


République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université AMO de Bouira

Faculté des Sciences et des Sciences Appliquées

Département d'Informatique

Mémoire de Master2

en Informatique

Spécialité : *Ingénierie de Système d'Information et Logiciel*

Thème

Détection de communautés thématiques dans un réseau social académique

---

Encadré par

— Boussaadi Smail

Réalisé par

— Bachouche Zineb

— Hettal Houda

2019/2020

Remerciements

Avant tout Dieu merci de nous avoir donné la force et le courage de mener à terme le present travail.

Nos remerciements s'étendent à notre promoteur M^r **Smail Boussaadi** pour sa gentillesse et sa spontanéité . Nous avons eu le grand plaisir de travailler sous votre direction, nous avons trouvé auprès de vous le conseil et le guide qui nous avons reçu durant la réalisation de ce mémoire.

Nous tenons à remercier chacun des membres du jury pour nous avoir fait l'honneur d'examiner et d'évaluer notre travail et de l'enrichir par leurs propositions.

Nous ne saurons épuiser ces remerciements sans gratifier nos enseignants qui ont su nous donner une formation appréciable durant tout notre cursus.

a notre famille pour leur soutien et encouragement tout au long notre cursus universitaire .

Dedicaces

Je veux dédier cet humble travail à la lumière de ma vie : Chers parents, Puisse Allah avoir pitié d'eux Pour leur plus grand amour, Soutenir, encourager la patience et aider à continuer pendant mes années études.

ce modeste travail est également destiné à :

A mes chères sœurs pour leurs encouragements permanents, et leur soutien moral,

A mes chers frères, pour leur appui et leur encouragement,

A mon très cher fiancée (Merwan),Pour tout l'encouragement, le respect tu m'as offert

Pour mes neveux et nièces adorable .je vous souhaite un avenir plein de réussite

À mes proches copines Houda , Mona et Imane, qui m'ont toujours soutenue à chaque fois que j'en avais besoin et a tous mes amis

A toute ma famille pour leur soutien tout au long de mon parcours universitaire,

Que ce travail soit l'accomplissement de vos vœux tant allégués, et le fuit de votre soutien infaillible,

Merci d'être toujours là pour moi.

Bachouche Zineb

Dedicaces

Je dédie ce travail à mes chers parents qui m'ont soutenu dans ma vie Scolaire, tous les remerciements et appréciations ne suffisent pas pour vous remercier.

A mes chères sœurs imane, samah,khadija et ma petite sœur lina , mon cher frère Amine et mon fiancée Amine, je vous souhaite une vie plein de bonheur, courage et surtout la réussite .

A ma proche copine zineb

A toute ma famille HETTAL et MECIEL de loin et de près Et tous les professeurs que j'ai rencontrés dans mon parcours scolaire

Merci !

Hettal Houda

Résumé

Un réseau social académique comme Researchgate ou Mendeley, apparaissent comme un écosystème qui offre aux scientifiques des fonctionnalités telle que la publications , le partage des articles ou la collaboration. Ces réseaux spécialisés sont nommés complex network, littéralement réseaux complexes, dont l'analyse structurelle repose essentiellement sur la théorie des graphes. Une tâche très courante de l'analyse de ces graphes, qui a engendré une littérature prolifique ces dernières années, est la détection de communautés . Il s'agit de trouver dans un graphe des ensembles d'éléments interagissant plus particulièrement entre eux qu'avec le reste du graphe, formant ainsi les dénommées communautés d'intérêts.

Dans le cadre de de ce mémoire de master en informatique, nous présentons une nouvelle méthode de détection de communautés qui repose principalement sur la modélisation thématique pour extraire les sujets d'intérêts aux sein du réseau objet de notre étude.la méthode comporte deux étapes et exploite la notion de cliques adjacents dans un graphe, puis les communautés sont identifiées par maximisation de la modularité.

Mots clés : Réseaux sociaux, Réseaux sociaux académique,modélisation thématique, communautés d'intérêt, détection de communautés , modularité, cliques adjacents.

Abstract

An academic social network such as Researchgate or Mendeley appear as an ecosystem that offers scientists functions such as publication, article sharing or collaboration. These specialized networks are called complex networks, literally complex networks, the structural analysis of which is based primarily on graph theory. A very common task of analyzing these graphs, which has spawned a prolific literature in recent years, is the detection of communities. It is about finding in a graph sets of elements interacting more specifically with each other than with the rest of the graph, thus forming the so-called communities of interest.

As part of this master's thesis in computer science, we present a new method for detecting communities which is based mainly on thematic modeling to extract the subjects of

interest within the network object of our study. The method has two stages and exploits the notion of adjacent cliques in a graph, then communities are identified by maximizing modularity.

Keywords : Social networks, Academic social networks, thematic modeling, communities of interest, community detection, modularity, adjacent cliques.

Table des matières

Table des matières **i**

Table des figures **iv**

Liste des tableaux **v**

Liste des abréviations **vi**

Introduction générale **1**

1 Généralités sur les réseaux sociaux **4**

1.1 Introduction 4

1.2 Les Réseaux Sociaux 4

1.2.1 Bref Historique 5

1.2.2 Définition 5

1.2.3 Catégories des réseaux sociaux 6

1.2.4 Analyse des réseaux sociaux 7

1.3 Les Réseaux Sociaux Académique 7

1.3.1 Définition 7

1.3.2 Les réseaux sociaux académiques les plus populaires 7

1.3.3 Analyse des réseaux sociaux académiques 8

1.4 La théorie des graphes 8

1.4.1 Définition d'un Graphe 9

1.4.2 Mode de représentation d'un graphe 9

1.4.3 Type de graphe 11

1.4.4 Orientation de graphe 12
1.4.5 Degré 13
1.5 Communauté 13
1.5.1 Définition 13
1.5.2 Structure de communauté 13
1.6 Conclusion 14

2 Détection de communauté dans les réseaux sociaux 15
2.1 Introduction 15
2.2 Détection de communauté thématique 16
2.3 Communauté thématique 17
2.3.1 Définition 17
2.3.2 Evaluation de la qualité d'une communauté 17
2.4 Communauté disjointe 18
2.4.1 Formulation du problème de détection des communautés disjointes . 18
2.4.2 Approche Hiérarchique 19
2.4.3 Les approches à base d'optimisation de la modularité 24
2.4.4 L'approche à base du modèle 26
2.5 Communautés chevauchantes 28
2.5.1 Formulation du problème de détection des communautés chevauchantes 28
2.5.2 Approches basées sur les cliques 29
2.5.3 Extension locale et optimisation 31
2.5.4 Approches basées sur la propagation de labels 32
2.5.5 Avantages et inconvénients des algorithmes de détection communautés 33
2.6 Les profils thématiques 35
2.6.1 Définition 35
2.6.2 Construction du profil chercheur 35
2.6.3 Modélisation thématique 36
2.7 Conclusion 40

3 Approche de détection de communautés thématiques chevauchantes 41
3.1 Introduction 41

3.2 Description de l'approche proposée 41
3.3 Algorithme proposée 48
3.4 Conclusion 50

4 Implémentation et évaluation **51**
4.1 Introduction 51
4.2 Outils et environnements de développement 51
4.2.1 Environnement matériel 51
4.2.2 Environnement logiciel 52
4.2.3 Langage utilisé 53
4.3 L'ensemble des données exploitées 53
4.3.1 Expérimentation et résultats 57
4.4 Evaluation et discussion 62
4.4.1 Métrique d'évaluation 62
4.5 Conclusion 64

Conclusion générale **65**

Bibliographie **65**

Table des figures

Liste des tableaux

Liste des abréviations

RSA	Réseau Social Académique
LSA	Latent Semantic Analysis
PLSA	Probabilist Latent Semantic Analysis
LDA	Latent Dirichlet Analysis
CPM	Clique Percolation Method

Introduction générale

- **Contexte du travail**

Les réseaux sociaux académiques comme researchgate ou Academia sont devenus des espaces d'interactions entres chercheurs du monde entiers, et ceux-ci grâces a des avantages qu'ils offrent comme le partage des connaissances et la collaboration. Ces réseaux sont décrits et modéliser par des graphes, un outil du domaine des mathématiques qui facilite l'étude et la compréhension de leur structure topologiques. Dans un graphe, un chercheur est représenté par un nœud et les interactions sociales sont représentés par des liens.

En générale la densité des liens du graphe représentant un réseau n'est pas homogène et varie d'une zone a une autre, ce qui indique l'existence de groupe de nœuds fortement connectés entre eux mais faiblement connectés aux autres nœuds du réseau. Cette densité peut être justifie par une tendance qu'ont les chercheurs à se lier avec d'autres aux caractéristiques similaires , ces zones sont appelées communautés d'intérêts ou communautés thématiques, la détection de communauté consiste à identifier et a extraire ces communautés. Dans la littérature scientifique on distingue deux types de communautés, communautés disjointes et communautés chevauchantes. Dans les communautés disjointes, chaque nœud ne peut appartenir qu'à une seule communauté, mais dans le cas de communautés chevauchantes, ils peuvent appartenir à plusieurs communautés.

Le concept de détection de communautés est lié d'une certaine façon au partitionnement de graphes, bien qu'il soit différent. Dans le cas de partitionnement de graphe, le nombre de groupes et la taille approximative de ces groupes sont connus d'avance et le partitionnement consiste à diviser le réseau en plusieurs sous graphes de nombre connu

et de taille presque identique. Dans le cas de détection de communautés, le nombre de communautés existantes dans le réseau est inconnu et leurs dimensions diffère d'une communauté a l'autre. Les applications de la détection de communautés sont nombreuses. Nous citons à titre d'exemple la recommandation de collaboration entre chercheur , le partage de connaissance entre les membres d'une même communauté, la détection de spam dans les courriers électroniques.

En général, dans tous les réseaux sociaux, une personne peut appartenir à plusieurs communautés différentes, et dans le contexte d'un réseau académique un chercheur peut avoir un ou plusieurs domaines de recherche et peut de ce fait appartenir à plusieurs sous-groupes, indiquant un chevauchement entre les communautés d'intérêts. Donc, pour les réseaux sociaux académiques, la détection de communautés chevauchantes doit être plus considérée par rapport à la détection de communautés disjointes.

•Problématique

La découverte de structure communautaire présente dans un réseau académique peut aider a comprendre et a visualiser la structure topologique du réseau. la découvertes de ces communautés thématiques, nous permet de déterminer le rôle et l'importance de chaque individu au sein de sa communautés et dans le réseau globale. Le but de la détection de communauté est de trouver une partition du réseau en ensembles de nœuds qui forment la structure de communautés. Chaque communauté représente un sous-groupe de nœuds qui sont fortement connectés plus qu'ailleurs dans le réseau.

• Objectifs attendus

L'objectif de notre travail, présenté dans ce mémoire est le développement d'une approche de détection de communautés d'intérêts dans le contexte d'un réseau académique. Pour cela nous avons défini une méthode qui fonctionne principalement en deux phases. Durant la première phase nous modélisons notre réseau objet de l'étude par un graphe pondéré sur la base de similarité sur les sujets (topics). Durant la seconde phase nous appliquons un algorithme de découvertes de communautés chevauchantes : Percolation de cliques, qui extrait d'abord les K-cliques adjacents dans notre graphe puis nous formons les communautés en question.

Pour évaluer les performances de notre approche nous utilisons la Modularité de Newman , la modularité eune métrique qui indique la qualité de partitionnement dans un graphe.

- **Organisation du manuscrit**

Ce mémoire est structuré en quatre chapitres comme suit :

Le premier chapitre intitulé «**Généralités sur les réseaux sociaux**» , est consacré à des notions de base (définitions, ...) relatifs aux réseaux sociaux et réseaux sociaux académique et on finalise ce chapitre par quelque concept fondamentaux relatifs aux communautés

Le second chapitre nommé «**détection de communauté dans les réseaux sociaux :Etat de l'art** » présente un état de l'art des approches et des méthodes existantes pour détecter des communautés disjointes et chevauchantes

Le troisième chapitre «**Approche de détection de communautés thématiques chevauchantes**» nous avons proposé un algorithme pour détecter les communautés thématique dans les réseaux sociaux académique.

Le dernier chapitre,nous avons présenté l'étude de cas et évaluer les performances de l'algorithme proposé.

Enfin, nous terminerons ce mémoire par une conclusion générale et quelques perspectives qu'on souhaite accomplir prochainement.

Chapitre 1

Généralités sur les réseaux sociaux

1.1 Introduction

Environ 3 milliards de personnes dans le monde utilisent les réseaux sociaux donc l'équivalent de 40 %
de la population mondiale .[1] Nous passons une moyenne de 2h par jour de navigation sur ses réseaux sociaux,[1] ces réseaux sont omniprésents depuis l'apparition d'Internet. Ils permettent aux différents utilisateurs d'interagir en communauté et de se regrouper selon des critères qui leur sont importants. Ces réseaux sociaux sont de différents types. Certains sont connus de tous (Facebook , Twitter , LinkedIn) et comptent des milliards de membres. D'autres sont moins connus qui exploite beaucoup plus par les chercheurs scientifique qui on les appelés les réseaux sociaux académique(Rearchget,Académia) . Dans ce chapitre, nous allons présenter les Online Social Network (OSNs), leur définition,leur intérêt et leurs enjeux. et les réseaux sociaux académique

1.2 Les Réseaux Sociaux

De nos jours, les Réseaux Sociaux (RS) sont devenus une partie très importante et omniprésente de notre vie quotidienne. 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.

1.2.1 Bref Historique

Les réseaux sociaux en ligne sont apparus en 2002 avec le site américain Friendster [2], c'est le premier qui à utiliser le principe du cercle d'amis en ligne. Puis vint Myspace [3], crée en 2003, qui dépassa Friendster. A l'origine, l'objectif de Myspace était de permettre à des musiciens de proposer certains de leurs morceaux de musique à l'écoute et de construire leur réseau en devenant amis avec d'autres membres. Ce site est rapidement devenu populaire auprès des jeunes qui l'adoptèrent pour rester en contact avec leurs amis et s'en faire de nouveaux, l'argument musical étant relégué au second plan.

Facebook est né à (Harvard) le 4 février 2004. Au début, seuls les étudiants de l'université pouvaient s'inscrire. On a ensuite ajouté quelques autres universités américaines et canadiennes.C'est en 2006 que le site s'est ouvert à tous. Son nom vient (évidemment) du mot anglais « album photo ».Le nom au début... Thefacebook.com.Le cas de Facebook marque un tournant dans la démocratisation des réseaux sociaux sur Internet. Pour beaucoup, ce fut une porte d'entrée vers l'univers du (Web 2.0) et des réseaux sociaux. Premier réseau social au monde en 2008, il comptait 350 millions de membre en 2009 et en compte 500 millions en Juillet 2010.[4],[5]

Twitter est apparu en 2006 . Cette plateforme repose sur le principe du (microblogging) :les messages postés par les utilisateurs sont limités à 140 caractères. Très populaire aux Etats-Unis ou de nombreuses

personnalités y ont un compte... le site connaît néanmoins une très forte croissance, supérieure à celle de Facebook. D'après (comScore7) (société américaine spécialisée dans l'analyse Web), le réseau aurait enregistré une croissance annuelle de 109 en termes de visiteurs uniques entre Juin 2009 et Juin 2010 . De nombreux internautes s'en servent comme une véritable source d'information, notamment pour faire de la veille.[5]

1.2.2 Définition

Dans la littérature, il existe plusieurs définitions de réseaux sociaux, nous présentons quelques-unes :

Définition 1

Le terme de réseaux sociaux désigne généralement l'ensemble des sites internet permettant de se constituer un réseau d'amis ou de connaissances professionnelles et fournissant à leurs membres des outils et interfaces d'interactions, de présentation et de communication.

Définition 2

Un réseau social est constitué à la fois par un ensemble de personnes liées entre elles et par la force de ces liens. On peut aussi dire qu'un réseau social est un ensemble d'individus liés entre eux par des liens caractérisés par un degré de familiarité variable qui va de simple connaissance aux liens familiaux les plus étroits.[6]

Définition 3

Selon Wasserman et Faust, auteurs de" Social Network Analysis : Methods and Applications" publié en 1994,un réseau social est un ensemble de relations entre des entités sociales (individus). Les contacts entre ces individus peuvent être, par exemple, des relations de collaboration, d'amitié, ou des citations bibliographiques. Ces ressources sont donc aussi bien formelles qu'informelles, matérielles qu'immatérielles.[7]

1.2.3 Catégories des réseaux sociaux

Le tableau ci dessus va représenter la classification par catégorie du réseaux sociaux :

Catégorie	**Exemples**
Les réseaux sociaux généralistes	facebook , twitter , google+ ,snapchat
Les réseaux sociaux professionnels	linkedin ,viadéo
Les réseaux sociaux communautaires	Reddit ,
Les réseaux sociaux "visuels"	Instagram,Pinterest
Les réseaux sociaux de vidéo	Youtube,Twitch,
Les réseaux sociaux de blogging	Medium, Tumblr

TABLE 1.1 – Catégories des réseaux sociaux

1.2.4 Analyse des réseaux sociaux

Définition

L'analyse des réseaux sociaux est avant tout une boite a outils permettant de visualiser,et modéliser les relations sociales. Elle est fondée sur une approche structurelle des relations entre membres d'un milieu social organisé. Elle s'attache a décrire les interdépendances entre acteurs et permet une simplification de leur représentation, a juste titre, Social Networking Sites que nous pourrions traduire en français par (Résautage) 1. En général c'est une représentation graphique ou les acteurs sont représentés par des et nœuds les relations entre ces acteurs sont représentées par des arcs. De ce fait, elle repose sur des visualisations graphiques issues d'algorithmes permettant de calculer des degrés de force ou de densité entre les différents acteurs d'un réseau. [8]

1.3 Les Réseaux Sociaux Académique

1.3.1 Définition

Les réseaux sociaux académiques (RSA) ou de recherche permettent de lister les publications des chercheurs, suives l'activité scientifique et collaborer les chercheurs qui travail dans le même domaine d'intérêt .Tous les contenus sont accessibles librement mais l'accès demeure restreint(accès par login). Or, l'open access [1]. [9]

1.3.2 Les réseaux sociaux académiques les plus populaires

Comme les réseaux sociaux académiques sont très utilises par les chercheurs et les institutions de recherche alors , ces Trois réseaux sont donc particulièrement connus et utilisés, parmi toute la flopée qui s'est depuis développée :

Researchgate

Ouvert en 2008 ,c'est un réseau social pour les chercheurs,orienté dans le domaine des sciences, techniques et médecine. Permettant de partager des documents scientifique ,son

1. un mode de diffusion des articles de recherche sous forme numérique, gratuite et dans le respect du droit d'auteur.

moteur de recherche sémantique interroge de nombreuses bases de données,il est classé le troisième base de données d'article après Google Scholar. [10]

Academia

Lancé en 2008 à USA , c'est actuellement le réseau le plus important au monde (41 millions d'inscrits en 2017). Orienté (SHS), il est centré sur le partage de documents selon une démarche de (peer-review) (post-publication). Détail important : l'extension « .edu » ne renvoie pas à un établissement d'enseignement supérieur à but non lucratif, mais a été déposée avant la régulation des noms de domaine en « .edu »,il est toujours disponible gratuitement.[10]

MyScienceWork

L'un des derniers venus, puisqu' ouvert en 2010, et il est français ! Il se veut multidisciplinaire mais quand même plutôt sciences « dures » pour le moment... Sa bibliothèque est constituée d'articles, de thèses et de rapports de recherche en Open Access moissonnés dans une vingtaine de bases de données. [10]

1.3.3 Analyse des réseaux sociaux académiques

L'analyse des réseaux sociaux académiques est décrite comme une nouvelle méthodologie d'étude de communautés académiques, permettant de visualiser et modéliser les relations sociales comme des nœuds (les individus...) et des liens (relations entre ces nœuds). on utilisant des métriques pour aider également à comprendre plus sur les réseaux sociaux académiques :

1.4 La théorie des graphes

La théorie des graphes est un ensemble d'outils puissants pour la représentation graphique ,la théorie des graphes n'est pas limité pour le domaine mathématique ou algébrique plutôt , mais on peut l'utiliser comme outils de modélisation en l'informatique est dans de nombreux autres domaines, on définir qu'est ce que c'est un graphe et leur types dans ce partie.

1.4.1 Définition d'un Graphe

Un graphe G noté G= (V , E) se définit un ensemble de sommets (en anglais Vertices) V{v1,v2,...,vi} ,et un ensemble d'arcs (edge) Ee1,e2,...,ei, où un arc (arête) relie un couple de sommets de V tel que : V != {ϕ}ET E != {ϕ} .

Définitions et terminologies

- Si une arête xy relie les deux sommets (x,y) ,Alors on dit que x,y sont des **extrémités** de l'arête xy .
- S un ensemble de sommets tel que Sx , y,on dit que x et y sont **adjacents**, s'il existe une arête xy les reliant.
- le sommet x d'une arête xy est dite **l'origine** de l'arête et la composante y est dite **la destination** de l'arête.
- une arête de la forme xx s'appelle une **boucle.**
- une arête est **incidente** à un sommet ou le contraire ,si le sommet est une extrémités de l'arête.
- Un graphe se définit par son **ordre** soit le nombre de sommets, et par sa **taille** définie par le nombre de liens. [11]

1.4.2 Mode de représentation d'un graphe

C'est naturel de intéresser aux différentes manières de les représenter des graphe. Il existe une infinité de représentations d'un graphe selon la nature des traitement l'on souhaite appliquer aux graphes ,On donne les trois grands modes :

1.La représentation matricielle

Matrice d'incidence : On représente un graphe G= (V , E) par une matrice d'incidence M est une matrice (n * m) avec n se définit les lignes où chaque ligne représente un sommet,Et m se définit les colonnes où chaque colonne représente un arc .Un élément i,α c'est l'intersection entre un ligne Ni et la colonne mα , Ce dernier indique le type de relation entre un arc et un sommet .

1 signifie que l'arc α admet un sommet i comme extrémité initiale.

-1 signifie que l'arc α admet un sommet i comme extrémité terminale , et 0 sinon.

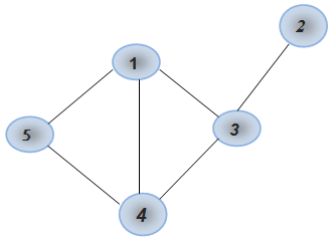

FIGURE 1.1 – Présentation non planaire du graphe G (des arêtes se croisent)

Matrice d'adjacence : On représente un graphe G= (V , E) par une matrice d'adjacence M (n*n) ,tel que M est une matrice booléenne avec chaque ligne et chaque colonne sont associées aux sommet du graphe ,chaque élément de M représente l'existence d'un arc entre deux sommet.

1 signifie que il existe un arc entre deux sommet adjacent (i, j),Et 0 sinon.

Si le graphe non orientés,on obtenu un matrice symétrique par rapport à se diagonal .

Matrice des degrés : On représente un graphe par une matrice des degrés M, est une matrice carrée diagonale qui contiens le degré ou le poids de chaque noeud du graphe a la place de diagonale.

Matrice Laplaciennes : La Matrice Laplacienne M d'un graphe G est la différence entre sa matrice des degrés et sa matrice d'adjacence.

2.La représentation des listes

Liste d'adjacence : peut être représenté un graphe par liste d'adjacence , représente la liste des successeurs et prédécesseurs de chaque sommet ,Ce type la plus simple et la plus économique.

3.Le graphe proprement dit

Les graphes tirent leur nom du fait qu'on peut les représenter par des dessins . À chaque sommet de G, on fait correspondre un point distinct du plan et on relie les points correspondant aux extrémités de chaque arête [12].

Remarque : Un graphe G = (V, E) est dit planaire lorsqu'on peut représenter les sommets de ce graphe comme des points du plan réel Vi = (x_i , y_i) et les arêtes < v_i , v_j > comme des lignes entre ces points qui ne se croisent jamais.[13]

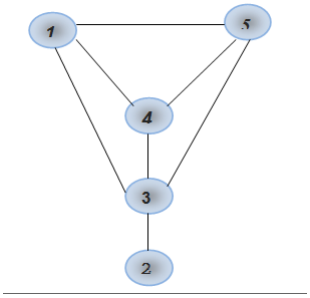

FIGURE 1.2 – Présentation planaire du graphe G

1.4.3 Type de graphe

On définir ici quelques type de graphes (orientés,non-orientés) :

Graphes isolés : Un graphe isolé est un graphe de n sommets qu'il n'existe aucune arête les reliant.

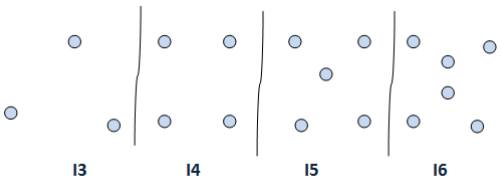

FIGURE 1.3 – Graphe isolé

Graphe est simple : Un graphe est simple s'il n'y a pas de boucle sur même sommet ou bien des arêtes multiple. On peut imaginer des graphes avec une arête qui relie un sommet à lui-même (une boucle), ou plusieurs arêtes reliant les deux mêmes sommets.On appellera ces graphes des multigraphes.[12]

Graphe connexe : Un graphe est dit connexe si y 'a un relation entre tous ses couple de sommet .

Graphe fortement connexe : Un graphe est dit fortement connexe s'il est possible à partir de n'importe quel sommet, de rejoindre tous les autres en suivant les arêtes. Un graphe non connexe se décompose en composantes connexes.[12]

Graphe est biparti : Un graphe est biparti (ou bipartite) s'il existe une partition de son ensemble de sommets en deux sous-ensembles X et Y telle que chaque arête ait une extrémité dans X et l'autre dans Y. On définit le graphe biparti complet entre un

ensemble de n sommets et un ensemble à m sommets comme le graphe simple tel que chaque sommet du premier ensemble est relié à chaque sommet su deuxième ensemble. On le note Kn,m.[15]

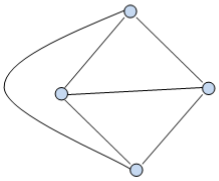

FIGURE 1.4 – Graphe biparti

1.4.4 Orientation de graphe

Un graphe est un ensemble fini de sommets reliés par des arêtes. Ces arêtes peuvent être orientées ou non, De plus une valeur peut être associée à chaque arête ou aux sommets.

Les graphes orientés

Un graphe orienté G est la donnée d'un couple G = (V, E) tel que :
∘ V est un ensemble fini de sommets.
∘ E un ensemble de couples ordonnés de sommets (x , y) ϵ V .
Un couple (x , y) est appelé un arc, et est représenté graphiquement par x $\rightarrow$ y .[14]
On dit que l'arc va de x à y . x est le sommet initial ou origine, et y le sommet terminal ou destinateur.

Les graphes non orientés

Un graphe non orienté G est la donnée d'un couple G = (V, E),Une paire{x,y} est appelée une arête, et est représentée graphiquement par x__y .On dit que les sommets x et y sont adjacents. L'ensemble des sommets adjacents au sommet x ϵ S,
est noté : Adj(x) ={x ϵ S, {x , y} ϵ A}.[14] Un graphe non orienté est complet si pour tous sommets x, y, il existe une arête entre x et y dans G.

1.4.5 Degré

Degré d'un sommet

◦ Le degré d'un sommet (s) d'un graphe non orienté, on note d(s), est le nombre des liens reliant ce sommet,ou bien le nombre de ses adjacence ,S'il existe une boucle sur un
◦ Le degré d'un sommet (s) d'un graphe orienté, on note d(s), est le nombre d'arcs entrant ,noté d+(s) plus le nombre d'arcs sortant ,noté d-(s)de ce sommet . d'où : d(s) = d+(s) + d-(s) sommet va compte deux.

Degré d'un graphe

Le degré d'un graphe est le degré maximum de tous ses sommets. Un graphe dont tous les sommets ont le même degré est dit régulier.[12]

1.5 Communauté

1.5.1 Définition

Une communauté est formée par un ensemble de sous-graphe dont les sommets sont plus liés entre eux qu'avec le reste du réseau. c'est sommets partagent le même sujet d'intérêt [16] on peut décrire une communauté de la manière suivante :C(V_cE_c)
Avec :
C : est une communauté de .G correspond au sous-graphe induit par $V_c$dans G.
V_c : est un ensemble de noeuds appartenant à V
E_c : est l'ensemble des liens appartenant à E

1.5.2 Structure de communauté

Structure de communauté ou bien (couverture de communauté) :est un groupe qui nous découvert dans un réseaux on peut la représenter comme suite : C=C_1,C_2...C_n
C : la structure de communauté
C_1,C_2...C_n : les communautés de réseau

1.6 Conclusion

Dans ce chapitre nous avons étudié les réseaux sociaux d'une manière générale ainsi que leur analyse, puis nous avons convergé vert les réseaux sociaux spécialiser dans les membres sont uniquement des scientifiques, c'est plateforme sont désigne par Réseaux Sociaux Académique.

Puis nous avons terminé ce chapitre par la théorie de graphe, un outil du domaine des mathématiques qui permet de présenter un réseau social par un graphe.

Chapitre 2

Détection de communauté dans les réseaux sociaux

2.1 Introduction

Dans ce chapitre nous présentant l'essentielle des travaux réaliser dans le domaine de la détection de communauté dans les réseaux sociaux les problème lie à la détection de communauté on fait l'objet des plusieurs travaux de recherche est qui apparaissent sous différent formulation qui ne posent pas nécessairement les mêmes méthodes de résolution

Nous commençons à présentons brièvement une revus de littérateur concernant la détection de communauté thématique dans les réseaux sociaux

Dans les deux sections suivantes nous aborder respectivement les deux catégories de communauté à savoir communauté disjoint et communauté chevauchent.

La dernière section sera consacrait aux profile thématique des utilisateurs d'une plateforme académique.

2.2 Détection de communauté thématique

Plusieurs travaux proposés pour la détection des communautés d'intérêt dans les Réseaux Sociaux ont été publiées. La plupart de méthode est basées sur les types des algorithmes de détection des communautés et leurs principes méthodologiques. Notons que, aujourd'hui, détection de communautés, partitionnement de graphe sont souvent utilisés indifféremment. Fortunato [17] Ces algorithmes sont principalement classés dans les catégories suivantes :

Algorithmes basés sur la modularité

La modularité et d'autres fonctions d'évaluation de la communauté caractérisent à quel point les groupes de nœuds du réseau ressemblent à des communautés. Algorithmes basés sur l'optimisation de la modularité tels que l'algorithme glouton de Newman [18] et sa version mise à jour par Clauset et al (Fast Greedy) joignent des sommets qui se traduisent par une plus grande augmentation de modularité. Après un processus itératif lorsque la modularité ne peut pas être maximisée de plus, le réseau est partitionné en communautés.

Une autre méthode d'optimisation de la modularité populaire est l'algorithme de Louvain qui trouve initialement de petites communautés en optimisant la modularité localement, puis en agrégeant les nœuds appartenant à la même communauté et en créant un réseau dont les nœuds représentent les communautés. Ce processus est itéré jusqu'à ce que la modularité maximale soit atteinte et qu'une hiérarchie de communautés soit produite [19].

Algorithme basé sur les vecteurs propres de la matrice de modularité

Cet algorithme de Newman (Leading Eigenvector) utilise eigenspectrum de la matrice de modularité. Initialement, cet algorithme crée initialement la matrice de modularité et trouve le vecteur propre de la plus grande valeur propre. Enfin, il étiquette les nœuds dans les communautés correspondantes en connaissant le signe des éléments dans le vecteur propre.[19]

Algorithmes de partitionnement 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.[17]

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. [20]

2.3 Communauté thématique

2.3.1 Définition

Il s'agit de structure regroupant un certain membre de chercheur qui partage les mêmes domaines de recherche ou bien les mêmes domaines d'expertises. Dans un réseau social académique les membres d'une même connaissance thématique ne sont pas nécessairement lie par une interaction explicite.

2.3.2 Evaluation de la qualité d'une communauté

Evaluation de la qualité d'une communauté est apparié pour évaluer les résultats de la communauté. Il existe des critères internes et d'autre externes pour évaluer la qualité d'une communauté , Parmi ces critère interne 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 « critères externes » se sert d'une partition comme référence de vérité de terrain. [21]

2.4 Communauté disjointe

Une communauté est définiecomme un ensemble d'entités au sein duquel il y a plus des relations internes qu'externes. Ondit que deux ensembles sont disjointes lorsque leurs intersection est un ensemble vide c à d (chaque nœud ne peut appartenir qu'à une seule communauté).la figure 2.1 illustre un exemple d'une communauté disjointe

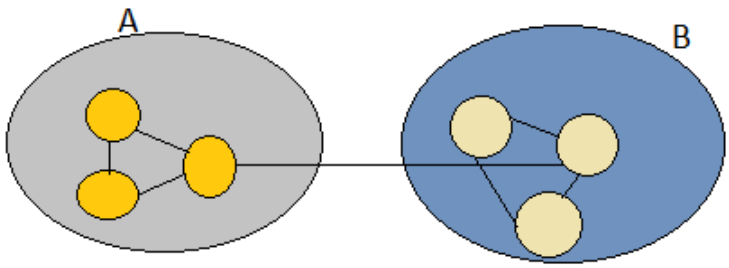

FIGURE 2.1 – Exemple d'une communauté disjointe .

2.4.1 Formulation du problème de détection des communautés disjointes

Considérons un réseau social représenté par un graphe G = (V, E). Le problème de détection de communauté dans sa forme générale consiste à trouver une partition P=$\{C_1,C_2,...C_r\}$ de l'ensemble des sommets V en r classes, avec $U_{k\epsilon\{1,2,...r\}}$,C_k = V , $C_k \bigcap C_l = \phi$, r $\geq$ k $\geq$ 1 et $C_k \neq \phi$, $\forall$ k ϵ {1,..r},de telle sorte que les sommets dans une communauté soient fortement connectés et faiblement avec le reste du graphe. [22]

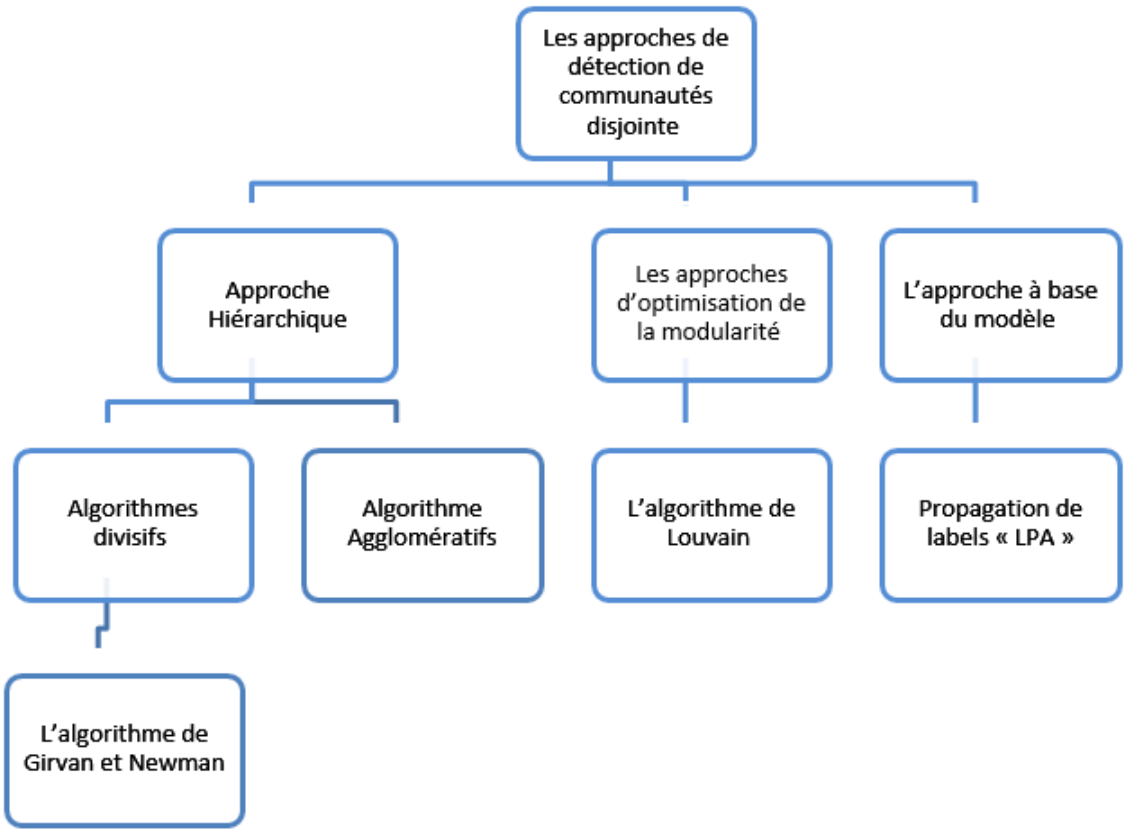

FIGURE 2.2 – Algorithmes de détection de communautés disjointes.

2.4.2 Approche Hiérarchique

Il existe traditionnellement deux types d'approches hiérarchiques

Algorithmes divisifs

L'idée principale dans les algorithmes divisifs est de trouver les liens entre les communautés et les supprimés afin de séparer les communautés. L'algorithme le plus populaire est celui proposé par Girvan et Newman [Newman and Girvan,2002] qui on le détail on suite. [23]

L'algorithme de Girvan et Newman :

C'est un algorithme proposé par [Girvan and Newman,2002], est basé sur la centralité d'intermédiarité (edge-betweeness-centrality). le but essentiel de l'algorithme étant alors de détecter des communautés partageant un intérêt commun. Cette mesure d'intérêt commun est basée sur la notion de plus court chemin dans un graphe.

Leurs principe est de supprimer progressivement l'arc qui a la plus forte « centralité des liens», puis refaire le même processus itérativement et à la fin on obtient un découpage des nœuds « chaque nœud considérer comme une communauté » , Les arcs entre nœuds sont ajoutés dans l'ordre inverse de celui dans lequel ils ont été supprimés. À chaque fois qu'un arc rejoint deux nœuds qui font partis de deux communautés différentes, tous les nœuds des deux communautés sont groupés dans une seule communauté. À chaque fois qu'un arc joint deux nœuds qui font partis d'une même communauté, l'ensemble des communautés reste inchangé [23]

Les mots clés que nous devons connaitre sont :la modularité , la centralité, le dendrogramme

La modularité : Le concept de la modularité à été proposé par Newman et Girvan en 2004 est une mesure pour comparer le nombre de liens dans un groupe de sommets à ce que l'on attendrait pour un graphe aléatoire similaire. elle est utilisée par plusieurs algorithmes comme fonction de qualité en l'optimisant, la déférence entre le nombre de liens présent dans Un module « communauté »,la valeur de la modularité est dans intervalle [-1,1], on la calculée comme suite [24]

$$Q = \frac{1}{2m} \sum_{ij} [A - \frac{k_i k_j}{2m}] \sigma(C_i, C_j)$$

m : le poids total des arcs du réseau

A : la matrice d'adjacence du réseau

K,k : le poids sommes i et j

σ : la fonction delta de Kronecker défini comme suit $\sigma(C_i, C_j)$ { 1 i=j 0 i≠j} .

La centralité d'intermédiarité (edge-betweeness-centrality) : est basée sur le calcule de plus court chemin et cherche les liens situés entre les communautés afin de les éliminées et de trouver les communautés, La centralité d'intermédiarité d'un Lien e_i est calculé comme suit :

$$ECB(e_i) = \sum_{i<j} \frac{NP_{V_jV_k(e_i)}}{NP_{V_jV_k}}$$

. $NP_{V_jV_k(e_i)}$: : est le nombre de plus courts chemins entre les nœuds V_j et V_k passant par le lien e_i

$NP_{V_jV_k}$:est le nombre de plus courts chemins entre les nœuds V_j et V_k

Dendrogramme : est un découpage du graphe en communautés

Les étapes principale de l'algorithme

- On calcule la centralité (betweenness) de chaque lien
- On enlève le lien de plus forte centralité
- On recommence jusqu'à ce qu'il n'y ait plus de lien
- les composantes connexes restantes sont les communautés
- On obtient une décomposition hiérarchique du réseau

.

L'algorithme de Girvan et Newman

Algorithm 1 Girvan et Newman

Input: un graphe non orienté Gu(V,E) d'ordre N

Output: C une structure de communautés

Begin

1 :initialisation :$T = \{V_1, V_2...V_n\} G' = G$;

2 :calculer la centralité d'intermédiarité pour chaque arête du graphe G' ;

3 :retiré du graphe G' l'arête em ayant la plus grande centralité d'intermédiarité

4 :identifier l'ensemble$C = \{C_1, C_2, C_3, ...C_l\}$ de toutes les composantes connexes de G'

5 :si T C$\notin$ lors rajouter C à T

6 :si $|E'|$ 0 $\neq$alors aller à 2

7 :retourner la structure de communauté C ayant la plus grande valeurs de modularité

End

La Figure (2.3)illustre le résultat d'exécution de l'algorithme Girven et Newman

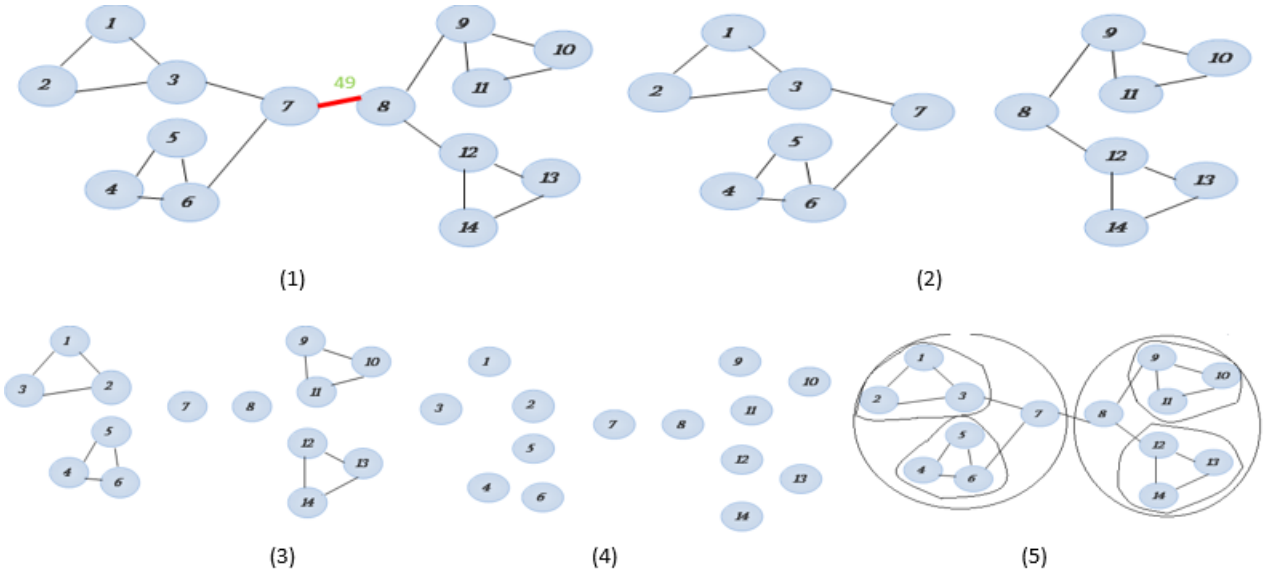

FIGURE 2.3 – Résultat d'exécution de l'algorithme Girvan et Newman .

Algorithme Agglomératif

L' algorithme de classification ascendante hiérarchique est basé sur le regroupement des nœuds ou des communautés itérativement dans des nouvelles communautés .Nous avons initialement n communautés (où n est le nombre de nœuds).

Le principe de l'algorithme est de calculer la distance entre les nœuds de la communauté et de fusionner les communautés les plus proches pour former une nouvelle communauté, à chaque étape, on recalcule toutes les distances entre les communautés et on fusionne deux communautés. Lorsqu'il n'y a qu'une seule communauté représentant le graphe entier, il n'existe plus de distance à calculer, les différentes étapes de ce processus peuvent être représentées par une forme arborescente appelée dendrogramme. Les feuilles sont les communautés avec un seul nœud et la racine représente le graphe entier.[25]

Les algorithmes agglomératifs hiérarchiques sont nombreux. Les premières études ont proposé les algorithmes SLINK (Single-LINK), CLINK (Complete-LINK) et ALINK (Average-LINK), tous trois basés sur le procédé SAHN (Sequential Agglomerative Hierarchical and Non-overlapping) [26]

Les étapes principale des algorithmes agglomératifs : [16]

1 :On définit une mesure de similarité, il existe deux mesure « indice de Jaccard » et

« similarité de cosinus »

∘ Indice de Jaccard : $Jaccard(V_i, V_j) = \frac{|N_i \bigcap N_j|}{|N_i \bigcup N_j|}$

∘ similarité de cosinus $cosinus(V_i, V_j) = \frac{|N_i \bigcap N_j|}{\sqrt{|N_i|.|N_j|}}$

Pour traiter le graphe qui est illustré dans la figure(2.4) on a utilisé la mesure de similarité de cosinus

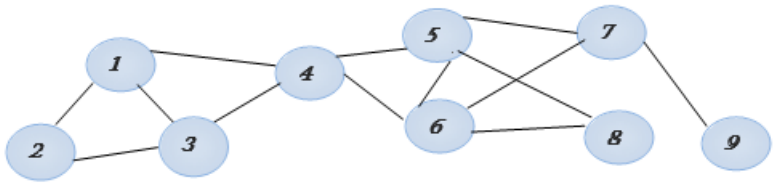

FIGURE 2.4 – Graphe de réseau.

La similarité pour les nœuds « V4 , V5 » :

$cosinus(V_i, V_j) = \frac{1,3,5,6 \bigcap 4,6,7,8|}{\sqrt{|1,3,5,6|.|4,6,7,8|}} = \frac{1}{4} = 0.25$

.

2 :après le calcul de similarité pour chaque nœuds on reproduit une matrice de similarité voir la matrice ci dessus qui illustre par la figure(2.5)

	V1	V2	V3	V4	V5	V6	V7	V8	V9
V1	1	0.408	0.667	0.289	0.289	0.289	0.000	0.000	0.000
V2	0.408	1	0.408	0.707	0.000	0.000	0.000	0.000	0.000
V3	0.667	0.408	1	0.289	0.289	0.289	0.000	0.000	0.000
V4	0.289	0.707	0.289	1	0.250	0.250	0.500	0.577	0.000
V5	0.289	0.000	0.289	0.250	1	0.750	0.500	0.577	0.500
V6	0.289	0.000	0.289	0.250	0.750	1	0.500	0.577	0.500
V7	0.000	0.000	0.000	0.500	0.500	0.500	1	0.577	0.000
V8	0.000	0.000	0.000	0.577	0.577	0.577	0.577	1	0.577
V9	0.000	0.000	0.000	0.000	0.500	0.500	0.000	0.577	1

FIGURE 2.5 – Matrice de similarité .

3 : on applique une méthode de CHA « classification hiérarchique ascendant » on obtenir un dendrogramme comme il est illustré dans la Figure (2.6)

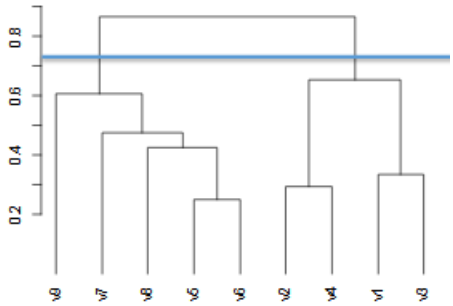

FIGURE 2.6 – Dendrogramme de réseau

D'après le dendrogramme on a deux communautés voir la Figure (2.7)

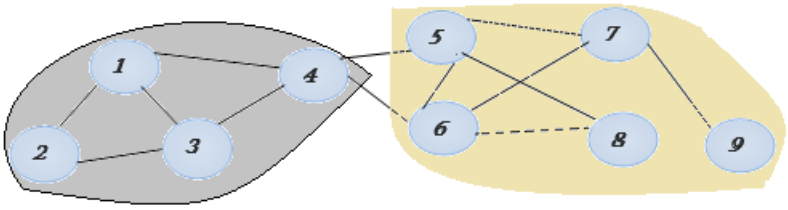

FIGURE 2.7 – Communautés de réseau .

2.4.3 Les approches à base d'optimisation de la modularité

Est une procédure mathématique qui permet d'obtenir les minimums (ou maximums) d'une fonction réelle f (que l'on appelle fonction objective). Ce type d'algorithme consiste à définir une fonction objective dont la valeur varie selon les communautés dégagées. La fonction est maximale pour la meilleure Structure de communauté.[22], [25]

L'approche d'optimisation est apparue pour corriger le problème de complexité de l'approche hiérarchique, le but essentiel de cette approche est de minimiser le temps de partitionnement .Des réseaux en communautés et de maximiser la modularité dans un délai raisonnable.il existe actuellement plusieurs algorithmes d'optimisation, qui sont utilisés selon le type de problème à résoudre, Un exemple de cette classe d'approches est l'algorithme Louvain

L'algorithme de Louvain

L'algorithme de Louvain a été introduit par Bland et al on 2008, le principe initial de l'algorithme est chaque sommet est affecté à une communauté différente des autres ,il est décomposé on deux phase déférentes :

• La première phase affectation des nœuds : Pour chaque nœud i on évalue le gain de la modularité si on le déplace dans la communauté de son voisin j. On affecte i à la communauté du voisin qui maximise le gain de la modularité, Si aucun gain n'est possible le nœud i reste dans sa communauté. [27]

• La deuxième phase la compression : la première étape dans cette phase, on remplace chaque communauté dans le graphe obtenu de la première phase par un seule nœud , s'il existe un lien entre un nœud de la communauté représentée par C_x et un nœud de la communauté représentée par C_yLe poids de lien entre deux communautés est égale à la somme des poids des liens reliant des nœuds de deux communautés.[27]

La Figure (2.8) illustre l'exécution de ces deux phases dans un double itération. L'algorithme s'arrête s'il a y plus de possibilité de réaffectation des nœuds ou si un maximum de modularité soit atteint

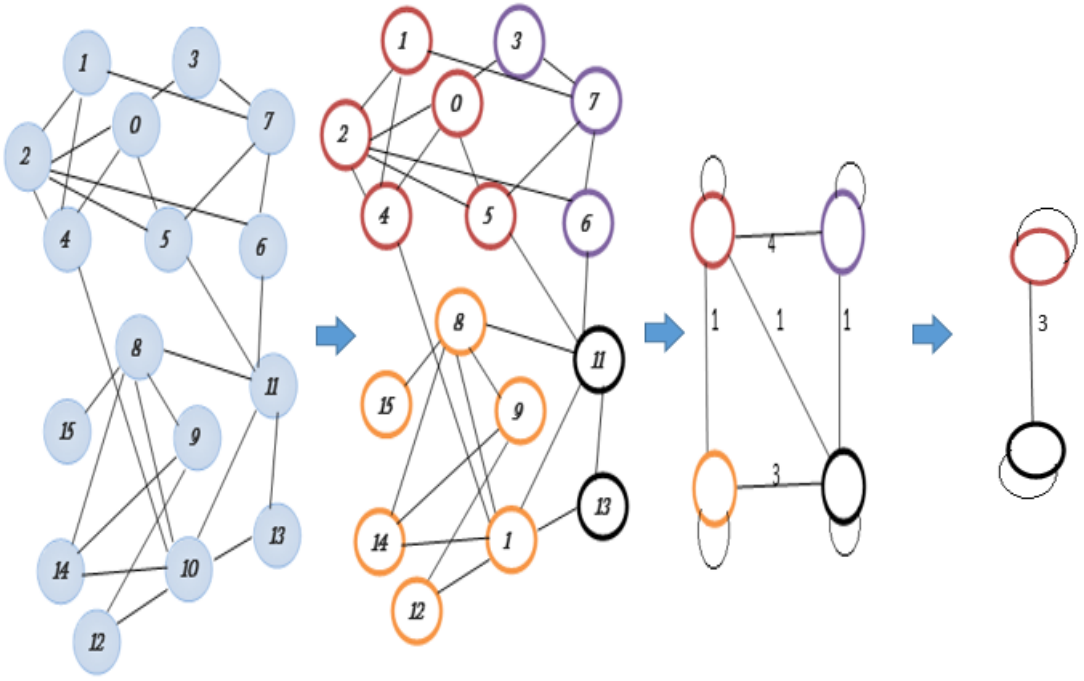

FIGURE 2.8 – Résultat d'exécution de l'algorithme de Louvain .

2.4.4 L'approche à base du modèle

Les algorithmes à base du modèle sont des algorithmes de classification non supervisée utilisant des méthodes à base de prototypes exprimés dans un formalisme de modèles. Pour chaque type de données, un modèle d'apprentissage adapté à la nature des données traitées est proposé, l'algorithme label propagation noté **LPA** c'est l'algorithme le plus utilisé pour traiter les graphes [25]

Propagation de labels « LPA »

L'algorithme de propagation de labels (Label Propagation Algorithme LPA) à été introduit pour la première fois par Raghavan et al. on 2007, C'est un algorithme itératif où à chaque itération un nœud envoi son label à ses voisins directs, Chaque nœud appartient à la communauté qui contient le plus grand nombre de ses voisins. Au début, chaque nœud est initialisé avec un label. Et à chaque itération, chaque nœud adopte le label majoritaire de ses voisins. Dans le cas d'égalité (cas de plus d'un label maximale), il choisit un au hasard.[27]

Les instructions de l'algorithmes de Propagation de labels :

Algorithm 2 l'algorithme LPA

Input: un graphe non orienté Gu(V,E) d'ordre N

Output: C une structure de communautés

1- t =0 ; Initialiser les nœuds avec un label unique, $\forall$ i $\epsilon N : c_i = l_i$,

2- t=1

3- Pour chaque nœud i trouver la communauté de iC_ien fonction des étiquettes des voisins de i

4- Si la structure de communauté devient stable($\forall$i ϵ N : $C_i(t) = C_i(t-1)$) Alors algorithme s'arrête

5- sinon t= t+1 et aller à (3).

.

On a un réseau de 7 nœuds , comme il est illustré dans la figure ce dessus , on applique l'algorithme LPA sur ce réseau la Figue 2.9 montre les étapes d'exécution et le résultat obtenu

Au début de l'algorithme chaque nœud du réseau est initialisé avec un label unique (a,b,c,d,e,f,g) , on choisi l'un des nœuds (le choix est aléatoire) par exemple on choisi le nœud 3 de la communauté c ,ses voisins appartient à des communautés différentes , donc ce nœud peut appartenir à l'une des communautés (a,b,d) , dans notre cas on associé le nœud 3 à la communauté b (Figure 2.9.(2)).

si le nombre de voisins d'un nœud qui appartient à la même communauté est plus que le nombre de voisins d'une autre communauté on associé le nœud à la communauté qui a plus grand des voisins on appliquant le même processus pour les autres nœuds , nous obtenons une structure de communauté avec deux communautés b et e pour le graphe entier (Figure 2.9.(6)). .

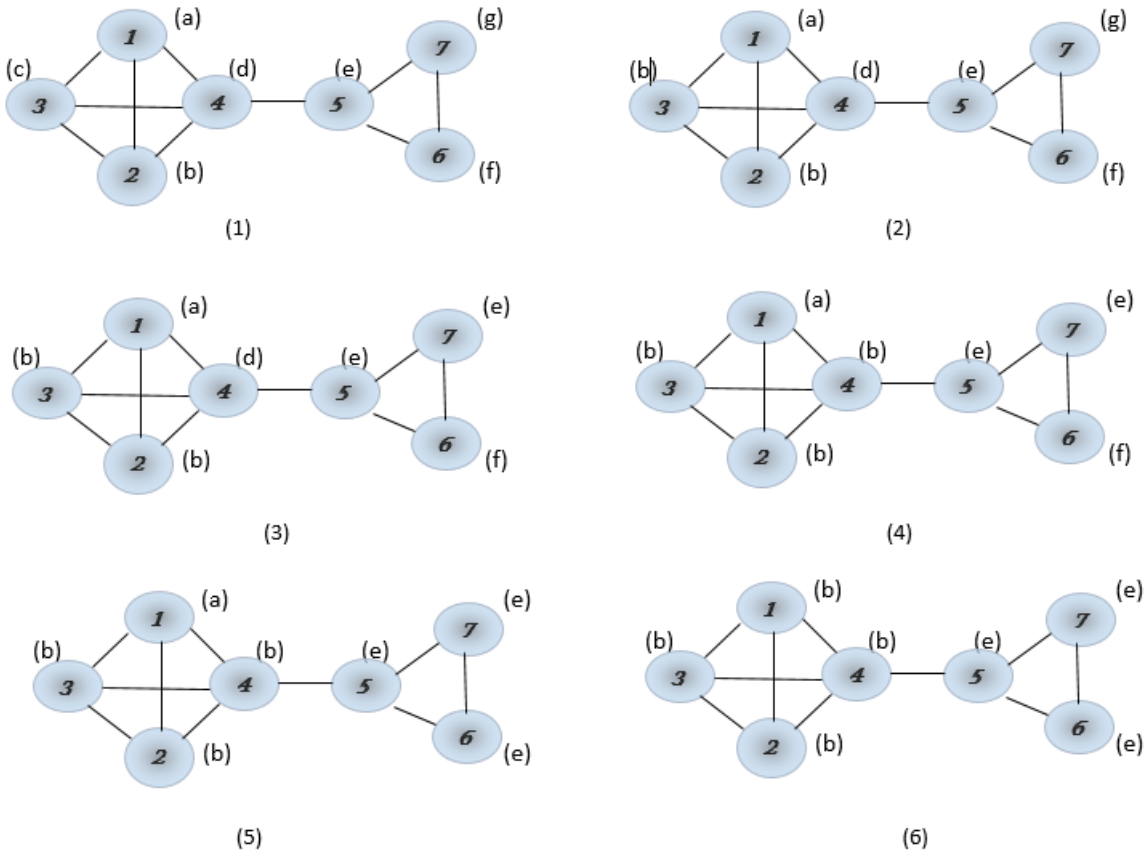

FIGURE 2.9 – Résultat d'exécution de l'algorithme LPA .

2.5 Communautés chevauchantes

Dans la section précédente nous a permis de trouver les méthodes pour la détection de communautés disjointes, C'est-à-dire le cas où chaque nœud ne peut appartenir qu'à une seule communauté, mais dans beaucoup de situations, dans la plupart des réseaux de terrain, il est possible de trouver un nœud dans un graphe qui peut appartenir à plusieurs communauté en même temps, on parle alors de communautés chevauchantes. Il existe des communautés chevauchantes dans un réseau sociaux académique, comme exemple, d'une chercheure à un intérêt multi-appartenance, qui publier dans un plusieurs domaine. Alors il appartenant au plusieurs communautés différentes donc c'est un chevauchement.

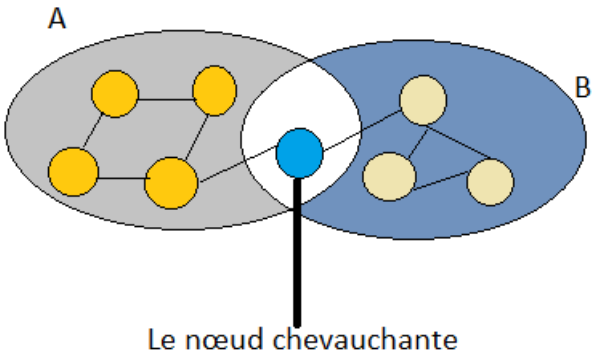

FIGURE 2.10 – Communautés chevauchantes.

2.5.1 Formulation du problème de détection des communautés chevauchantes

La détection de communauté chevauchante et difficile à résoudre lorsqu'un nœud peut appartenir à plusieurs communautés, le problème consiste à trouver un recouvrement de communauté [28], en considérant un graphe G = (V, E) où E est l'ensemble des liens. L'ensemble de couvertures trouvées est appelé $C = \{c_1, c_2, \ldots, c_k\}$ de l'ensemble V des sommets, tel que :

$I \epsilon \{1, 2, \ldots, k\}, c_i \subset V et c_i \neq \phi$;

$U_{i=1}^{k} C_i = V$

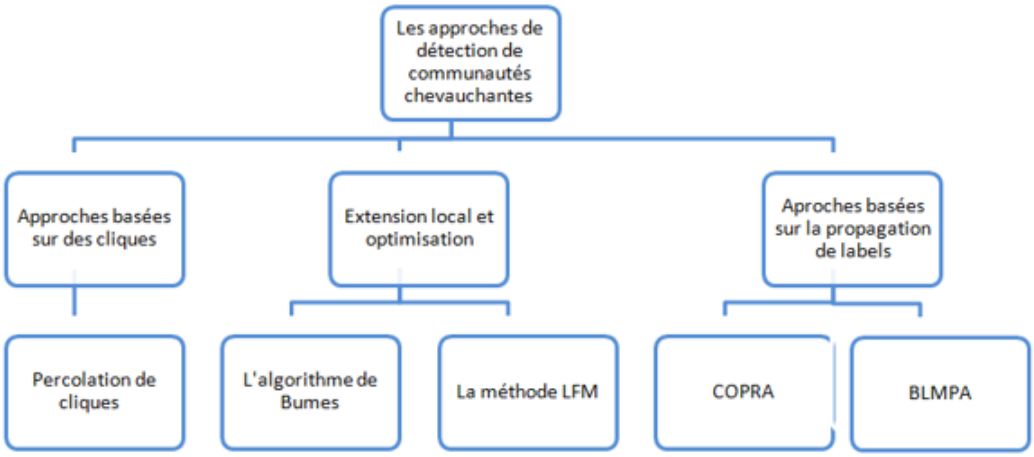

FIGURE 2.11 – Algorithmes de détection de communautés chevauchantes .

2.5.2 Approches basées sur les cliques

Parmi les approches de détection de communautés chevauchante basées sur l'utilisation de cliques, en trouve plusieurs méthodes « méthode de percolation, EAGLE ». Dans un graphe il est possible de former des communautés définies comme une chaîne de cliques adjacentes, et un sommet peut appartenir dans plusieurs cliques situées dans des communautés différentes ce qu'on appelle nœud chevauchant.

Percolation des cliques

Est un algorithme de détection de communautés chevauchantes , il basé sur l'utilisation des cliques, son principe est de former des communautés en percolant (adaptation du sens utilisé en physique) par la recherche de clique adjacente [29]. Cette approche nécessite d'abord la fixation du paramètre k ($k\epsilon\{1, 2, \ldots, 6\}$) pour donner de bon résultats), qui définit la taille des cliques (la taille des cliques est le nombre de sommet). Passant premièrement par l'identification de toutes les K- cliques présentes dans le réseau. Après les avoir identifiées, un nouveau graphe GC appelé clique-graph est construit de telle sorte que les sommets de celui-ci représentent les k-clique identifiées. Deux nœuds sont connectés si les K cliques les représentants partagent K-1 éléments [28]. La dernière étape présente l'identification des composantes connexes du graphe GC dont les cliques composent les communautés, un sommet compris dans plusieurs cliques situées dans des communautés différentes est alors un chevauchement entre les communautés.

Par exemple, le réseau représenté par la Figure (2.12) contient sept 3-cliques qui représentent les nœuds du nouveau graphe clique. Les sept 3-cliques sont : a : (1, 2,3) ; b : (1, 3,4) ; c :

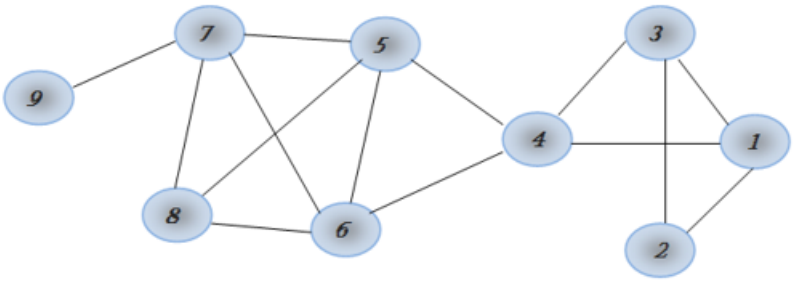

FIGURE 2.12 – Exemple de réseau.

(4,5,6) ; d : (5,6,7) ; e(5,6,8) ;f(5,7,8) ;g(5,7,8)
Si deux cliques partagent (k-1) nœuds (k=3) alors elles seront connectées avec un lien. Clique (1,2,3) et (1,3,4) partagent deux nœuds commun (1,3). Donc ces deux nœuds seront reliés. De la même manière d'autre cliques font également la connexion entre eux pour former le graphe des cliques de la Figure (2.13) .

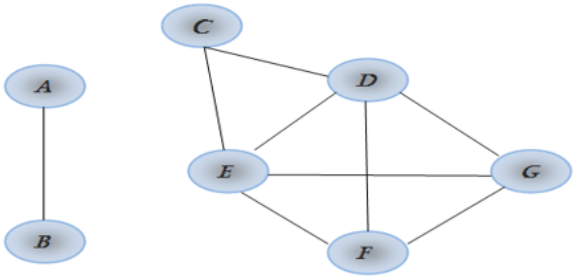

FIGURE 2.13 – Graphe des cliques.

Les composante connexe dans le nouveaux graphe sont (a, b) et (c ,d ,e ,f ,g). Les composantes connexes représentent les communautés.
c1 : (1, 2, 3,4)
c2 : (4,5, 6, 7,8)
Le nœud 4 est chevauchant entre les deux communautés.

2.5.3 Extension locale et optimisation

Les algorithmes utilisant l'extension locale et l'optimisation sont basés sur la croissance d'une communauté naturelle ou une communauté partielle [28]. La plupart d'entre eux s'appuyer sur une fonction de bénéfice local qui caractérise la qualité d'un groupe de nœud densément connectés.

L'algorithme de Baumes

Un des premiers algorithmes de détection de communauté chevauchante proposer par Baumes et al,ils sont Proposés un processus s'articule autour de deux algorithmes [30] : Algorithme RankRemoval (RaRe) : qui permet d'isoler un ensemble de noyaux de communautés pour la deuxième étape.

• La recherche des nœuds les plus importants.

• Ensuit, les retire du graphe de manière à ne conserver qu'un ensemble de petites composantes connexes, Ces composantes connexes donc deviennent les communautés.

• Les nœuds « importants » sont alors rajoutés tour à tour à ces composantes connexes pour finaliser les communautés et mettre en place les recouvrements.

Algorithme Itératif Scan (IS) : qui permet de construire une partition du graphe localement optimale au sens de la densité.

• Un nœud quelconque du graphe est ajouté ou supprimer à la communauté afin augmente la densité.

La fonction de densité est donnée comme suite :

$$f(c) = \frac{W_{in}^c}{(W_{in}^c + W_{out}^c)}$$

ou ($W_{in}^c et W_{out}^c$ représentent le poids total interne et externe de la communauté C.

La méthode LFM

LFM (Lancichinetti et al. 2009) élargit une communauté à partir d'un nœud de graine aléatoire pour former une communauté naturelle jusqu'à la fonction fitness.

$$f(c) = \frac{k_{in}^c}{(k_{in}^c + k_{out}^c)^\alpha}$$

où $k_{in}^c et k_{out}^c$ sont le degré interne et externe total de la communauté c, et α est le paramètre de résolution contrôlant la taille des communautés. Après avoir trouvé une

communauté, LFM dépend considérablement du paramètre de résolution α. La complexité de calcul pour une valeur α fixe est approximativement $O(n_c s^2)$.
où ,n_c est le nombre de communautés et s est la taille moyenne des communautés. La complexité la plus défavorable est O (n2).

2.5.4 Approches basées sur la propagation de labels

La propagation de labels LPA est une méthode qui a déjà été proposée pour détecter les communautés disjointes, Ces algorithmes nécessitent de parcourir tous les nœuds du graphe, et les partitionnements résultats dépendent beaucoup de l'ordre dans lequel on parcourt ces nœuds. Nous présentons deux algorithmes pour la détection de communautés chevauchantes à base de propagations de labels.

La méthode COPRA

L'algorithme COPRA posée par Gregory (2010), est une adaptation de la méthode propagation de labels aux cas avec recouvrement. Pour ce faire, il propose, de ne plus choisir seulement le label le plus courant chez ses voisins, mais de maintenir une liste des labels les plus courants dans son entourage. Un paramètre de l'algorithme fixe le nombre maximum de labels qu'un nœud peut retenir (sans quoi chaque label s'étendrait à l'infini). La limite de cet algorithme est que le choix de nœud à traiter est aléatoirement et avec une condition d'arrêt (non pas une mesure), et plusieurs résultats finaux peuvent être obtenus [31]. La Figure suivant montre le résultat d'exécution de COPRA avec $v = 2$

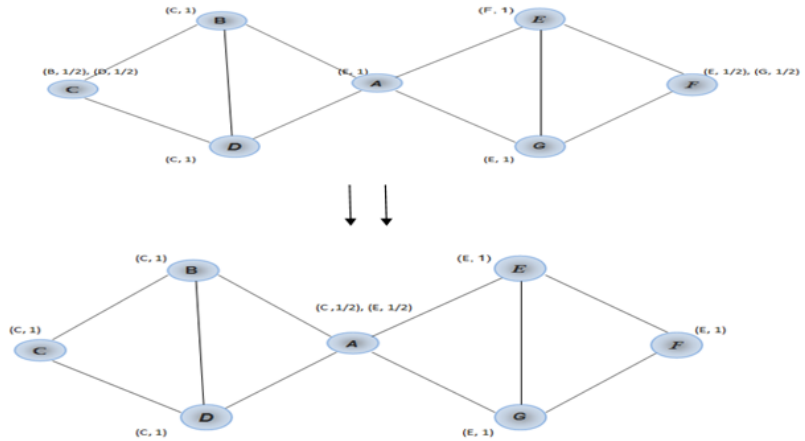

FIGURE 2.14 – Résultat d'exécution de l'algorithme Copra (extrait de Gregory (2010))

La méthode BMLPA

BMLPA pour "balanced multi-label propagation algorithme" (Wu et al. (2012)) est une amélioration de la méthode COPRA. Les auteurs proposent une stratégie de mise à jour, qui demande que les nœuds ayant le même label aient un coefficient d'appartenance vis-à-vis d'autres communautés. Cela a pour conséquence que certains nœuds puissent appartenir à plusieurs communautés. Les auteurs proposent également de générer des "cœurs bruts", qui sont utilisés pour l'initialisation des vecteurs pour la propagation de labels multiples. Cela permet d'aller plus vite et de stabiliser une partie du graphe lors de la propagation. Les résultats en termes de qualité de partionnement sont encourageants et permettent de Limi- ter le nombre de mauvaises propagations (sans toutefois les faire disparaitre) [32].

2.5.5 Avantages et inconvénients des algorithmes de détection communautés

Le tableau ce dessus montre Les avantages et les inconvénients des algorithmes de détection de communauté :

.

Algorithmes	Avantages	Inconvénients
Girvan et Newman	— L'algorithme GN est simple et facile à implémenter. — la connaissance préalable du nombre de communauté non requise.	— le temps de calcul trop élevé. — Il est inutilisable dans les réseaux sociaux qui contiennent des miliaires nœuds. — La complexité algorithmique est plus forte O(m2n), ou n le nombre de nœuds est m le nombre de liens. — Il est utilisable dans les petits graphes.
des l'algorithme agglomératifs.	— le dendrogramme produit est très utile pour comprendre les données	— la complexité temporelle peut entraîner des temps de calcul très longs, en comparaison avec des algorithmes efficaces
LPA	— la simplicité de compréhension. — la complexité temporelle linéaire.	— l'instabilité des résultats(produit de nombreuse solution).
Louvain	— calculer des communautés sur de très grands graphes .	— la sensibilité du résultat à l'ordre de traitement des sommets.
Cobra	— La communauté qui se chevauche peut être détectée	— C'est très incertain

Extension locale et optimisation	— Ciblage de niche efficace et précis — La communauté qui se chevauche peut être détectée, peut bien fonctionner également pour les et réseaux dynamiques.	— Les résultats peuvent être moins précis que les autres méthodes.

TABLE 2.1 – Les avantages et les inconvénients des algorithmes de détection de communautés.

2.6 Les profils thématiques

2.6.1 Définition

C'est une structure qui permet de modéliser et stocker des informations relatives à l'utilisateur et leur domaine Professional. [33]

2.6.2 Construction du profil chercheur

Indépendamment de son modèle de représentation, la construction du profil chercheur repose sur deux phases principales : la phase de collecte des sources d'informations et la phase d'exploitation de ces sources d'informations pour construire et représenter le profil chercheur avant son utilisation par des techniques de personnalisation d'information voir (Figure 2.15) [33]
le profil d'un chercheur est modéliser par un vecteur comme suit :
$\vec{V} = (topic0, topic1, topic2, ...topic_k)$
la contraint de modélisation d'un profil de chercheur est comme suite :

$$\sum_i^k p(Z_i/d_j) = 1$$

avec : d_j : document de chercheur j Z_i : un topic / tel que $i = 0, 1, 2, 3, ..., k$

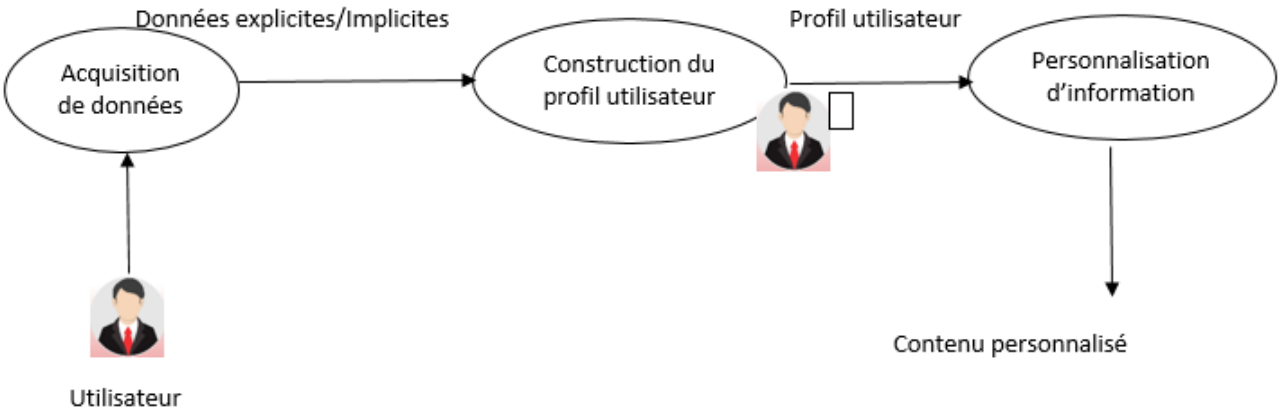

FIGURE 2.15 – Les phases de construction du profil utilisateur

2.6.3 Modélisation thématique

Définition de la modélisation des intérêts

La modélisation de sujets est l'un des outils que nous utilisons pour analyser les données textuelles de manière structurée, ordonnée et quantifiable. Au début du processus, l'analyste est confronté à une masse de documents non organisés. Après l'analyse des données (des données explicites de l'utilisateur ou des données implicite à partir des données collectées., On construire une liste représentant des intérêts de l'utilisateur(point de vue). [34]

Avantages et inconvénients de la modélisation des intérêts

le tableau suivant va montre les avantages et les inconvénients de la modélisation thématique :

Avantages	**Inconvénients**
beaucoup plus utiles que la simple fréquence des mots ou les approches basées sur un dictionnaire	les modèles de sujet ne produiront pas une classification très nuancée des textes
Les modèles de sujet tendent à produire les meilleurs résultats (lorsqu'ils sont appliqués à des textes qui ne sont pas trop courts)	
la modélisation de sujet développera la représentation vectoriel et introduire un nouveau terme qui est le sens de mot	

TABLE 2.2 – Avantages et inconvénients de la modélisation des intérêts

Modèles thématiques

- Analyse sémantique latente (LSA)

Également appelé LSI, cet algorithme construit un espace sémantique dans lequel les mots et les documents associés sont placés les uns à côté des autres. Il utilise SVD comme technique. [34]

- Analyse sémantique latente probabiliste (PLSA)

Également appelé modèle d'aspect, il s'agit d'un modèle générateur probabiliste. Il n'utilise pas SVD. Il examine la probabilité d'un sujet donné un document et la probabilité d'un mot donné un sujet. Ce sont des distributions multinomiales qui peuvent être entraînées avec l'algorithme EM. [34]

- Allocation latente de Dirichlet (LDA)

Il s'agit d'une approche bayésienne. Le document est modélisé comme un mélange fini de sujets. Chaque sujet est modélisé comme un mélange infini de probabilités de sujet. Les probabilités de sujet constituent la représentation d'un document. Le mélange de sujets est une distribution de Dirichlet. [34]

le principe de fonctionnement du LDA :

LDA fonctionne comme suit :

Premièrement, LDA exige que la recherche spécifie une valeur de k ou le nombre de sujets dans le corpus. Dans la pratique, il s'agit d'une décision très difficile - et conséquente -. Nous discuterons des procédures qui peuvent être utilisées pour identifier la valeur appropriée de k dans le scénario commun où l'on n'a pas a priori une forte théorie sur le nombre de thèmes latents qui pourraient exister dans un corpus ;

Chaque mot qui apparaît dans le corpus est assigné à l'aléatoire à l'un des k sujets. Cette affectation n'est techniquement pas aléatoire, car elle implique une distribution de Dirichlet qui utilise un simplexe de probabilité au lieu de nombres réels (cela signifie simplement que les nombres attribués à travers les k sujets totalisent 1).

Les affectations de sujet pour chaque mot sont mises à jour de manière itérative en mettant à jour la prévalence du mot parmi les k sujets, ainsi que la prévalence des sujets dans le document. Cette étape de LDA utilise la métrique terme fréquence-fréquence de document inverse discutée dans un didacticiel précédent. Les affectations de sujets sont mises à jour jusqu'à un seuil spécifié par l'utilisateur, ou lorsque les itérations commencent à avoir peu d'impact sur les probabilités attribuées à chaque mot du corpus.

LDA, et la plupart des autres formes de modélisation de sujet, produisent deux types de sortie. Tout d'abord, on peut identifier les mots les plus fréquemment associés à chacun des k sujets spécifiés par l'utilisateur. Deuxièmement, LDA produit la probabilité que chaque document du corpus soit également associé à chacun des k sujets spécifiés par l'utilisateur. Les chercheurs affectent souvent ensuite chaque document au sujet auquel il ressemble le plus ou définissent un seuil de probabilité pour définir le document comme contenant un ou plusieurs des k sujets. pour bien comprendre le fonctionnement du LDA,nous introduisons quelques notations :

- M : représente le corpus de document.
- N : représente les termes de corpus
- K : représente le nombre des topics associe au corpus
- α :est le paramètre du dirichelt prior sur les distributions de sujets par documents

— β :est le paramètre du dirichelt prior sur les distributions de mots par thème

— θ_d : est la distribution des sujets pour chaque document d

— Z_d : esZ_i, i=1...N , sont les topics associés à chaque mot W_i d'un document

— W_d :est le terme apparaissant dans le document d

la figure ce dessus c'est une représentation graphique du modèle LDA .

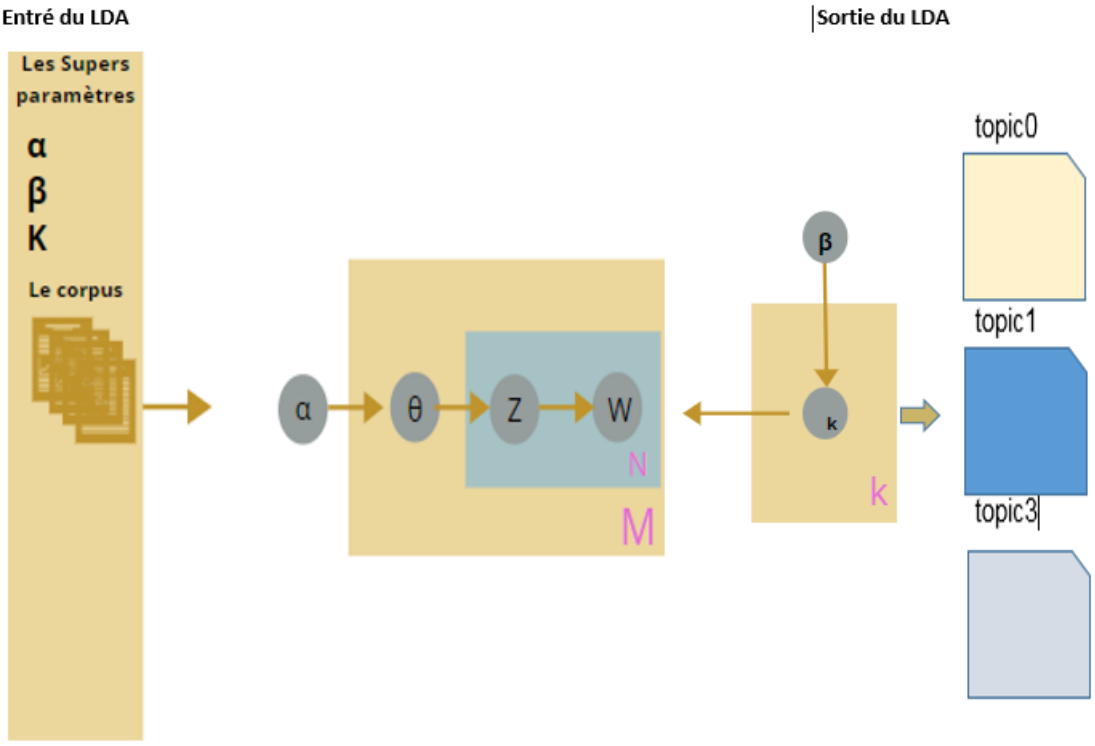

FIGURE 2.16 – Représentation de LDA sous forme de modèle graphique

Les avantages et les inconvénients de modèle LDA

le tableau suivant va montre les avantages et les inconvénients du modèle LDA :

Avantages	**Inconvénients**
Le modèle LDA est considéré plus complet par rapport à d'autres modèles	nécessité de déterminer des estimateurs pour ces paramètres
le LDA permet de généré un corpus,au niveau des mots et pas seulement des documents.	

TABLE 2.3 – Avantages et inconvénients du modèle LDA

2.7 Conclusion

Nous avons présentées dans le deuxième chapitre un état de l'art concernant les défirent algorithmes et approches les plus utile pour détecter les communautés dans un réseau social académique que ce soit une communauté disjoint ou bien chevauchant.Un tableau qui d'écrit les avantages et les inconvénients entre des défirent approche. Nous finissons ce chapitre par des notions des profil thématique des chercheurs et la modélisation thématique.

Dans le chapitre suivant, nous allons proposer une nouvelle méthode de détection des communautés thématique .

Chapitre 3

Approche de détection de communautés thématiques chevauchantes

3.1 Introduction

Nous présentons dans ce chapitre, une approche de détection de communautés chevauchantes dans les réseaux sociaux académiques cette approche est basé sur l'utilisation d'un algorithme de détection de communauté désigner par percolation des clique (clique percolation) qui nous avons modifier et produit une nouvelle variante dans la quelle nous avons intégrer une technique d'extraction des topics [1] . Nous présentons tout d'abord un schéma et une description général concernant notre approche proposée. Ensuite, nous détaillons chaque phase de notre approche.

3.2 Description de l'approche proposée

Notre approche "détection de communautés thématiques chevauchantes ", basée sur les centres d'intérêts des chercheurs, puis on détecte les communautés chevauchantes existant dans notre réseau social académique, cette approche est organisée comme suite : capture et collecte d'informations, construction des profils thématiques des chercheurs, construction du graphe et la dernière phase qui est la détection de communautés thématiques.

1. Ensembles des sujets existant dans un article

Les phases de l'approche

la figure ce dessus (3.1) va montre les différents phase de notre approches

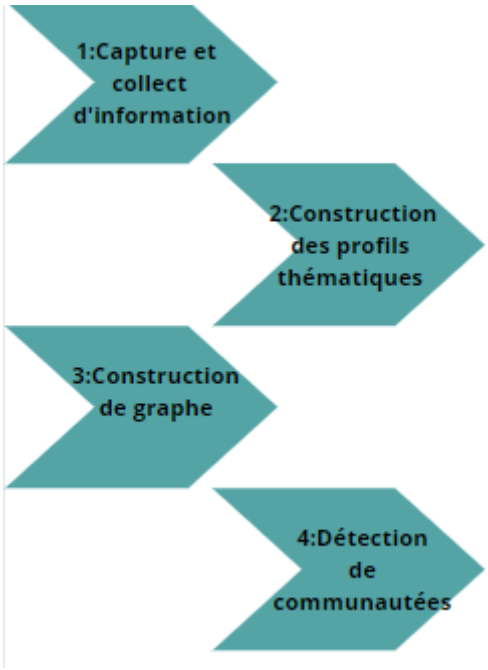

FIGURE 3.1 – Les phases de l'approche proposée

Phase 1 : capture et collecte d'informations

Chaque chercheur posté plusieurs articles concernant son domaine d'intéressement dans un réseau social académique. Pour ce là, nous capturons ces articles avec une technique du web crawling [2] .

L'ensemble des articles collectés est appelé un corpus [3] . Notre Data set est un fichier csv qui contient un ensemble des documents qui possède le nom de chercheure 'author', le titre d'article 'title', le sujet 'subject', les mots clé 'keywords' et le résumer du texte 'abstract'.

.

2. technique permet d'extraire un maximum d'informations possibles afin de connaître la structure d'un site

3. Un ensemble d'articles

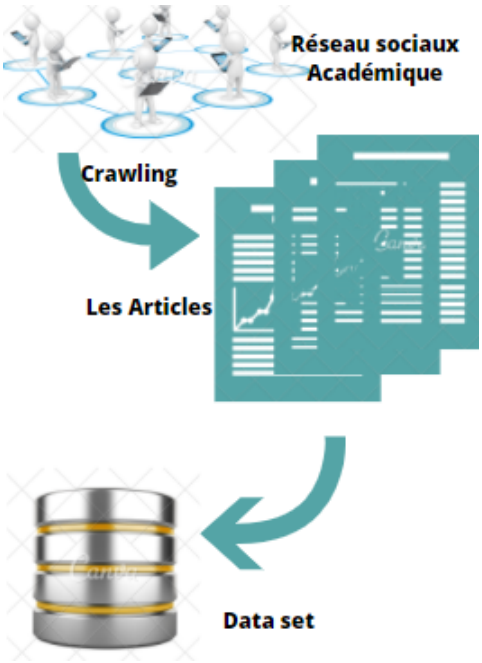

FIGURE 3.2 – Schéma résumant la première phase

Phase 2 : construction des profils thématiques des chercheurs

La (Figure 3.3)va représenter la construction des profils thématiques des chercheurs

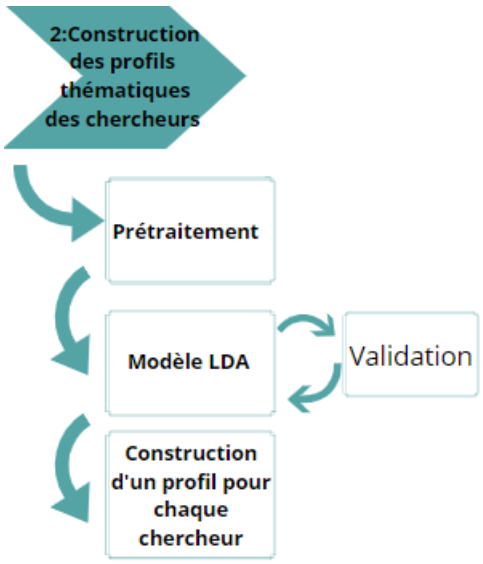

FIGURE 3.3 – Les phases de construction des profils thématiques des chercheurs

pré-traitement :

Dans cette étape nous appliquons sur chaque document (article) de notre data set la technique du pré-traitement (preprocessing) qui permet d'éliminer les bruit et de repérer les données incomplètes, nettoyer notre data set et représenter chaque article par son sac de mot(bac of word)[4] D'après les opérations suivant :

•Éliminations des mots vides (Stop words) : éliminer les mots très courants qui ne rendent pas compte de la signification propre du document.

•tokenisation et normalisation : découpage du texte en « termes » avec la suppression des accents les ponctuation par exemple, ou normalisation des acronymes

•stimming, lemmatisation : rendre la racine des mots pour éviter le biais des variations autour d'un même sens

Ensuite, nous représentant chaque article par un ensembles des qu' il contient,sans souci de contexte ,donc chaque article est représenté par un vecteur composé de fréquences des mots Article/Termes.

Modèle LDA : Après le prétraitement de notre date set nous appliquons le modèle LDA qui permet d'extraire un ensemble des topices.

Le choix est porté sur cette technique a cause de :

- LDA est un modèle probabiliste génératif, qui suppose un a priori de Dirichlet sur les sujets latents.
- En pratique, le LSI est beaucoup plus rapide à entraîner que le LDA, mais sa précision est moindre.
- LDA est la version bayésienne de PLSI.

Construction des profils thématiques des chercheurs :

Et enfin, d'après les topics qui construit par LDA, nous construisons une matrice chercheur/topic' qui définit le profile thématique de chaque chercheur la (Figure 3.4) va montre le résultat lorsque on applique le modèle LDA .

4. chaque document est représenté par un vecteur de la taille du vocabulaire

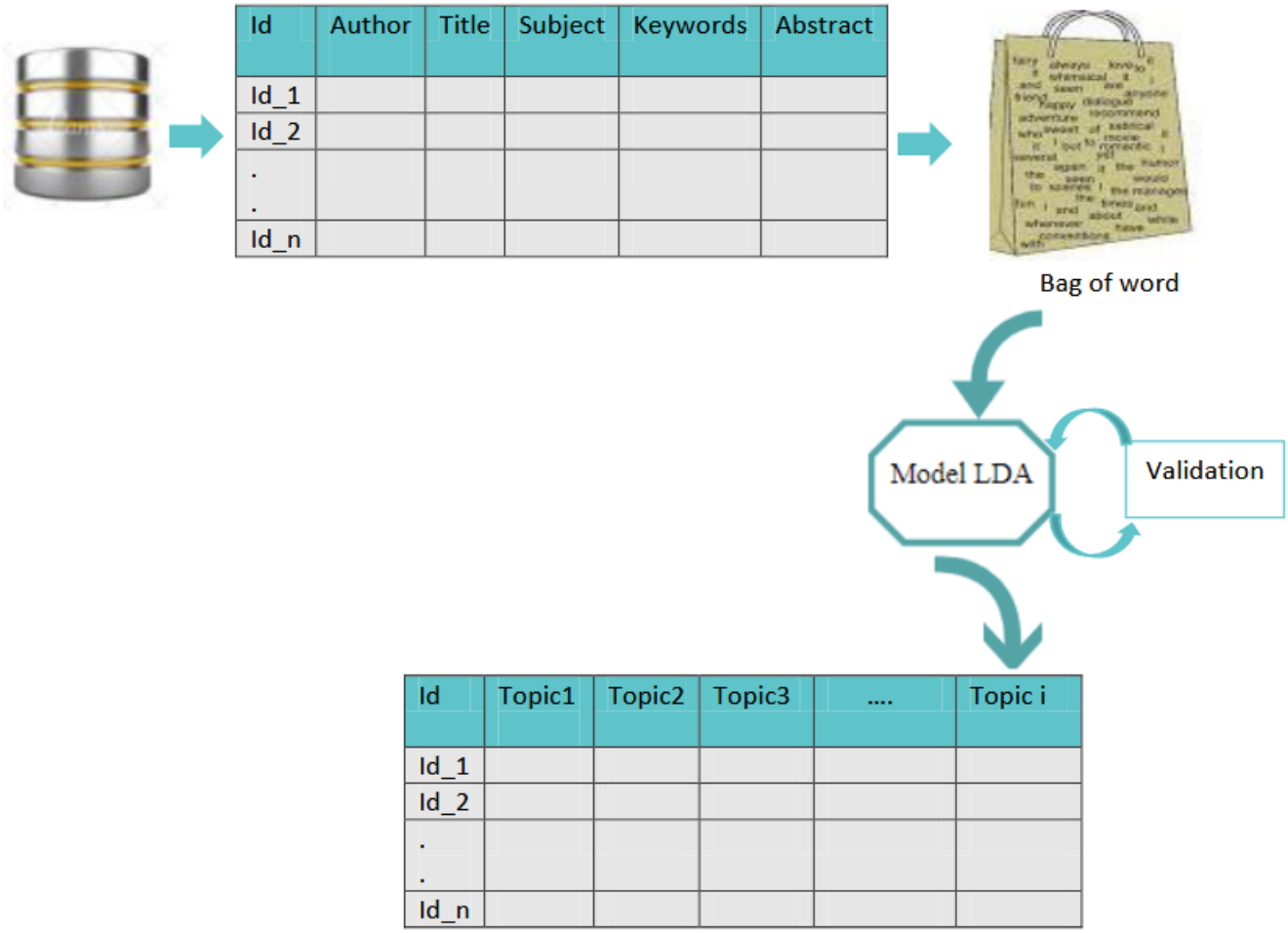

FIGURE 3.4 – Schéma résumant la deuxième phase.

Phase 3 : construction du graphe

le détail de cette phase est montré dans la figure ce dessus

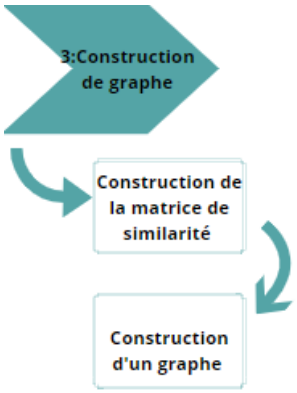

FIGURE 3.5 – Les phases de construction du graphe

Construction de la matrice de similarité chercheur-chercheur :

A partir de matrice 'chercheur-topic', Nous construisons une matrice de similarité de cosinus 'chercheur'- chercheur, Nous utilisons pour cela une fonction qui permet de cherche la similarité entre les profils des chercheurs. comme il est montre ce dessus.

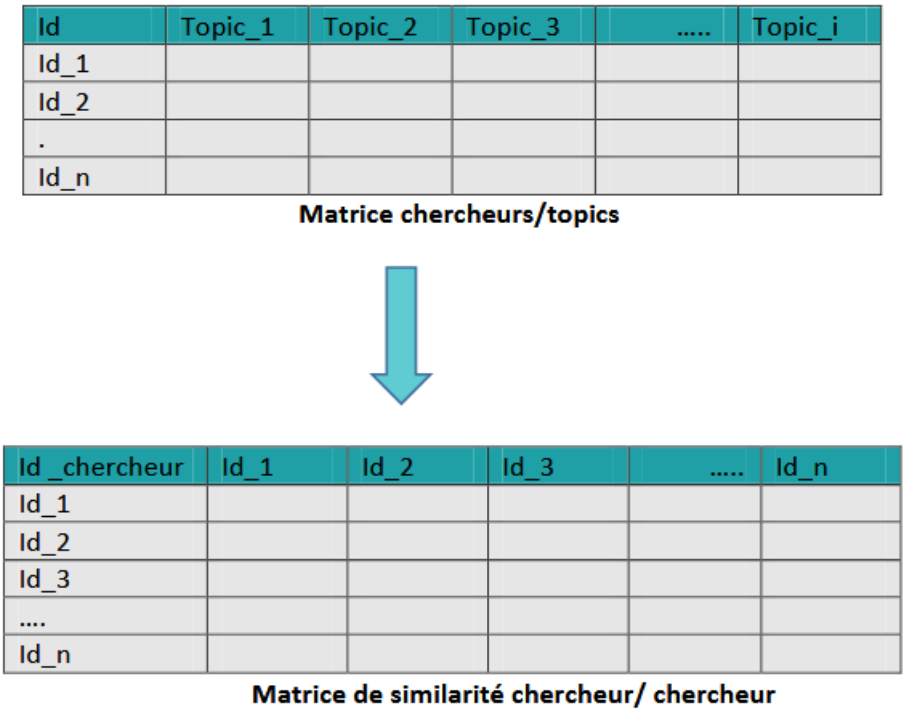

FIGURE 3.6 – La matrice de similarité chercheur-chercheur

Construction de graphe :

Nous avons converti la matrice de similarité 'chercheur- chercheur' en graphe, nous représentons chaque chercheur par un nœud où chaque deux chercheur similaire y'a toujours un lien. Pour cela on utilise un seuil (seuil >0.5) qui permet d'éliminer les arcs les plus faible poids entre les nœuds.

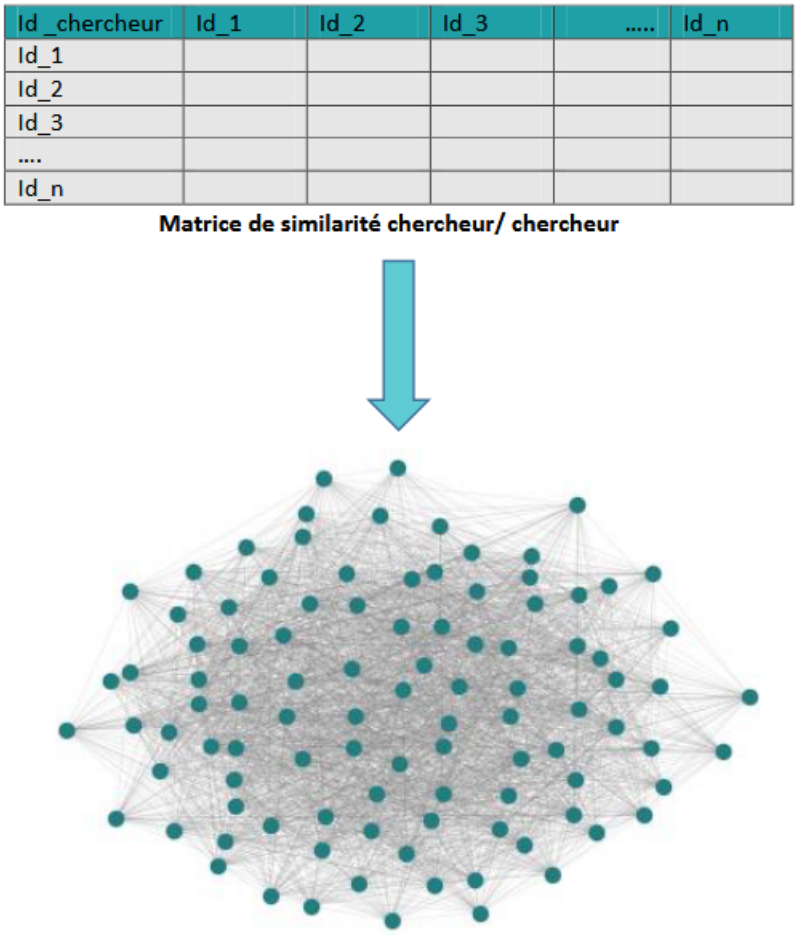

FIGURE 3.7 – Construction du graphe

Phase4 : détection de communautés :

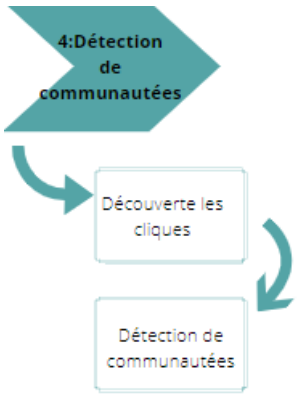

FIGURE 3.8 – Les phases de détection de communautés

Cette phase consiste à détecter les communautés thématiques à partir du graphe de la phase précédente, Pour découvrir ces communautés nous utilisons l'algorithme de détection de communautés thématiques chevauchante la Percolation des cliques :

Découverts les cliques :

Nous identifiant dans cette étape toutes les K- cliques [5]adjacent présentes dans le graphe.

Découverts les communautés :

Nous construisons un nouveau graphe G_c appelé clique-graph, de telle sorte que les sommets de celui-ci représentent les k-clique identifiées déjà dans la phase précédente.
Ensuite, Nous formons les communautés, par l'identification des composantes connexes du graphe G_c dont les cliques composent les communautés, un sommet compris dans plusieurs cliques situées dans des communautés différentes est alors un chevauchement entre les communautés.

3.3 Algorithme proposée

L'algorithme proposé a pour entrée un réseau social académique . il détecte une structure de communautés thématiques chevauchantes en sortie.

Premièrement nous extrayons le corpus des chercheurs, puis nous extrayons les meta données de chaque article(Crawling), en suit nous construisons la représentation vectoriel de chaque article. Après l'application du la technique du LDA nous extrayons les topics, puis nous validons le modèle LDA pour obtenu le TopModel, une fois les topics son extraire, nous construisons le vecteur 'document/topic' est nous appliquons l'équation du moyenne pour agréger les chercheurs ,et extraire la matrice 'chercheurs/topics'. Après, nous générons une matrice de similarité à partir de la matrice 'chercheurs/topics'. En suit nous transformons Cette matrice à un graphe. Enfin, nous appliquons l'algorithme de détection de communautés thématique sur le graphe.

5. K est la taille du clique. Clique :sous graphe induit complet

Algorithm 3 Approche de détection des communautés thématique chevauchante

Input: Academic social network (Data set)

Output: thematic community

Description :

Begin

1 : for each researcher $r_j \epsilon$ N { //N : represents all researchers,

2 : Corpus = ExtractCorpus(R) ;

3 : while Corpus is notnull() {

4 : for each article a_i { //a_i item belongs to set M , M : set of documents from all researchers

5 : MetaData = ExtractMetaData(a_i) ;

6 : DataV=DataVectorized(MetaData)

7 : Model = LDAMethod(DataV) ; //Extracting the topics

8 :TopModel=validation-Model //perplexity

9 :}

10 :}

11 : V = ConstructVector(DataV, Model) ; // Construct a vector document-topics.

12 :Ag=AggregateTopics(Moy) ; //Aggregate topics for each researcher

13 :}

14 : Matrix-topics =ConstructMatrix researcher topic(V) ;

15 : Matrix-similarity = Construct similarity matrix(Matrix-topics) ;

16 :graph=ConstructGraphe(Matrix-similarity) ;

17 : Comminity = DetectCommunities(graph) ;

18 : Return community

End

3.4 Conclusion

Nous avons présentés dans ce chapitre notre approche pour la détection de communauté chevauchante dans les réseaux sociaux académiques. Nous avons développé un algorithme à base de Percolation de cliques entres les chercheurs du réseau. Avant d'exécuter cette algorithme, Nous avons extraire tout d'abord les profile thématique de chaque chercheur à partir de modèle LDA ,et on a construire un graphe qui collaborer tout les chercheurs de notre réseau, qu'il est l'entrée de notre approche, lorsque on exécuter cette approche on va regrouper les chercheurs qui partage le même intérêt dans une communauté ou plus La méthode de percolation de clique qui est une méthode flexible pour la détection de communautés thématiques chevauchante.

Chapitre 4

Implémentation et évaluation

4.1 Introduction

Après avoir vu le fonctionnement des algorithmes proposés dans le chapitre précédent, dans ce chapitre nous allons voir tout ce qui concerne l'évaluation des algorithmes proposés.Ce chapitre est organisé comme suit : premièrement, nous décrivons Outils et environnements de développement. Ensuite, nous présentons l'étude de cas . Enfin nous allons donné les résultats et évaluer les performances de l'algorithme proposé

4.2 Outils et environnements de développement

4.2.1 Environnement matériel

Au cours de ce présent projet, tous les travaux ont été effectués sur ordinateur qui présente les caractéristiques techniques suivantes :

Matériels	**Laptop**
CPU	Intel Core i5-7200@2.50GHz
RAM	8 GO
GPU	AMID Radeon R7 M440
OS	Windows 10 pro x64
STORAGE	HDD 1Tb 7200RPM

Table 4.1 – Environnement matériel

4.2.2 Environnement logiciel

Anaconda

Anaconda [1] est une distribution libre et open source des langages de programmation Python et R appliqué au développement d'applications dédiées à la science des données et à l'apprentissage automatique (traitement de données à grande échelle, analyse prédictive, calcul scientifique), qui vise à simplifier la gestion des paquets et de déploiement [**?**]

Jupyter notebook

Jupyter Notebook [2] est une application Web open source qui permet de créer et de partager des documents contenant du code en direct, des équations, des visualisations et du texte narratif. Les utilisations comprennent : le nettoyage et la transformation des données, la simulation numérique, la modélisation statistique, la visualisation des données, l'apprentissage automatique et bien plus encore. [**?**]

Les Avantages de Jupyter notebook

- Flexible
- Figures / Graphiques / Tracés / Visualisation
- Il s'agit d'un package complet qui peut être converti en pdf
- Préparation des documents plus facile

Google Colab

Colab [3] est un environnement de notebook Jupyter gratuit qui s'exécute entièrement dans le cloud. Plus important encore, il ne nécessite pas de configuration et les blocs-notes que vous créez peuvent être modifiés simultanément par les membres de votre équipe - tout comme vous modifiez des documents dans Google Docs. Colab prend en charge de nombreuses bibliothèques de machine learning populaires qui peut être facilement chargées dans votre notebook.[**?**]

1. https ://www.venturelessons.com/what-is-anaconda/
2. https ://www.quora.com/What-exactly-is-E2-80-98Jupyter-Notebook-E2-80-99
3. https ://www.tutorialspoint.com/google-colab/what-is-google-colab.htm

Les avantages de Google colab

— Écrire et exécuter du code en Python
— Documentez votre code qui prend en charge les équations mathématiques
— Importer / publier des blocs-notes depuis GitHub
— Importez des ensembles de données externes, par ex. depuis Kaggle

4.2.3 Langage utilisé

Python

Python [4] est un langage interprété. Cela signifie qu'il n'est pas converti en code lisible par ordinateur avant l'exécution du programme, mais au moment de l'exécution. Dans le passé, ce type de langage était appelé un langage de script, laissant entendre que son utilisation était pour des tâches insignifiantes. Cependant, les langages de programmation tels que Python ont forcé un changement dans cette nomenclature. De plus en plus, les applications volumineuses sont écrites presque exclusivement en Python.[**?**]
Voici quelques façons d'appliquer Python :

— Programmation CGI pour applications Web
— Lire et écrire sur MySQL
— Créer des calendriers en HTML
— Travailler avec des fichiers

4.3 L'ensemble des données exploitées

Pour exécuté notre algorithme proposé nous avons besoin d'une dataset publier par « Daneil kershaw et Rob koeling » en août 2020 [5]. Il s'agit d'un réseau de 745 chercheurs, chaque chercheur publier au minimum 3 articles donc on a 2671 articles. (Figure4.1)

4. https ://www.thoughtco.com/what-is-python-2813564
5. https ://data.mendeley.com/datasets/zm33cdndxs/draft ?a=ef6d0d03-1102-4b7e-9195-43dd9d1be3b1

	author	tilte	subject	keywords	abstract
0	A. Bachmaier	Structural evolution and strain induced mixing...	ENGI;MATE;PHYS	Atom probe tomography;Cu-Co;High-pressure tors...	A Cu-Co composite material is chosen as a mode...
1	A. Bachmaier	On the remarkable thermal stability of nanocry...	ENGI;MATE;PHYS	Cobalt;Nanostructured materials;Severe plastic...	Nanostructured Co materials are produced by se...
2	A. Bachmaier	On the process of co-deformation and phase dis...	MATE	Atom probe tomography;Composites;Mechanical al...	In this study, dual phase Cu-Co composites wit...
3	A. Banos	Corrosion of uranium in liquid water under con...	CENG;CHEM;MATE	Bulk-UH 3;FIB;Threshold pressure;Uranium;Water...	The reaction of unirradiated-U with liquid wat...
4	A. Banos	Corrosion of uranium in liquid water under vac...	CENG;CHEM;MATE	Bulk-UH 3;FIB;Threshold pressure;Uranium;Water...	The reaction of unirradiated-U with liquid wat...
5	A. Banos	The effect of work-hardening and thermal annea...	CENG;CHEM;MATE	B. SEM;C. effects of strain;Hydrogen permeatio...	The characteristics of hydride formation on me...
6	A. Banos	A review of uranium corrosion by hydrogen and ...	CENG;CHEM;MATE	B. SEM;B. SIMS;C. Effects of strain;C. Hydroge...	Uranium hydride (UH3) is the direct product of...
7	A. Banos	The effect of sample preparation on uranium hy...	CENG;CHEM;MATE	B. SEM;B. SIMS;C. Effects of strain;C. Hydroge...	The influence of sample cleaning preparation o...
8	A. Bertei	A novel approach for the quantification of inh...	CHEM;ENER;ENGI	3D tomography;Advanced characterization;Guidel...	The electrode microstructural properties signi...
9	A. Bertei	The fractal nature of the three-phase boundary...	ENER;ENGI;MATE	Anode degradation;Coarsening;Electrochemical i...	Nickel/zirconia-based nanostructured electrode...
10	A. Bertei	Validation of a physically-based solid oxide f...	ENER;PHYS	Anode;Charge-transfer;Electrochemical impedanc...	This study presents a physically-based model f...

FIGURE 4.1 – Partie du dataset.

nous appliquons le pré traitement sur la colonne "text" qui fusionne les deux colonnes "kekwords" et "title" pour donne une bonne résultat de modularité lorsque on applique le modèle LDA . la figure ce dessus montre résultat de fusionnement

	author	tilte	subject	keywords	abstract	text
0	A. Bachmaier	Structural evolution and strain induced mixing...	ENGI;MATE;PHYS	Atom probe tomography;Cu-Co;High-pressure tors...	A Cu-Co composite material is chosen as a mode...	Structural evolution and strain induced mixing...
1	A. Bachmaier	On the remarkable thermal stability of nanocry...	ENGI;MATE;PHYS	Cobalt;Nanostructured materials;Severe plastic...	Nanostructured Co materials are produced by se...	On the remarkable thermal stability of nanocry...
2	A. Bachmaier	On the process of co-deformation and phase dis...	MATE	Atom probe tomography;Composites;Mechanical al...	In this study, dual phase Cu-Co composites wit...	On the process of co-deformation and phase dis...
3	A. Banos	Corrosion of uranium in liquid water under con...	CENG;CHEM;MATE	Bulk-UH 3;FIB;Threshold pressure;Uranium;Water...	The reaction of unirradiated-U with liquid wat...	Corrosion of uranium in liquid water under con...
4	A. Banos	Corrosion of uranium in liquid water under vac...	CENG;CHEM;MATE	Bulk-UH 3;FIB;Threshold pressure;Uranium;Water...	The reaction of unirradiated-U with liquid wat...	Corrosion of uranium in liquid water under vac...
...	...	...	...	...	...	...
2666	Zoi C. Tetta	Textile-reinforced mortar (TRM) versus fiber-r...	ENGI;MATE	Carbon fibre;Concrete strengthening;Debonding;...	This paper presents an experimental study on s...	Textile-reinforced mortar (TRM) versus fiber-r...
2667	Zoi C. Tetta	Shear strengthening of full-scale RC T-beams u...	ENGI;MATE	A. Carbon fibre;A. Fabrics/textiles;A. Glass f...	This paper presents a study on the effectivene...	Shear strengthening of full-scale RC T-beams u...
2668	Zoi C. Tetta	Shear strengthening of concrete members with T...	ENGI;MATE	Basalt fibres;Carbon fibre;Debonding;Fracture;...	An experimental work on reinforced concrete (R...	Shear strengthening of concrete members with T...
2669	Zoi C. Tetta	TRM vs FRP jacketing in shear strengthening of...	ENGI;MATE	Carbon fibre;Debonding;Fabrics/textiles;Glass ...	This paper presents the first study on the per...	TRM vs FRP jacketing in shear strengthening of...
2670	Zoi C. Tetta	On the design of shear-strengthened RC members...	ENGI;MATE	Design models;Reinforced concrete;Shear streng...	Textile reinforced mortar (TRM) is a promising...	On the design of shear-strengthened RC members...

FIGURE 4.2 – Partie du dataset après le fusionnement .

Lorsque on applique le modèle LDA sur notre dataset nous a permis d'extraire 10 topics (Figure4.3)

```
Topic 0: analysis use particle element single drug image model datum recovery
Topic 1: model energy control surface simulation power storage source gas uncertainty
Topic 2: structure method vaccine age development protein transport wave market strength
Topic 3: property use metal field long term distribution study quality effect
Topic 4: model base emission economic value urban solid service growth organic
Topic 5: cell fuel temperature microbial power solar thermal application performance size
Topic 6: energy change low carbon water policy transition cycle climate management
Topic 7: response state self high function loss local base treatment spatial
Topic 8: risk stress rate food health policy global study product brain
Topic 9: phase heat flow thermal transfer factor alloy numerical material high
```

FIGURE 4.3 – Les topics extraits.

Après l'extraction des topics on les représenté sous forme d'une matrice Documents-Topics(Figure4.4) Il ya toujours un topic dominant pour chaque documents

	Topic_0	Topic_1	Topic_2	Topic_3	Topic_4	Topic_5	Topic_6	Topic_7	Topic_8	Topic_9	Dominant_topic
Doc0	0.005	0.005	0.645	0.005	0.005	0.005	0.005	0.315	0.005	0.005	2
Doc1	0.01	0.01	0.01	0.01	0.91	0.01	0.01	0.01	0.01	0.01	4
Doc2	0.006	0.006	0.27	0.006	0.474	0.006	0.006	0.217	0.006	0.006	4
Doc3	0.008	0.931	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	1
Doc4	0.007	0.94	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	1
Doc5	0.006	0.085	0.006	0.398	0.006	0.153	0.006	0.327	0.006	0.006	3
Doc6	0.006	0.324	0.006	0.161	0.112	0.006	0.103	0.272	0.006	0.006	1
Doc7	0.009	0.186	0.009	0.268	0.009	0.009	0.009	0.483	0.009	0.009	7
Doc8	0.007	0.381	0.007	0.007	0.283	0.007	0.007	0.007	0.289	0.007	1
Doc9	0.007	0.007	0.007	0.007	0.264	0.007	0.308	0.007	0.079	0.309	9

FIGURE 4.4 – Partie de la matrice documents/topics..

Nous avons agrégé les topics pour chaque chercheur afin de trouver le profile thématique pour chaque chercheur on a utiliser l'équation du moyenne , on la représenter mathématiquement comme suite :

$$\sum_{i=1}^{N_j} \frac{P(Z_k/d_i)}{N_j}$$

avec : **k** :nombre de topics , **j** :identifiant de chercheur , N_j :nombre de document du chercheur j ,
la figure (4.5)représentre la matrice des topics (chercheur/topics)

	Unnamed: 0	Topic_0	Topic_1	Topic_2	Topic_3	Topic_4	Topic_5	Topic_6	Topic_7	Topic_8	Topic_9
0	A. Bachmaier	0.007000	0.007000	0.308333	0.007000	0.463000	0.007000	0.007000	0.180667	0.007000	0.007000
1	A. Banos	0.007200	0.493200	0.007200	0.168400	0.028400	0.036600	0.026600	0.219400	0.007200	0.007200
2	A. Bertei	0.006333	0.131000	0.006333	0.006333	0.210000	0.006333	0.331333	0.006333	0.190667	0.107000
3	A. Beth	0.009333	0.009333	0.009333	0.873000	0.051667	0.009333	0.009333	0.009333	0.009333	0.009333
4	A. D. Anastasiou	0.015500	0.015500	0.015500	0.363000	0.015500	0.015500	0.015500	0.198000	0.330250	0.015500
...	...	...	...	...	...	...	...	...	...	...	...
740	Zhaolong Yu	0.005667	0.005667	0.005667	0.005667	0.078000	0.263667	0.621333	0.005667	0.005667	0.005667
741	Zhen Zhang	0.304857	0.207286	0.153286	0.029714	0.089857	0.008571	0.120000	0.018286	0.008571	0.058571
742	Zhiwei Ma	0.006200	0.006200	0.126200	0.370400	0.155200	0.049000	0.150800	0.006200	0.006200	0.125000
743	Zoe Berk	0.008333	0.008333	0.495667	0.008333	0.008333	0.008333	0.439667	0.008333	0.008333	0.008333
744	Zoi C. Tetta	0.007000	0.007000	0.069600	0.232000	0.007000	0.029400	0.022400	0.612000	0.007000	0.007000

FIGURE 4.5 – Matrice chercheurs/topics utilisant l'aggrégation par moyenne.

Afin de construire les profile thématique de chaque chercheur nous avons crée une matrice de similarité (Figure4.6) on utilisons la similarité de cosinus ,la formule de similarité est écrit comme suite :

$$Cos(x, y) = \frac{(x,y)}{||x||.||y||}$$

	res001	res002	res003	res004	res005	res006	res007	res008	res009	res010	...	res736	res737	res738	res739	res740
0	1.000000	0.181167	0.386238	0.068340	0.171038	0.586220	0.222933	0.484505	0.703230	0.707703	...	0.090280	0.151127	0.797210	0.194322	0.429092
1	0.181167	1.000000	0.317057	0.313803	0.385816	0.080808	0.076450	0.249140	0.351780	0.102448	...	0.367586	0.357272	0.072823	0.413575	0.406019
2	0.386238	0.317057	1.000000	0.057899	0.317594	0.498457	0.591776	0.110802	0.488646	0.742612	...	0.301627	0.764558	0.470163	0.339658	0.071523
3	0.068340	0.313803	0.057899	1.000000	0.696613	0.044486	0.033805	0.064462	0.220271	0.131372	...	0.949151	0.039330	0.102197	0.718861	0.032185
4	0.171038	0.385816	0.317594	0.696613	1.000000	0.078456	0.184121	0.236839	0.406751	0.102126	...	0.666511	0.082256	0.081324	0.535760	0.396637
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
740	0.113853	0.086883	0.712060	0.029522	0.056354	0.583142	0.459856	0.318706	0.437341	0.543562	...	0.311054	0.846054	0.124716	0.298685	0.037501
741	0.384985	0.493284	0.479427	0.102618	0.138866	0.502354	0.381442	0.328320	0.648731	0.338429	...	0.216591	0.586318	0.225455	0.721364	0.163055
742	0.427685	0.297353	0.464164	0.821818	0.594280	0.489019	0.478628	0.317623	0.531844	0.511198	...	0.883554	0.387310	0.372996	0.744069	0.094087
743	0.416598	0.061929	0.497430	0.029138	0.064140	0.967274	0.572285	0.506853	0.622876	0.338287	...	0.230126	0.766940	0.023904	0.329396	0.120732
744	0.356215	0.478627	0.058030	0.363999	0.600060	0.133746	0.113421	0.457099	0.684537	0.063366	...	0.352362	0.085738	0.033968	0.298818	0.937291

FIGURE 4.6 – Partie de la matrice de similarité chercheurs/chercheurs.

Cette matrice on la transforme au array liste afin de construire le graphe comme il est montre dans la (Figure4.7) . cet graphe est composé de 745 nœuds (chaque chercheur va représente par un nœud) et 277885 arcs (les arcs va représente la similarité entre chaque chercheur du graphe)

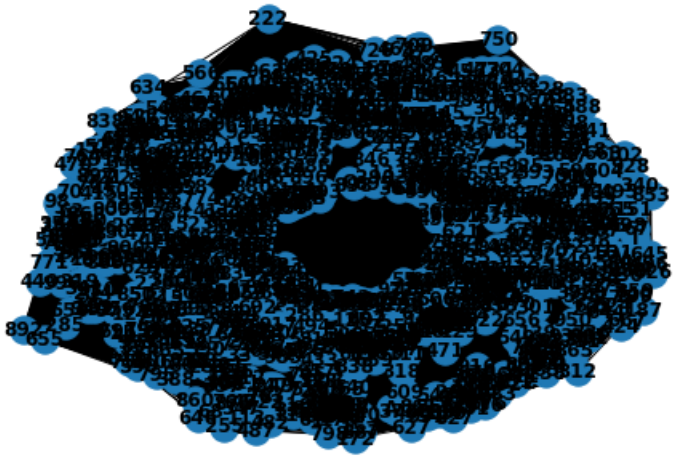

FIGURE 4.7 – Graphe représentant les chercheurs .

pour diminuer le nombre des arcs on utilisant un seuil, on va couper chaque arcs qui a un poids <= 0.5 ,Et ça devient le nombre d'arcs (61071),le nouveau graphe il est dans la figure ce dessus

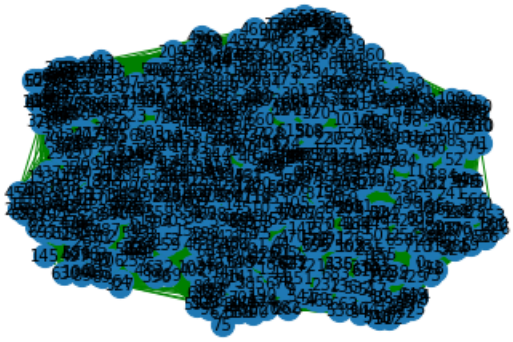

FIGURE 4.8 – Graphe représentant les chercheurs après l'élimination des arcs faibles

4.3.1 Expérimentation et résultats

Détection de communautés thématique

D'après l'exécution de notre algorithme nous avons définie tout d'abord les clique qui sont présenter dans notre graphe on a prendre le paramètre K =35 , qui signifie on a 35 nœuds et K-1 arcs dans chaque clique a partir de ces clique on a extrait les communautés

(13 communautés)qui sont monter ce dessus

les communauté détecter représentant se forme des listes :

communauté :0 [0, 512, 514, 2, 517, 519, 9, 10, 11, 524, 17, 532, 538, 541, 543, 34, 549, 37, 40, 562, 53, 58, 573, 579, 71, 73, 597, 602, 94, 607, 102, 616, 104, 112, 119, 125, 126, 129, 136, 649, 648, 138, 654, 655, 658, 148, 662, 151, 152, 155, 158, 161, 162, 678, 679, 166, 169, 683, 686, 177, 182, 696, 697, 699, 190, 191, 703, 197, 717, 207, 721, 214, 737, 232, 235, 236, 239, 241, 249, 266, 270, 275, 276, 336, 337, 339, 346, 375, 388, 391, 393, 398, 411, 414, 422, 423, 429, 432, 438, 447, 465, 480, 491, 496, 501]

communauté :1 [0, 513, 5, 519, 8, 7, 534, 545, 555, 44, 50, 562, 566, 54, 56, 58, 59, 570, 63, 575, 577, 70, 584, 73, 588, 77, 82, 595, 600, 604, 605, 101, 109, 621, 623, 118, 634, 636, 639, 127, 131, 646, 135, 656, 657, 658, 155, 161, 673, 165, 169, 171, 696, 697, 187, 191, 710, 717, 206, 719, 208, 205, 721, 724, 725, 727, 732, 743, 235, 242, 250, 254, 255, 263, 272, 279, 281, 282, 284, 288, 293, 307, 308, 317, 319, 324, 326, 330, 334, 341, 342, 348, 354, 355, 362, 372, 376, 381, 383, 391, 394, 425, 426, 435, 449, 453, 462, 463, 476, 485, 489, 498, 501, 511]

communauté :2 [1, 261, 518, 519, 270, 271, 272, 18, 20, 534, 24, 539, 540, 33, 290, 549, 295, 42, 558, 304, 561, 565, 311, 56, 568, 316, 589, 596, 344, 95, 609, 361, 110, 111, 370, 117, 118, 636, 382, 638, 128, 389, 391, 393, 650, 399, 401, 147, 409, 154, 412, 668, 671, 416, 673, 426, 176, 433, 177, 181, 694, 441, 445, 189, 192, 708, 716, 205, 719, 727, 728, 729, 227, 484, 493, 246, 247, 506, 507, 508, 509, 510]

communauté :3 [2, 3, 4, 5, 6, 8, 13, 15, 19, 23, 26, 27, 28, 32, 34, 36, 37, 38, 39, 40, 43, 47, 49, 51, 57, 61, 62, 63, 66, 68, 72, 74, 76, 77, 80, 88, 92, 96, 97, 105, 107, 109, 110, 111, 113, 119, 121, 129, 132, 133, 134, 137, 138, 140, 142, 143, 146, 148, 149, 150, 154, 158, 159, 160, 162, 163, 168, 172, 175, 176, 180, 181, 183, 184, 188, 189, 195, 197, 198, 199, 201, 202, 204, 207, 210, 215, 216, 218, 220, 221, 223, 224, 225, 228, 230, 231, 236, 237, 238, 243, 244, 245, 248, 249, 252, 253, 256, 259, 260, 262, 265, 268, 270, 277, 279, 283, 284, 287, 291, 292, 293, 295, 296, 300, 303, 309, 311, 314, 317, 318, 320, 322, 323, 328, 331, 338, 339, 341, 343, 344, 345, 351, 352, 355, 356, 357, 358, 359, 364, 365, 366, 367, 368, 371, 374, 378, 383, 384, 387, 389, 390, 395, 400, 401, 403, 404, 405, 406, 407, 408, 410, 411, 413, 417, 418, 419, 420, 421, 424, 433, 434, 437, 440, 443, 444, 445, 446, 448, 450, 454, 456, 457, 458, 459, 460, 462, 464, 467, 468, 469, 471, 474, 477, 478, 480, 484, 486, 487, 493, 499, 500, 505, 507, 513, 515, 516, 520, 521, 522, 523, 525, 527, 530, 531,

533, 534, 537, 540, 542, 546, 549, 551, 552, 553, 554, 556, 560, 563, 564, 568, 571, 575, 578, 583, 585, 587, 590, 591, 594, 597, 598, 599, 601, 602, 603, 606, 609, 610, 612, 613, 615, 620, 626, 627, 628, 633, 634, 635, 640, 644, 646, 647, 648, 651, 652, 653, 654, 660, 666, 668, 672, 677, 681, 685, 686, 688, 690, 691, 693, 696, 698, 702, 704, 705, 707, 709, 712, 713, 714, 717, 718, 722, 724, 725, 726, 727, 728, 729, 731, 733, 735, 736, 738, 740, 742, 743]

communauté :4 [515, 7, 526, 528, 24, 539, 29, 30, 543, 544, 32, 37, 550, 552, 557, 45, 46, 48, 561, 564, 53, 569, 60, 574, 65, 581, 590, 79, 80, 592, 90, 607, 98, 611, 101, 615, 106, 107, 619, 115, 631, 122, 634, 635, 641, 642, 643, 644, 645, 139, 660, 664, 153, 667, 156, 669, 670, 160, 166, 167, 680, 174, 687, 689, 180, 693, 182, 694, 185, 195, 198, 711, 200, 201, 203, 204, 722, 211, 726, 217, 730, 219, 731, 732, 228, 230, 234, 247, 260, 262, 263, 264, 267, 269, 285, 286, 294, 297, 318, 320, 322, 329, 331, 332, 333, 337, 344, 349, 360, 377, 380, 384, 386, 392, 395, 397, 412, 421, 434, 439, 447, 448, 460, 466, 469, 479, 484, 488, 495, 504]

communauté :5 [8, 12, 525, 14, 21, 535, 546, 35, 554, 42, 556, 46, 48, 570, 60, 69, 80, 599, 90, 608, 609, 101, 617, 619, 114, 117, 120, 637, 641, 129, 130, 132, 657, 146, 663, 153, 668, 160, 673, 171, 684, 173, 686, 178, 179, 694, 183, 184, 697, 186, 188, 711, 199, 715, 719, 209, 723, 727, 728, 216, 218, 734, 225, 739, 229, 744, 237, 238, 258, 260, 264, 270, 271, 283, 284, 286, 299, 301, 302, 321, 323, 325, 350, 352, 363, 375, 378, 382, 384, 387, 392, 397, 400, 405, 408, 428, 436, 451, 455, 468, 473, 485, 502, 506]

communauté :6 [8, 264, 270, 284, 546, 35, 549, 554, 560, 597, 352, 609, 374, 378, 384, 129, 387, 132, 400, 146, 403, 150, 663, 160, 686, 181, 183, 184, 186, 188, 199, 711, 468, 216, 728, 218, 733, 225, 237, 238, 499, 244, 245]

communauté :7 [384, 129, 387, 132, 8, 270, 146, 403, 150, 284, 546, 35, 549, 38, 554, 686, 560, 181, 183, 186, 188, 199, 468, 597, 216, 728, 733, 352, 609, 237, 238, 499, 244, 245, 374, 378]

communauté :8 [519, 8, 526, 527, 529, 530, 531, 20, 22, 539, 540, 30, 31, 543, 545, 550, 41, 45, 49, 52, 54, 567, 55, 57, 572, 60, 576, 65, 67, 580, 69, 78, 81, 85, 87, 607, 95, 611, 613, 105, 108, 116, 119, 139, 142, 665, 667, 157, 672, 161, 673, 675, 164, 676, 166, 170, 179, 181, 194, 197, 711, 716, 206, 207, 213, 222, 738, 227, 741, 233, 234, 241, 251, 260, 263, 268, 273, 274, 280, 283, 289, 298, 304, 305, 306, 307, 310, 313, 320, 335, 340, 360, 363, 366, 367, 373, 379, 380, 382, 397, 401, 417, 422, 430, 433, 452, 455, 472, 476, 492, 496,

499, 511]

communauté :9 [256, 257, 4, 262, 265, 11, 13, 15, 529, 531, 20, 19, 535, 25, 291, 547, 294, 39, 296, 43, 44, 302, 559, 49, 567, 312, 315, 316, 575, 64, 325, 582, 327, 74, 586, 77, 591, 83, 596, 89, 603, 93, 605, 95, 353, 354, 99, 356, 100, 614, 103, 357, 620, 113, 626, 369, 628, 629, 630, 372, 632, 638, 642, 643, 644, 645, 396, 401, 145, 407, 664, 154, 157, 415, 671, 162, 676, 678, 167, 190, 706, 198, 200, 460, 461, 725, 214, 472, 474, 477, 482, 226, 483, 486, 495, 497, 252]

communauté :10 [6, 14, 16, 533, 534, 21, 536, 541, 548, 38, 552, 555, 49, 563, 565, 574, 575, 68, 75, 76, 593, 84, 86, 602, 91, 605, 103, 105, 618, 106, 621, 622, 623, 624, 625, 114, 110, 123, 124, 134, 648, 138, 140, 141, 144, 659, 147, 661, 151, 674, 682, 692, 695, 696, 699, 700, 701, 190, 193, 194, 196, 200, 713, 720, 721, 212, 218, 732, 223, 228, 232, 238, 240, 249, 259, 266, 273, 278, 283, 286, 287, 330, 345, 347, 369, 370, 373, 381, 385, 391, 397, 400, 402, 406, 410, 412, 427, 431, 435, 442, 451, 459, 465, 468, 470, 472, 475, 481, 487, 488, 490, 494, 502, 503, 510, 511]

communauté :11 [513, 391, 639, 281, 165, 555, 44, 562, 435, 187, 319, 575, 577, 449, 324, 710, 70, 330, 588, 77, 462, 206, 342, 604, 605, 355, 485, 235, 109, 622, 624, 498, 242, 381, 255]

communauté :12 [513, 391, 511, 639, 534, 281, 165, 555, 171, 435, 570, 187, 319, 575, 577, 449, 324, 70, 710, 330, 588, 77, 206, 462, 604, 605, 355, 485, 235, 109, 622, 621, 624, 623, 498, 242, 636, 381, 255]

La figure (4.9) ,c'est un Graphe représentant les chercheurs pour chaque communauté.

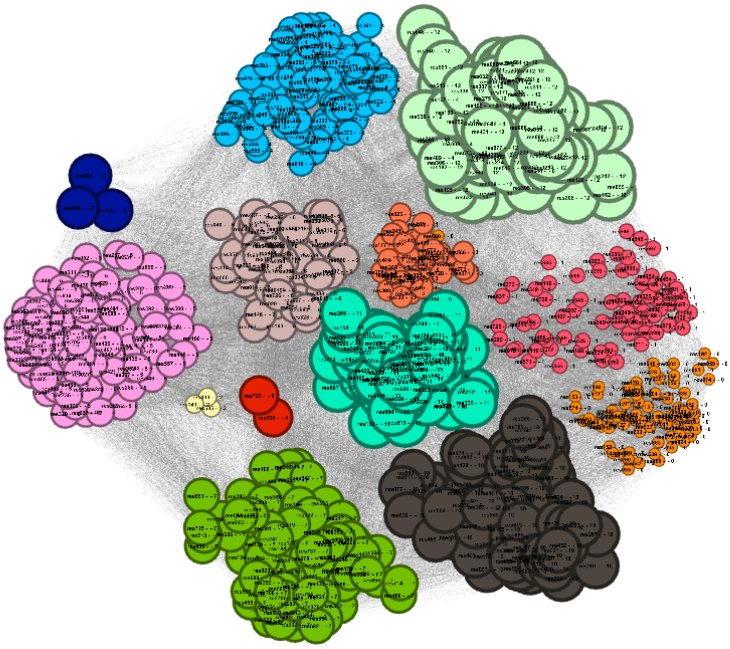

FIGURE 4.9 – Graphe représentant les chercheurs pour chaque communauté.

La (figure 4.10) va représenté un graphe qui illustre les communauté détecter

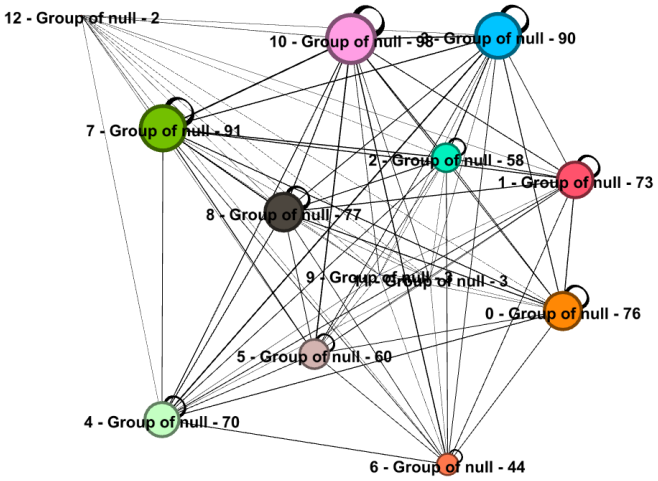

FIGURE 4.10 – Graphe représentant les communautés détecter.

4.4 Evaluation et discussion

4.4.1 Métrique d'évaluation

Après avoir vu le fonctionnement de l'algorithme proposé ,nous allons évaluer les performances de algorithme proposé :

Le degré de distribution

Permet de calculer le degré d'une relation au sein du réseau.comme il est montre dans la figure ce dessus

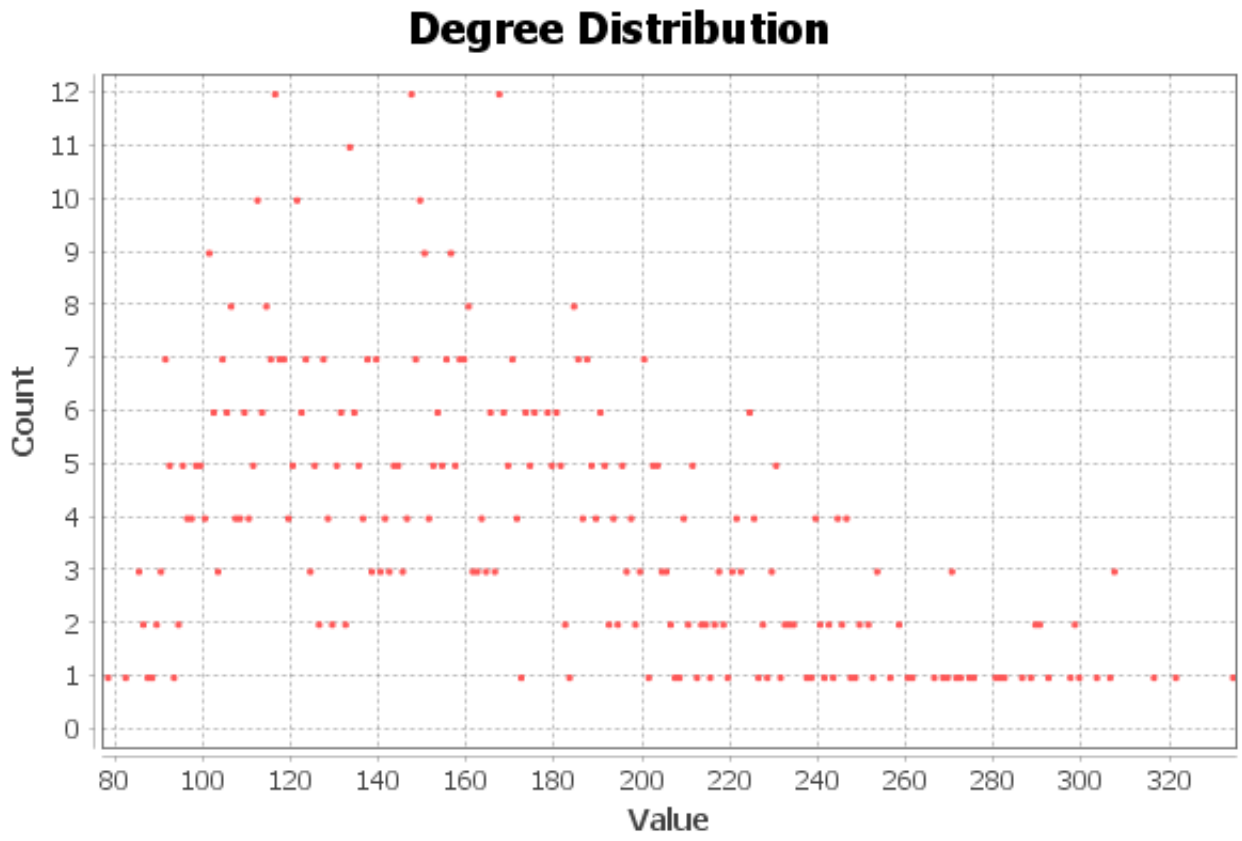

FIGURE 4.11 – Degré de distribution

La modularité

pour calculer la modularité on a applique l'algorithme Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, Fast unfolding of communities in large networks . la formule qui permet de calculer la modularité est :

$$Q = \frac{1}{2m}\sum_{ij}[A - \frac{k_i k_j}{2m}]\sigma(C_i, C_j)$$

m : le poids total des arcs du réseau

A : la matrice d'adjacence du réseau

K,k : le poids sommes i et j

σ : la fonction delta de Kronecker défini comme suit $\sigma(C_i, C_j)$ { 1 i=j 0 i≠j}

.

la valeur de Modularité $= 0,420$

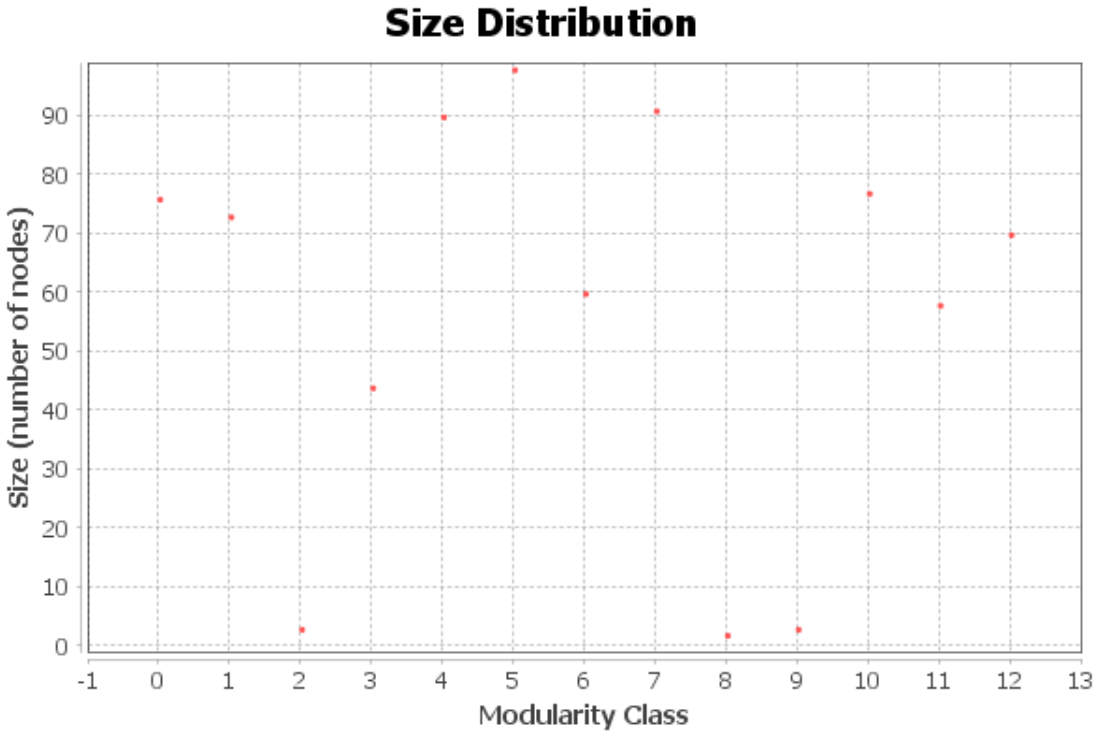

FIGURE 4.12 – La modularité

La densité

la valeur de densité est égale à : val-densité$= 0,218$

Le diamètre

la Figure 4.13 va montre le diametre .

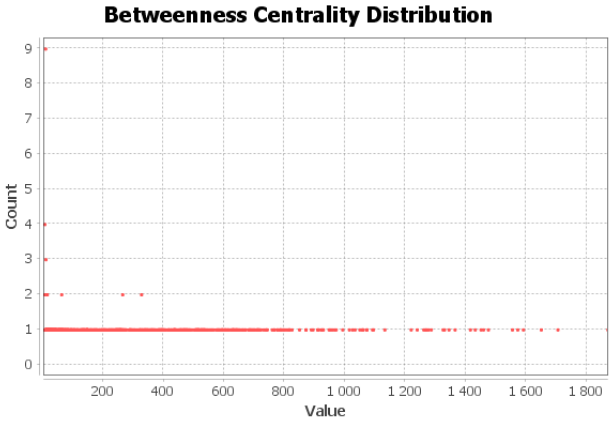

FIGURE 4.13 – Le diametre

Donc nous avons présenté une nouvelle approche de détection de communautés thématiques dans les réseaux sociaux académique dont la propriété principale est le partage d'un même domaine d'intérêt des chercheurs. Avec une modularité de 0.42, le nombre de communautés découvertes confirme une grande précision de l'approche sachant qu'avec une modularité supérieure à 0,3 le résultat est considéré comme précis.

4.5 Conclusion

Dans ce chapitre, nous avons d'abord traité une dataset , puis dans l'évaluation, nous avons présenté quatre résultats d'analyse ,la première c'est Le degré de distribution,la deuxième c'est de type modularité et la dernière la densité et Le diamètre.

Conclusion générale

Ce travail visait à détecté les communautés thématiques dans un réseau social académique. Pour cela, nous nous sommes d'abord intéressés aux dernières méthodes de la détection de communautés chevauchantes.

Le travail a débuté par une étude axée sur les généralités des réseaux sociaux et les réseaux sociaux académiques, Ensuite un deuxième chapitre a été consacré sur état de l'art des méthodes de détection des communautés .

Le troisième chapitre a été entamé par une proposition d'une nouvelle méthode de détection des communautés thématiques, dont la propriété principal est le partage d'un même intérêt pour une ou plusieurs thématiques. En effet, 'a travers la méthode proposée, les chercheurs sont représentés par l'ensemble des thématiques qui traduisent leur intéressement. Notre approche permet de regrouper les chercheurs en utilisant la méthode du percolation des cliques. Enfin les résultats de simulation sont été présentés dans le dernier chapitre.

En guise de perspectives, nous envisageons de faire les points suivants :
La première perspective,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 pour construire le profil thématique de chercheur.
La deuxième perspective, sera de faire l'objet de recommandation sur les publications récentes concernant le domaine d'intérêt.

Bibliographie

[1] https ://www.blogdumoderateur.com/chiffres-reseaux-sociaux/ consulter le 01/02/2020

[2] FRIENDSTER. [En ligne].http ://www.friendster.com/

[3] MYSPACE. [En ligne].http ://fr.myspace.com/

[4] https ://www.francoischarron.com/ consulter le 11/02/2020

[5] CREFF Marie, "Réseaux sociaux : quelles opportunités pour les services Le cas de l'assistance en ligne d'Orange" Institut national des techniques de la documentation, Mémoire de fin d'étude INTD 2010.

[6] ZEMMAR Nisrine,Réseaux Sociaux numériques : essai de catégorisation et cartographie des controverses,Université Rennes 2, Thèse de doctorat 2012.

[7] TORLOTING Philippe, Enjeux et perspectives des réseaux sociaux, Institut Supérieur du Commerce, Paris, mémoire de fin d'étude 2006

[8] Maria Mercanti Guérin.Analyse des reseaux sociaux et communautés en lign :Quelles applications en marketing ? 2010.

[9] " Réseaux sociaux académiques : fonctionnalités principales et enjeux(Academia, ResearchGate) ".

[10] « DeleteAcademicSocialNetworks ? Les réseaux sociaux académiques en 2016 ».

[11] Graphes, réseaux, réseaux sociaux : vocabulaire et notation Laurent Beauguitte

[12] Didier Müller .Introduction à la théorie des graphes. CRM,2011

[13] Romain Demangeon,Méthode Combinatoire pour le Calcul du Pfaffien

[14] Quelques rappels sur la théorie des graphes,iut lyon ,informatique,2011-2012

[15] Mathieu SABLIK.GRAPHE. 2008

[16] Ricco Rakotomalala , Détection de communautés dans les réseaux sociaux, Université Lyon 2.Octobre 2016

[17] Thèse de doctorat, Une approche de détection des communautés d'intérêt dans les réseaux sociaux : application à la génération d'IHM personnalisées Nadia Chouchani, 07/12/2018, à Valenciennes

[18] Newman M., Girvan M., "Finding and evaluating community structure in networks", Phys. Rev. E 69, 026113, 2004.

[19] Mariam Haroutunian, Karen Mkhitaryan , Josiane Mothe,Community Detection : Comparison of State of the Art Algorithms,2017

[20] C. BOTHOREL,et al , Clustering attributed graphs : models, measures and methods, 2015.

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

[22] Jean-Philippe Attal, Nouveaux algorithmes pour la détection de communautés disjointes et chevauchantes basés sur la propagation de labels et adaptés aux grands graphes. informatique .Université de Cergy Pontoise, 2017. Français.

[23] Auteurs : Raphaël Fournier-S'niehotta, Michel Crucianu, Marin Ferecatu

[24] Anaïs Correc , Comparaison d'algorithmes de détection de communautés dynamiques dans les réseaux évolutifs, Mémoire présenté en vue de l'obtention du grade de maîtrise ès sciences, Novembre 2015

[25] Mohamed Talbi, Une nouvelle approche de détection de communautés dans les réseaux sociaux, Août 2013

[26] Sneath (P. H. A.) et Sokal (R. R.). – Numerical Taxonomy - The Principles and Practice of Numerical Classification. – San Francisco, W. H. Freeman and Compagny

[27] Rushed Kanawati. Détection de communautés dans les grands graphes d'interactions (multiplexes) :état de l'art. 2013.

[28] Overlapping Community Detection in Networks : the State of the Art and Comparative Study1

[29] : Mael Canu, Détection de communautés orientée sommet pour des réseaux mobiles opportunistes sociaux, de l'Université Pierre et Marie Curie ,20 décembre 2017

[30] christian BELBEZE, agrégats de mots sémantiquement cohérents issus d'un grand graphe de terrain, DOCTORAT DE L'Université DE TOULOUSE ,2012

[31] NEDIOUI MOHAMED ABDELHAMID, fouille et apprentissage automatique dans les réseaux dynamique magister, université Mohamed khider-BISKRA ,2015

[32] Jean-Philippe Attal, Nouveaux algorithmes pour la détection de communautés disjointes et chevauchantes basés sur la propagation de labels et adaptés aux grands graphes, 'Université de Cergy Pontoise Spécialité : Sciences et Technologies de l'Information et de la Communication STIC,2017

[33] Sirinya ON-AT,Temporalité et réseaux sociaux : prise en compte de l'évolution dans la construction du profil utilisateur,DOCTORAT DE l'Université Toulouse 3 Paul Sabatier ,29/05/2017

[34] https ://www.gavagai.io/text-analytics/topic-modelling/ vu le 07/07/2020