structure réelle des Réseaux Sociaux. Il serait intriguant d'étudier l'évolution temporelle des membres d'une communauté.

- En ce qui concerne le modèle d'Analyse des Sentiments, nous pensons qu'il serait intéressant d'inférer les polarités des sentiments aux deux niveaux utilisateur et publication.

- Dans les Réseaux Sociaux, en plus des textes, ils existent beaucoup de données multimédias telle que les images et les vidéos. Il serait donc pertinent de les intégrer dans le traitement des données sociales.

- D'un point de vue général, nous pensons que la perspective primordiale est récupérer plus de données sur différents sujets d'intérêt. L'exploitation d'autres domaines garantira mieux la généricité de nos propositions.

- Si dans notre travail, nous nous sommes basés sur la méthode *LDA* pour la détection des sujets d'intérêt dans les publications des utilisateurs, il semble essentiel de trouver un nombre "correct" de sujets S. Des travaux se sont proposés d'optimiser un critère afin de chercher automatiquement la meilleure valeur de S [Rosen-Zvi et al., 2004]. Cependant, d'autres travaux ont tenté de se passer complètement de ce paramètre [Blei et al., 2003a]. Ceci s'avère donc un important axe de recherche.

- Développer une application de simulation pour évaluer l'approche de génération des applications sociales sur différents scénarios. Pour aller plus loin dans les mécanismes d'adaptation de l'information, plusieurs manières d'utiliser les deux dimensions proposés sont envisageables.

Bibliographie

Ifttt : Put the internet to work for you. URL https://ifttt.com/.

K. Abbas. Systeme d'acces a l'information : application au domaine medical, 2008.

G. D. Abowd, Anind K. Dey, Peter J. Brown, Nigel Davies, Mark Smith, and Pete Steggles. Towards a better understanding of context and context-awareness", booktitle="handheld and ubiquitous computing. pages 304–307, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg.

G. Agarwal and D. Kempe. Modularity-maximizing graph communities via mathematical programming. *Eur. Phys. J. B*, 66(3) :409–418, 2008. doi : 10.1140/epjb/e2008-00425-1.

C. Aggarwal, editor. *Social Network Data Analytics*. Springer, 2011. ISBN 978-1-4419-8461-6.

K. Akshi and M. S. Teeja. Sentiment analysis on twitter. *IJCSI International Journal of Computer Science Issues*, 9 :372–373, 2012.

G. Alec, B. Richa, and H. Lei. Twitter sentiment classification using distant supervision. *Processing*, pages 1–6, 2009.

N. A. Alves. Unveiling community structures in weighted networks. *Phys. Rev. E*, 76 : 036101, Sep 2007. doi : 10.1103/PhysRevE.76.036101.

A. Armstrong and J. Hagel. Creating value in the network economy. chapter The Real Value of On-line Communities, pages 173–185. Harvard Business School Press, Boston, MA, USA, 1999. ISBN 0-87584-911-3. URL http://dl.acm.org/citation.cfm?id=303444.303455.

A. M. Auvinen. Social media - the new power of political influence. https://www.martenscentre.eu/publications/social-media-and-politics-power-political-influence, 2012.

F. Bacha, K. Oliveira, and M. Abed. Using context modeling and domain ontology in the design of personalized user interface. *International Journal on Computer Science and Information Systems*, 6 :69–94, 2011.

A.L. Barabasi and R. Albert. Emergence of scaling in random networks. 286 :509–512, October 1999.

N. Barbieri, F. Bonchi, and G. Manco. Cascade-based community detection. In Stefano Leonardi, Alessandro Panconesi, Paolo Ferragina, and Aristides Gionis, editors, *WSDM*, pages 33–42. ACM, 2013. ISBN 978-1-4503-1869-3.

T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, 284 (5) :34–43, 2001.

J. W. Berry, Randall A. Bruce, Hendrickson, and Cynthia A. Phillips LaViolette. Tolerating the community detection resolution limit with edge weighting. *Physical Review E*, 2009. doi : 10.1103/PhysRevE.83.056119.

M.A. Beyer and D. Laney. The importance of "big data" : A definition. 2012.

D. Blei and M. Jordan. Variational inference for Dirichlet process mixtures. *Journal of Bayesian Analysis*, 1 :121–144, 2006.

D. M. Blei, M. I. Jordan, T. L. Griffiths, and J. B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS'03, pages 17–24, Cambridge, MA, USA, 2003a. MIT Press.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3 :993–1022, March 2003b. ISSN 1532-4435.

D. M. Blei, T. L. Griffiths, and M. I. Jordan. The nested Chinese restaurant process and Bayesian inference of topic hierarchies. arXiv :0710.0845v2 [stat.ML], March 2008.

V.D. Blondel, J.L. Guillaume, R. Lambiotte, and E.L.J.S. Mech. Fast unfolding of communities in large networks. *J. Stat. Mech*, page P10008, 2008.

S. Boccaletti, M. Ivanchenko, V. Latora, A. Pluchino, and A. Rapisarda. Detecting complex network modularity by dynamical clustering. *Phys. Rev. E*, 75 :045102, 2007.

J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *J. Comput. Science*, 2(1) :1–8, 2011.

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

D. M. Boyd and N. B. Ellison. Social network sites : Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1), 2007.

M. Brambilla and A. Mauri. Model-driven development of social network enabled applications with webml and social primitives. In Michael Grossniklaus and Manuel Wimmer, editors, *ICWE Workshops*, volume 7703 of *Lecture Notes in Computer Science*, pages 41–55. Springer, 2012. ISBN 978-3-642-35622-3.

M. Brambilla, P. Fraternali, and C. Vaca. Bpmn and design patterns for engineering social bpm solutions. In Florian Daniel, Kamel Barkaoui, and Schahram Dustdar, editors, *Business Process Management Workshops (1)*, volume 99 of *Lecture Notes in Business Information Processing*, pages 219–230. Springer, 2011. ISBN 978-3-642-28107-5.

U. Brandes, D. Delling, M. Gaertler, R. G
"orke, M. Hoefer, Z. Nikoloski, and D. Wagner. On modularity-np-completeness and beyond. 2006. URL http://scholar.google.de/scholar.bib?q=info:qBbwVpcNTssJ:scholar.google.com/&output=citation&hl=de&as_sdt=0,5&ct=citation&cd=0.

J. G. Breslin, A. Harth, U. Bojars, and S. Decker. Towards semantically-interlinked online communities. In A. Gomez-Perez and J. Euzenat, editors, *European Semantic Web Conference (ESWC)*, volume 3532 of *Lecture Notes on Computer Science*, pages 500–514. Springer, 2005.

D. Brickley and L. Miller. FOAF Vocabulary Specification. Namespace Document 2 Sept 2004, FOAF Project, 2004. http ://xmlns.com/foaf/0.1/.

A. Brossard. Percomom : une methode de modelisation des applications interactives appliquees a l'information voyageur dans le domaine des transports, 2008.

C. Brun, F. Chevenet, D. Martin, J. Wojcik, A. Guenoche, and B. Jacq. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. 5 :R6, 02 2003.

S. Ceri, P. Fraternali, A. Bongio, M. Brambilla, S. Comai, and M. Matera. *Designing Data-Intensive Web Applications*. Elsevier, Amsterdam, Netherlands, December 2002. ISBN 1558608435.

William Y. C. Chen, Andreas W. M. Dress, and Winking Q. Yu. Community structures of networks. *Mathematics in Computer Science*, 1(3) :441–457, 2008.

N. Chouchani and M. Abed. Online social network analysis : detection of communities of interest. *Journal of Intelligent Information Systems*, Aug 2018a. ISSN 1573-7675. doi : 10.1007/s10844-018-0522-7.

N. Chouchani and M. Abed. A user-centered approach for integrating social actors into communities of interest. In *2018 12th International Conference on Research Challenges in Information Science (RCIS)*, pages 1–11, May 2018b. doi : 10.1109/RCIS.2018.8406681.

A. Clauset. Finding local community structure in networks. *Phys. Rev. E*, 72 :026132, August 2005. doi : 10.1103/PhysRevE.72.026132.

A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70 :066111, 2004.

D. Combe. *Détection de communautés dans les réseaux d'information utilisant liens et attributs. (Community detection in information networks using links and attributes).* PhD thesis, Jean Monnet University, Saint-Etienne, France, 2013.

D. Combe, C. Largeron, E. Egyed-Zsigmond, and M. Géry. Getting clusters from structure data and attribute data. In *ASONAM*, pages 710–712. IEEE Computer Society, 2012. ISBN 978-0-7695-4799-2.

I.D. Constantiou and J. Kallinikos. New games, new rules : big data and the changing context of strategy. *JIT*, 30(1) :44–57, 2015.

J. Darlay, N. Brauner, and J. Moncel. Partition en sous graphes denses pour la détection de communautés. 02 2010.

I. Davis and E. Vitiello. RELATIONSHIP : A vocabulary for describing relationships between people. Technical report, vocab.org.

A. Degenne and M. Forsé. Les réseaux sociaux. *Mathematics and Social Sciences*, 42 (168) :5–9, 2004.

Y. Ding. Community detection : Topological vs. topical. *J. Informetrics*, 5(4) :498–514, 2011.

L. Donetti and M. A. Muñoz. Detecting network communities : a new systematic and efficient algorithm. *Journal of Statistical Mechanics : Theory and Experiment*, 2004 (10) :P10012, October 2004. ISSN 1742-5468. doi : 10.1088/1742-5468/2004/10/P10012.

A. Doucet, N. Lumineau, C. Berrut, N. Denos, B. Rumpler, Boughanem M. Soule-Dupuy C. Mouaddib N. Rocacher, D., and D. Kostadinov. Action spécifique sur la personnalisation de l'information, 2004.

G. Erétéo, F. Gandon, and M. Buffa. Semtagp : Semantic community detection in folksonomies. In Olivier Boissier, Boualem Benatallah, Mike P. Papazoglou, Zbigniew W. Ras, and Mohand-Said Hacid, editors, *Web Intelligence*, pages 324–331. IEEE Computer Society, 2011. ISBN 978-0-7695-4513-4.

M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology. *SIGCOMM*, pages 251–262, Aug-Sept. 1999.

U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. The kdd process for extracting useful knowledge from volumes of data. *Commun. ACM*, 39(11) :27–34, November 1996. ISSN 0001-0782. doi : 10.1145/240455.240464.

E. Ferrara, P. D. Meo, S. Catanese, and G. Fiumara. Detecting criminal organizations in mobile phone networks. *Expert Syst. Appl.*, 41(13) :5733–5750, 2014.

S. Fortunato. Community detection in graphs. *Physics Reports*, 486 :75–174, 2010.

E. Galanaki. The decision to recruit online : a descriptive study. *Career Development International*, 7(4) :243–251, 2002. doi : 10.1108/13620430210431325.

L. Garton, C. Haythornthwaite, and B. Wellman. Studying online social networks. *Journal of Computer-Mediated Communication*, 3(1) :0–0, 1997. ISSN 1083-6101. doi : 10.1111/j.1083-6101.1997.tb00062.x.

M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *PNAS*, 99(12) :7821–7826, June 2002.

A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *Processing*, pages 1–6, 2009.

J. Goldenberg, B. Libai, and E. Muller. Talk of the network : A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 2001.

M. Granovetter. Threshold models of collective behavior. *Am. Journal of Sociology*, 83 (6) :1420–1443, 1978.

S. Greengard. Picking–and keeping–the cream of the crop : Smart strategies are needed for both recruitment and retention of talent. *Human Resource Management International Digest*, 20(3) :26–29, 2012. doi : 10.1108/09670731211224357.

T. R. Gruber. Ontology. In Ling Liu and M. Tamer Özsu, editors, *Encyclopedia of Database Systems*, Springer Reference, pages 1963–1965. Springer, New York, 2009. doi : 10.1007/978-0-387-39940-9_1318.

K. Gueorgi and J. W. Duncan. Origins of homophily in an evolving social network. *American Journal of Sociology*, 115 :405–450, 2009.

A. Guille, H. Hacid, C. Favre, and D. A. Zighed. Information diffusion in online social networks : a survey. *SIGMOD Record*, 42(2) :17–28, 2013.

U. Hanani, B. Shapira, and P. Shoval. Information filtering : Overview of issues, research and systems. *User Modeling and User-Adapted Interaction*, 11(3) :203–259, August 2001. ISSN 0924-1868. doi : 10.1023/A:1011196000674.

F. Harary, R. Z. Norman, and D. Cartwright. *Structural models : an introduction to the theory of directed graphs*. Wiley, New York, 1965.

T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer Verlag, August 2001.

M. B. Hastings. Community detection as an inference problem, April 2006. URL http://arxiv.org/abs/cond-mat/0604429.

D. Heckmann, T. Schwartz, B. Brandherm, M. Schmitz, and Wilamowitz-Moellendorff M. Gumo - The General User Model Ontology. In *User Modeling 2005*, volume 3538, pages 428–432. Springer Berlin / Heidelberg, 2005.

M. Hepp. Hypertwitter : Collaborative knowledge engineering via twitter messages. In Philipp Cimiano and Helena Sofia Pinto, editors, *EKAW*, volume 6317 of *Lecture Notes in Computer Science*, pages 451–461. Springer, 2010. ISBN 978-3-642-16437-8.

Y. Hu, M. Li, P. Zhang, Y. Fan, and Z. Di. Community detection by signaling on complex networks. *Physical Review E*, 78(1) :016115, 2008. doi : 10.1103/PhysRevE.78.016115.

Y. E. Ioannidis and G. Koutrika. Personalized systems : Models and methods from an ir and db perspective. In Klemens Böhm, Christian S. Jensen, Laura M. Haas, Martin L. Kersten, Per-Åke Larson, and Beng Chin Ooi, editors, *VLDB*, page 1365. ACM, 2005. ISBN 1-59593-177-5.

K. S. Jones, S. Walker, and S. Robertson. A probabilistic model of information retrieval : development and status. Technical report, October 1998.

D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD '03*, 2003.

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

N. Khan, I. Yaqoob, I. Abaker, and T. Hashem. Big data : Survey, technologies, opportunities, and challenges. 2014.

J. Kim, J. Yoo, H. Lim, H. Qiu, Z. Kozareva, and A. Galstyan. Sentiment prediction using collaborative filtering. In Emre Kiciman, Nicole B. Ellison, Bernie Hogan, Paul Resnick, and Ian Soboroff, editors, *ICWSM*. The AAAI Press, 2013. ISBN 978-1-57735-610-3.

A. Kobsa, J. , Koenemann, and W. Pohl. Personalized hypermedia presentation techniques for improving online customer relationships. *The Knowledge Engineering Review*, 16 (2) :111–155, 2001.

J. Laconich, F. Daniel, F. Casati, and M. Marchese. From a simple flow to social applications. In Quan Z. Sheng and Jesper Kjeldskov, editors, *ICWE Workshops*, volume 8295 of *Lecture Notes in Computer Science*, pages 39–50. Springer, 2013. ISBN 978-3-319-04243-5.

D. Laney. 3D data management : Controlling data volume, velocity, and variety. Technical report, META Group, February 2001.

M. S. Laura, Z. Linhong, L. Kristina, and K. Zornitsa. The role of social media in the discussion of controversial topics. In *SocialCom*, pages 236–243. IEEE Computer Society, 2013. ISBN 978-0-7695-5137-1.

P. Lazarsfeld and R. Merton. Friendship as a social process : a substantive and methodological analysis. *Freedom and Control in Modern Society*, pages 18–66, 1954.

Roger Th. A. J. Leenders. Modeling social influence through network autocorrelation : constructing the weight matrix. *Social Networks*, 24(1) :21–47, 2002.

D. Leitch and M. Sherif. Twitter mood, ceo succession announcements and stock returns. *J. Comput. Science*, 21 :1–10, 2017.

J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. 2008.

H. Li, Z. Nie, W. Lee, C. L. Giles, and J. Wen. Scalable community discovery on textual data with relations. In James G. Shanahan, Sihem Amer-Yahia, Ioana Manolescu, Yi Zhang, David A. Evans, Aleksander Kolcz, Key-Sun Choi, and Abdur Chowdhury, editors, *CIKM*, pages 1203–1212. ACM, 2008. ISBN 978-1-59593-991-3.

J. P. Li, A. Vishwanath, and H. R. Rao. Retweeting the fukushima nuclear radiation disaster. *Commun. ACM*, 57(1) :78–85, 2014.

Rao H.R. Li, J. Twitter as a rapid response news service : An exploration in the context of the 2008 china earthquake. *The Electronic Journal on Information Systems in Developing Countries*, 42 :1–22, 2010.

D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM 2003)*, pages 556–559, New Orleans, LA, USA, 2003. ACM. ISBN 1-58113-723-0. doi : 10.1145/956863.956972.

S. Lin, Q. Hu, G. Wang, and P.S. Yu. Understanding community effects on information diffusion. In Tru Cao, Ee-Peng Lim, Zhi-Hua Zhou, Tu-Bao Ho, David Wai-Lok Cheung, and Hiroshi Motoda, editors, *PAKDD (1)*, volume 9077 of *Lecture Notes in Computer Science*, pages 82–95. Springer, 2015. ISBN 978-3-319-18037-3.

B. Liu. Sentiment analysis and subjectivity. *Handbook of Natural Language Processing*, 2, 2010.

J. Long, Y. Mo, Z. Ming, L. Xiaohua, and Z. Tiejun. Target-dependent twitter sentiment classification. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *ACL*, pages 151–160. The Association for Computer Linguistics, 2011. ISBN 978-1-932432-87-9.

B. Luciano and F. Junlan. Robust sentiment detection on twitter from biased and noisy data. In Chu-Ren Huang and Dan Jurafsky, editors, *COLING (Posters)*, pages 36–44. Chinese Information Processing Society of China, 2010.

J. Maisonneuve. Psychosociologie des affinités. Paris, PUF, 1966.

M. B. Margaret and J. L. Peter. Affective norms for english words (anew) : Instruction manual and affective ratings, 1999.

A. K. McCallum. Mallet : A machine learning for language toolkit. 2002. URL http://mallet.cs.umass.edu.

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

M. Mernik, J. Heering, and A. M. Sloane. When and how to develop domain-specific languages. *ACM Comput. Surv.*, 37(4) :316–344, December 2005. ISSN 0360-0300. doi : 10.1145/1118890.1118892.

S. Michael, S. Nikita, U. Sid, and B. Jason. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First Workshop on Unsupervised Learning in NLP*, EMNLP '11, pages 53–63, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-937284-13-8.

A. A. Middleton and D. S. Fisher. Three-dimensional random-field ising magnet : Interfaces, scaling, and the nature of states. *Physical Review B-Condensed Matter*, 65(13) : 1344111–13441131, 4 2002. ISSN 0163-1829.

S. Milgram. The small world problem. *Psychology Today*, 1 :62–67, 1967.

J. Moody. Race, school integration, and friendship segregation in america. *American Journal of Sociology*, 107 :679–716, 2002. doi : doi:10.1086/338954.

J. L. Moreno. Emotions mapped by new geography. http://diana-jones.com/wp-content/uploads/Emotions-Mapped-by-New-Geography.pdf, 1933.

M. Ringel Morris, J. Teevan, and S. Bush. Enhancing collaborative web search with personalization : Groupization, smart splitting, and group hit-highlighting. Association for Computing Machinery, Inc., November 2008.

F. Moser, R. Ge, and M. Ester. Joint cluster analysis of attribute and relationship data withouta-priori specification of the number of clusters. In *KDD '07 : Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 510–519, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-609-7. doi : http://doi.acm.org/10.1145/1281192.1281248.

N. Natarajan, P. Sen, and V. Chaoji. Community detection in content-sharing social networks. In Jon G. Rokne and Christos Faloutsos, editors, *ASONAM*, pages 82–89. ACM, 2013. ISBN 978-1-4503-2240-9.

A. Natasha and H. Lisa. Building a positive candidate experience : towards a networked model of e-recruitment. *Journal of Business Strategy*, 34(5) :36–47, 2013. doi : 10.1108/JBS-11-2012-0072.

T. Nepusz, A. Petróczi, L. Négyessy, and F. Bazsó. Fuzzy communities and the concept of bridgeness in complex networks. E-print, 2007. URL http://www.arxiv.org/abs/0707.1646.

M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*.

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

M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review*, E 69(026113), 2004.

M. E. J. Newman and E. A. Leicht. Mixture models and exploratory analysis in networks. *Proceedings of the National Academy of Sciences*, 104(23) :9564–9569, 2007. doi : 10.1073/pnas.0610537104. URL http://www.pnas.org/content/104/23/9564.abstract.

D. T. Nguyen and J. J. Jung. Real-time event detection on social data stream. *MONET*, 20(4) :475–486, 2015.

D. Nozza, D. Maccagnola, V. Guigue, E. Messina, and P. Gallinari. A latent representation model for sentiment analysis in heterogeneous social networks. In Carlos Canal and Akram Idani, editors, *SEFM Workshops*, volume 8938 of *Lecture Notes in Computer Science*, pages 201–213. Springer, 2014. ISBN 978-3-319-15200-4.

G. B. Orgaz, J. J. Jung, and D. Camacho. Social big data : Recent achievements and new challenges. *Information Fusion*, 28 :45–59, 2016.

G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435 :814–818, 2005.

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

S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos. Community detection in social media. *Data Mining and Knowledge Discovery*, 24(3) :515–554, May 2012. ISSN 1573-756X. doi : 10.1007/s10618-011-0224-z.

P. Pons and M. Latapy. Computing communities in large networks using random walks. In *International Symposium on Computer and Information Sciences*, pages 284–293. Springer, 2005.

F. A. Pozzi, D. Maccagnola, E. Fersini, and E. Messina. Enhance user-level sentiment analysis on microblogs with approval relations. In Matteo Baldoni, Cristina Baroglio, Guido Boella, and Roberto Micalizio, editors, *AI*IA*, volume 8249 of *Lecture Notes in Computer Science*, pages 133–144. Springer, 2013. ISBN 978-3-319-03523-9.

R. D. Putnam. *Democracies in Flux : The Evolution of Social Capital in Contemporary Society*. 2002.

L. Recalde. Detection of trending topic communities : Bridging content creators and distributors. In Ana Freire and Ricardo A. Baeza-Yates, editors, *FDIA*, Workshops in Computing. BCS, 2017.

P. K. Reddy, M. Kitsuregawa, P. Sreekanth, and S. S. Rao. A graph based approach to extract a neighborhood customer community for collaborative filtering. In Subhash Bhalla, editor, *DNIS*, volume 2544 of *Lecture Notes in Computer Science*, pages 188–200. Springer, 2002. ISBN 3-540-00264-2.

J. Reichardt and S. Bornholdt. Detecting fuzzy community structures in complex networks with a potts model. *Phys. Rev. Lett.*, 93(21) :218701, November 2004. doi : 10.1103/PhysRevLett.93.218701.

J. Reichardt and S. Bornholdt. Statistical mechanics of community detection. *Arxiv preprint cond-mat/0603718*, 2006.

J. Reichardt and D. R. White. Role models for complex networks, 2007. URL http://arxiv.org/abs/0708.0958. cite arxiv :0708.0958.

D. Richard. Getting social with recruitment. *Strategic HR Review*, 9(6) :11–15, 2010. doi : 10.1108/14754391011078063.

M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, pages 487–494, Arlington, Virginia, United States, 2004. AUAI Press. ISBN 0-9749039-0-6.

M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4) :1118–1123, 2008.

J. Scott. *Social Network Analysis : A Handbook*. Sage Publications, 2000.

A. M. Segura, J. de Lara, and J. S. Cuadrado. Rapid development of interactive applications based on online social networks. In Boualem Benatallah, Azer Bestavros, Yannis Manolopoulos, Athena Vakali, and Yanchun Zhang, editors, *WISE (2)*, volume 8787 of *Lecture Notes in Computer Science*, pages 505–520. Springer, 2014. ISBN 978-3-319-11745-4.

M. Sendi, M. N. Omri, and M. Abed. Discovery and tracking of temporal topics of interest based on belief-function and aging theories. *Journal of Ambient Intelligence and Humanized Computing*, Sep 2018. ISSN 1868-5145. doi : 10.1007/s12652-018-1050-6. URL https://doi.org/10.1007/s12652-018-1050-6.

G. Shmueli and O.R. Koppius. Predictive analytics in information systems research. *MIS Quarterly*, 35(3) :553–572, 2011.

G. Simmel. *Sociologie. Études sur les formes de la socialisation*. 1908.

G. Simmel. Sociologie et épistémologie. Paris, PUF, 1981, 1917.

I. Simonsen. Diffusion and networks : A powerful combination ! *Physica A : Statistical Mechanics and its Applications*, 357(2) :317–330, November 2005.

S. Staab and R. Studer, editors. *Handbook on Ontologies*, volume 10 of *International Handbooks on Information System*. Springer, 2004.

C. Stegehuis, R. Van der Hofstad, and J. Van Leeuwaarden. Epidemic spreading on complex networks with community structures. *CoRR*, abs/1611.06092, 2016.

Y. Sun and J. Han. *Mining Heterogeneous Information Networks : Principles and Methodologies*. Synthesis Lectures on Data Mining and Knowledge Discovery. Morgan and Claypool Publishers, 2012.

C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, and P. Li. User-level sentiment analysis incorporating social networks. In Chid Apté, Joydeep Ghosh, and Padhraic Smyth, editors, *KDD*, pages 1397–1405. ACM, 2011. ISBN 978-1-4503-0813-7.

J. Tang, J. Zhang, L. Yao, J. Li, Z. Li, and Z. Su. Arnetminer : extraction and mining of academic social networks. In Ying Li, Bing Liu 0001, and Sunita Sarawagi, editors, *KDD*, pages 990–998. ACM, 2008. ISBN 978-1-60558-193-4.

H. Tapio, M. K. Jussi, K. Kimmo, and S. Jari. Detecting modules in dense weighted networks with the potts method. *Journal of Statistical Mechanics : Theory and Experiment*, 2008(08) :P08007, 2008.

D. Tchuente, M. Canut, N. Jessel, A. Péninou, and F. Sèdes. Visualizing the relevance of social ties in user profile modeling. *Web Intelli. and Agent Sys.*, 10(2) :261–274, April 2012. ISSN 1570-1263.

J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 449–456, New York, NY, USA, 2005. ACM. ISBN 1-59593-034-5. doi : 10.1145/1076034.1076111. URL http://doi.acm.org/10.1145/1076034.1076111.

M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12) :2544–2558, 2010.

J.C. Valverde-Rebaza and A. Andrade Lopes. Link prediction in online social networks using group information. In *ICCSA (6)*, volume 8584 of *Lecture Notes in Computer Science*, pages 31–45. Springer, 2014. ISBN 978-3-319-09152-5.

S. Wasserman and K. Faust. *Social network analysis : Methods and applications*, volume 8. Cambridge university press, 1994.

M. Wick, K. Rohanimanesh, A. Culotta, and A. McCallum. Samplerank : Learning preferences from atomic gradients. In Neural Information Processing Systems (NIPS) Workshop on Advances in Ranking, editor, *booktitle*, 2009.

G. Xu, Tsoka S., and Papageorgiou L. G. Finding community structures in complex networks using mixed integer optimisation. *The European Physical Journal B*, 60(2) : 231–239, November 2007. ISSN 1434-6036. doi : 10.1140/epjb/e2007-00331-0.

Z. Xu, Y. Ke, Y. Wang, H. Cheng, and J. Cheng. A model-based approach to attributed graph clustering. In K. Selçuk Candan, Yi Chen, Richard T. Snodgrass, Luis Gravano, and Ariel Fuxman, editors, *SIGMOD Conference*, pages 505–516. ACM, 2012. ISBN 978-1-4503-1247-9.

B. Yang and J. Liu. Discovering global network communities based on local centralities. *TWEB*, 2(1) :9 :1–9 :32, 2008.

B. Yang, D. Liu, and J. Liu. Discovering communities from social networks : Methodologies and applications. In Borko Furht, editor, *Handbook of Social Network Technologies*, pages 331–346. Springer, 2010. ISBN 978-1-4419-7141-8.

J. Yang, J. McAuley, and J. Leskovec. Community detection in networks with node attributes. In *Data Mining (ICDM), 2013 IEEE 13th international conference on*, pages 1151–1156. IEEE, 2013.

D. Zhou, E. Manavoglu, J. Li, C. L. Giles, and H. Zha. Probabilistic models for discovering e-communities. In Les Carr, David De Roure, Arun Iyengar, Carole A. Goble, and Michael Dahlin, editors, *WWW*, pages 173–182. ACM, 2006. ISBN 1-59593-323-9.

H. Zhou. Distance, dissimilarity index, and network community structure. *Phys. Rev. E*, 67(6) :061901, June 2003. doi : 10.1103/PhysRevE.67.061901.

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