République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université de Bordj Bou Arreridj

MEMOIRE

Présenté en vue l'obtention du diplôme de

MASTER en Informatique

Spécialité : Ingénierie de l'informatique décisionnelle

THEME

Application d'algorithmes d'apprentissage supervisé pour la prédiction d'interactions dans les réseaux complexes

Présenté par :

Khoudour Taous

Encadré par

Dr M. Charikhi

Promotion 2019/2020

Remerciement

D’abord et avant tout je remercie mon Dieu, d’avoir donné la force et la volonté pour réaliser ce modeste travail.

Mes vifs et sincères remerciements, accompagnés de toute ma gratitude vont à mon encadreur Dr. Charikhi Mourad, qui s’est toujours montré à l’écoute ainsi son précieux conseil et son aide durant toute la période du travail, pour le temps qu’il a consacré à m’apporter les outils méthodologiques indispensables à la conduite de cette recherche et dirigé dans mon travail, pour sa disponibilité et son avis éclairés.

J’exprime mes gratitudes aux mes parent pour leur soutien et patience. Je n’oublie pas mes amis et mes proches qui m’ont toujours soutenue et encouragé.

Merci à toute et à tous

Dédicaces

الى النموذج المثالي عن الحب و الانسانية .. الى امي و امي فوزية, الى ابي و ابي عبد الكريم .. دائما و ابدا

الى خالتي احلام , ملهمة الهجوم ,الصمود ,الألف محاولة ,عدم الاستسلام و خطي الدفاع الاول ..

الى اختي مريم و ايمان ,الى خالتي دنيا و أولادها, الى جدتي و جدي

الى اخي ايمن ,شد عضدي ,لكل انسان بديل الا انت

الى كل من مر من هنا و لكل روح دعمتني بدعائها

Table des matières

Liste des figures

Liste des tableaux

Abstract

Predicting interactions in complex networks is important for the analysis and exploration of network evolution.

After having made an overview of the existing methods, the mathematical formulation and the fields of application of this research theme, we implemented two types of solution for this problem and to apply them on three Datasets from different fields. Finally, we have presented the results of the experiments carried out.

Résumé

La prédiction des interactions dans les réseaux complexe est importante pour l'analyse et l’exploration de l'évolution des réseaux.

Après avoir faire un tour d’horizon sur les méthodes existées, la formulation mathématique et les domaines d’application de ce thème de recherche, nous avons implémenté deux types de solution pour ce problème ainsi de les appliquer sur trois Datasets de différents domaines. Enfin, nous avons exposé les résultats des expérimentations réalisées.

.

Introduction

L’utilisation des réseaux complexes est devenue une partie indispensable de la vie des entreprises, des laboratoires de recherches et des humains, ce qui est également une conséquence du développement rapide de la technologie numérique.

De nombreux systèmes de sociologie, de biologie ou d'information peuvent utiliser le réseau pour décrire l'état actuel des interactions entre ses éléments, dans lequel les nœuds représentent les individus et les arêtes représentent les relations entre ces individus. Par conséquent, l'étude des réseaux complexes a été une branche importante de nombreux domaines scientifiques.

Nous allons travailler dans le cadre de ce projet sur la prédiction de liens qui est une tâche importante dans l'analyse et l'exploration de liens « Link Mining ». La prédiction de lien donc, consiste à prédire s'il y aura des liens entre deux nœuds dans le futur en se basant sur des informations existantes observées.

La prédiction de lien peut non seulement être utilisée dans le domaine des réseaux sociaux pour aider les gens à trouver des nouveaux amis, mais peut également être appliquée dans d'autres domaines. Par exemple, il peut être appliqué aux systèmes de recommandation dans le e-commerce pour trouver de nouveaux collaborateurs potentiels, fournir des articles/produits intéressants dans les achats en ligne, recommander des partenaires de brevets dans les réseaux sociaux d'entreprise et les partenaires, trouver des experts ou des co-auteurs dans les réseaux sociaux universitaires et prédire les contacts des téléphones portables dans les réseaux de communication à grande échelle.

Elle peut également être utilisé pour déduire les réseaux complets sur la base de réseaux partiellement observés, mieux comprendre l'évolution des réseaux et prédire les hyperliens dans les réseaux sociaux hétérogènes.

Enfin, les techniques de prédiction des liens peuvent être appliquées en bio-informatique et en biologie, par exemple dans les réseaux de soins de santé et d'expression génique, en prédisant les spécialistes les plus susceptibles de recevoir de

futures références et en trouvant des interactions protéine-protéine. Même dans d'autres domaines tels que le domaine lié à la sécurité, il peut être utilisé pour identifier des communications anormales. Par conséquent, ces dernières années, de nombreux algorithmes ont été proposés pour résoudre le problème de la prédiction des liens.

Problématique

Considérons un réseau social $G\,(V,E)$ à un instant t particulier, où V et E sont des ensembles de nœuds et de liens, respectivement. La prédiction de liens vise à prédire de nouveaux liens entre les nœuds du réseau pour un temps futur t' avec $(t' > t)$.

Ce problème peut être expliqué en utilisant un simple ensemble de données d'un réseau social de sept personnes (nœuds) avec des relations d'amitiés représentants les liens entre les nœuds du graphe de la figure (Figure 1.a).

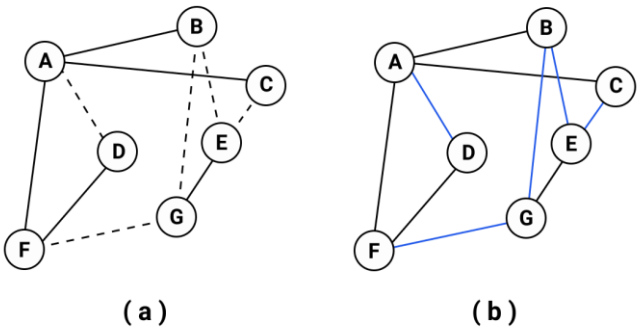

Figure 1 : Un exemple pour expliquer le problème de prédiction de liens

Le but de la prédiction de liens ici, est de prédire de nouvelles relations d'amitiés entre les personnes (nœuds) du réseau (Figure 1.b).

Objectifs du projet

Nous allons implanter deux types de solutions dans le cadre de ce projet. La première solution est basée sur les métriques de similarités. Nous testons plusieurs métriques de similarités sur différentes dataset pour obtenir des scores qui nous aide à faire la prédiction des futurs liens.

La deuxième solution est basée sur les algorithmes d'apprentissage supervisé. Nous allons utiliser deux algorithmes : *K Nearest Neighbors* ou K les plus proches voisins (KNN) et *Decision Trees* ou arbre de décisions

Enfin, nous faisons une évaluation des performances pour chaque méthode programmée sur les différentes dataset.

Organisation du rapport

Pour atteindre les objectifs de notre travail, on a organisé notre mémoire en trois chapitres :

Dans le premier chapitre, nous allons présenter la définition et les différents types existants ainsi que les concepts fondamentaux des réseaux complexes.

Le deuxième chapitre est consacré aux définitions importantes, ainsi que les différentes méthodes utilisées dans l'analyse des réseaux complexes. Nous introduisons le problème de la prédiction des liens et nous décrivons les principales approches pour résoudre ce problème.

Le dernier chapitre présente la conception, la réalisation et les expérimentations réalisées. Il expose les différents tests effectués et les résultats obtenus.

Enfin nous marquons l'accomplissement de notre travail par une conclusion générale suivi par des perspectives pour un travail futur.

CHAPITRE I

Introduction aux réseaux complexes

I.1 Introduction

Nous allons présenter dans ce premier chapitre, la définition d'un réseau complexe et ces différents types existants. Pour mieux comprendre Le problème traité dans ce projet de fin d'études, nous allons également décrire la modélisation d'un réseau en un graphe.

I.2 Définition

Un réseau est un ensemble de nœuds et de liens. Il est considéré comme un ensemble d'éléments en interactions mutuelles. Le nombre d'éléments ou de nœuds déterminent la taille d'un réseau. Dans un contexte de modélisation, la présence d'un lien entre deux nœuds indique une interaction entre les deux éléments correspondants du système. Pour représenter tout type de réseau, peu importe sa complexité on monde réel on utilise les graphes. [1]

Donc, un réseau complexe est représenté sous forme d'un graphe, où les nœuds représentent les acteurs / participants (individus, objets, organisations, etc.) et les liens correspondent aux interactions / relations entre les acteurs. Le graphe qui représente le réseau possède des caractéristiques topologiques non triviales, ses caractéristiques ne se produisent pas dans les réseaux simples mais se produisent dans la modélisation de graphe des systèmes réels. [2]

L'étude des réseaux complexes est un domaine nouveau et actif dans la recherche scientifique depuis les années 2000 [3]. Il est inspiré en grande partie par l'étude empirique des réseaux du monde réel tels que les réseaux d'informations, les réseaux biologiques, les réseaux technologiques, les réseaux sociaux, etc… [4]

I.3 Types des réseaux complexes

Les domaines concernés par les réseaux complexes sont aussi divers et nombreux tel que la biologie, la sociologie, la psychologie et l'informatique. Nous allons introduire dans les sections ci-dessous les deux types les plus traités dans la littérature scientifique.

I.3.1 Réseaux sociaux

Un réseau social est une structure sociale composée d'un ensemble d'acteurs sociaux (des individus ou des organisations) et des ensembles de liens dyadiques ainsi que d'autres interactions sociales entre les acteurs.

La perspective des réseaux sociaux fournit un ensemble de méthodes pour analyser la structure d'entités sociales entières on utilisant une variété de théories expliquant les modèles observés dans ces structures.

L'étude de ces structures utilise l'analyse des réseaux sociaux (SNA) pour identifier les modèles locaux et mondiaux, localiser les entités influentes et examiner la dynamique des réseaux.

Le réseau social est une construction théorique utile en sciences sociales pour étudier les relations entre individus, groupes, organisations ou même des sociétés entières.

En sciences sociales, les domaines d'études comprennent, sans s'y limiter, l'anthropologie, la biologie, les études en communication, l'économie, la géographie, les sciences de l'information, les études organisationnelles, la psychologie sociale, la sociologie et la sociolinguistique. [5]

Les réseaux d'information sont une généralisation de l'espace de réseaux, mais le concept des réseaux sociaux ne se limite pas au cas particulier d'un réseau social basé sur internet tel qu'Instagram, TikTok, Twitter... Etc.

Il y'a deux catégorie de réseaux sociaux :

- **Réseaux sociaux humains** : indique un regroupement de personnes ou d'organisations qui communiquent et échangent leurs idées et interagirent entre eux
- **Réseaux sociaux sur internet** : où dit les médias sociaux sont des technologies interactives informatisées qui facilitent la création ou le partage d'informations, d'idées, d'intérêts professionnels et d'autres formes d'expression via des communautés et des réseaux virtuels. [6][7]

  La variété des services de médias sociaux autonomes et intégrés en évolution rend difficile leur définition. [6] Cependant, les experts en marketing et en médias sociaux conviennent généralement que les médias sociaux incluent les 13 types de médias sociaux suivants: blogs, réseaux d'entreprises (business networks), projets collaboratifs (collaborative projects), réseaux sociaux d'entreprise (entreprise social networks), forums, microblogs, partage de photos (photo sharing), Revue des produits / services (products/services review), référencement social (social bookmarking), jeux sociaux (socail gaming), réseaux sociaux (social networks), partage de vidéos (video sharing), et mondes virtuels (virtual worlds). [8]

  Chaque utilisateur se crée un profile/compte (nœud). Chaque utilisateur peut publier (des liens, texte, images, story/snap, vidéo...etc.) ou interagi avec une publication ou avec d'autre utilisateur (j'aime, abonnement, repartage, commentaire, mention d'autre utilisateur, tag, etc…)

En sciences social, les réseaux sociaux sont fréquemment étudiés sous l'angle des interactions génériques entre tous les acteurs, où ses interactions peuvent être dans n'importe quelle forme conventionnelle. [9]

I.3.2 Réseaux biologiques

Les réseaux sont largement utilisés dans de nombreuses branches de la biologie comme une représentation pratique des modèles d'interaction entre les éléments biologiques appropriés. Ces réseaux biologiques comprennent les réseaux biochimiques, les réseaux de neurones et les réseaux écologiques.

Les réseaux biochimiques représentent les modèles d'interaction au niveau moléculaire et les mécanismes de contrôle dans la cellule biologique. Les principaux types de ces réseaux sont les réseaux métaboliques, les réseaux protéine-protéine et les réseaux de régulation génétique [10]. Figure 2 représente un sous-réseau métabolique biochimique

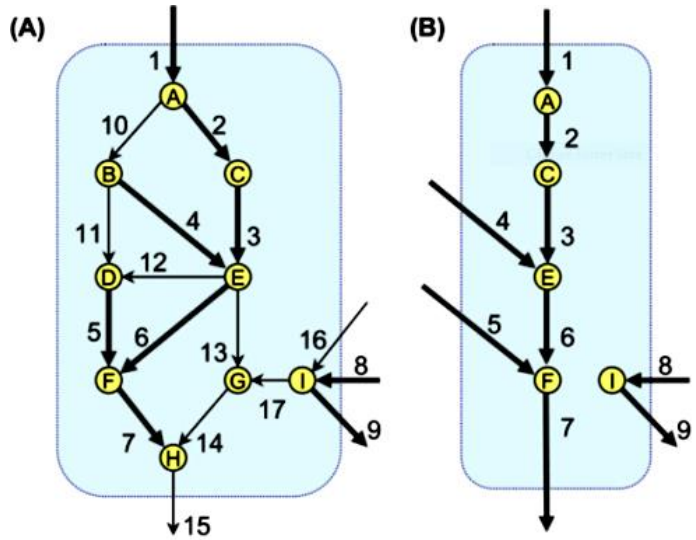

Figure 2 : *Un exemple de sous-réseau métabolique. (A): Un petit réseau métabolique avec 17 réactions. Les métabolites sont représentés par des nœuds, tandis que les réactions sont représentées par des flèches. Les réactions 1, 8, 9, 15 et 16 sont des réactions aux limites, tandis que toutes les autres réactions sont des réactions internes. Nous pourrions être intéressés uniquement par un sous-réseau contenant neuf réactions: 1...9, qui sont représentés par des flèches épaisses. Ce sous-réseau sera appelé SuN. (B): Le sous-système réduit ne comprenant que les neuf réactions intéressantes. [11]*

Les réseaux de neurones : une des principales fonctions du cerveau est de traiter l'information et l'élément principal de traitement de l'information est le neurone, une cellule cérébrale spécialisée qui combine plusieurs entrées pour générer une seule sortie.

Un neurone typique se constitue d'un corps cellulaire ou soma, ainsi que d'un certain nombre de tentacules saillants, appelés dendrites, qui sont des fils d'entrée pour transporter des signaux dans la cellule. La plupart des neurones n'ont qu'une seule sortie, appelée axone, qui est généralement plus longue que les dendrites. Il se ramifie généralement près de son extrémité en terminaux axonaux pour permettre à la sortie de la cellule d'alimenter l'entrée de plusieurs autres. Il existe un petit espace, appelé synapse, entre le terminal et la dendrite à travers lequel le signal de sortie du premier neurone (présynaptique) doit être acheminé pour atteindre le second neurone (postsynaptique) [10].

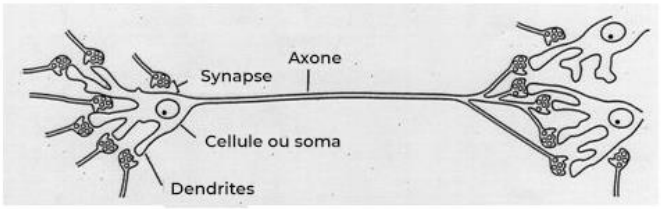

Figure 3 : un exemple sur le réseau neurone dans un cerveau humain

Les réseaux écologiques sont des réseaux d'interactions écologiques entre espèces. Les espèces d'un écosystème peuvent interagir de différentes manières: elles peuvent se manger les unes les autres, elles peuvent se parasiter les unes les autres, ou elles peuvent avoir une variété d'interactions mutuellement avantageuses, telles que la pollinisation ou la dispersion des graines [10]

I.4 Terminologie et notation

Considérons un réseau complexe $G\,(V,E)$ à un instant t particulier, où V et E sont respectivement des ensembles de nœuds et des liens. La prédiction de liens vise à prédire de nouveaux liens ou des liens supprimés entre les nœuds pour un temps futur t' avec $t' > t$ dans le réseau. Par exemple, nous utilisons un simple ensemble de données (dataset) d'un réseau social sur sept (7) personnes (nœuds), nous le structurons sous forme d'un graph, des liens existants solides indiquent que des interactions existaient déjà au temps t et des liens en pointillés indiquent des liens qui sont apparus récemment pendant l'intervalle de temps $[t, t']$.

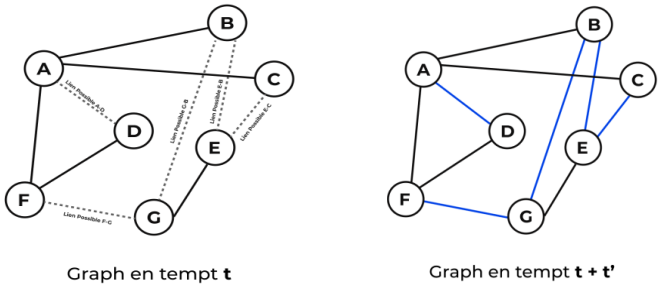

Figure 4 : Un exemple pour expliquer le problème de prédiction de lien.

Le but du problème de prédiction de lien ici est de prédire de nouveaux liens entre les liens nouvellement apparus entre les personnes (nœuds). Après avoir analysé les données, nous obtenons le graph à droite, de nouvelles connexions ont été formées.

I.5 Conclusion

Dance ce chapitre, nous avons défini qu'est-ce qu'un réseau complexe, nous avons présenté les différents types des réseaux ainsi qu'une terminologie sur la modélisation d'un réseau en un graphe. Dans le chapitre suivant nous présentons l'état de l'art relatif à la prédiction des liens.

CHAPITRE II

Prédiction des interactions dans les réseaux complexes

II.1 Introduction

Dans ce chapitre, nous allons tout d'abord introduire le problème de la prédiction d'interactions dans les réseaux complexes par une formulation mathématique. Ensuite, nous allons expliquer les différentes solutions présentes dans la littérature. Nous allons donc, présenter les mesures de similarité basées sur les nœuds ou basées sur les chemins, les modèles probabilistes et les approches discriminantes. Enfin, nous décrivons les domaines d'applications de ce projet de fin d'études ainsi que les orientations et les défis futurs.

II.2 Formulation mathématique

Un graphe ou réseau G est une paire ordonnée $G = (V, E)$ où V est un ensemble de sommets ou nœud éventuellement étiquetés et E est un ensemble de liens entre des paires d'éléments de l'ensemble V. Un lien entre deux nœuds x et y est noté $e_{x,y}$. Le nombre de nœuds dans le réseau, également appelé taille du réseau, est noté $|V|$. Le nombre de liens est noté $|E|$.

Nous pouvons distinguer les liens dirigés (notés comme des arcs), qui connecte un nœud source à un nœud de destination, et les liens non dirigés (notés comme des arêtes), lorsqu'il n'y a pas de concept de source et de destination. Un graphe orienté est composé uniquement d'arcs. Un graphe non orienté est composé uniquement d'arêtes. Un graphique mixte peut contenir les deux types de liens (arcs et arêtes).

L'ensemble des nœuds connectés via un lien à un nœud $x \in V$ est appelé le voisinage de x et est noté Γx.

Dans les graphes non orientés, le degré d'un nœud x est défini comme le nombre d'arêtes connectées au nœud et sera noté $|\Gamma x|$.

Dans les graphes dirigés, le degré d'un nœud est la somme du degré sortant et du degré intérieur, qui sont respectivement le nombre d'arcs sortants et les arcs entrants. Le degré moyen d'un réseau est noté $\langle \Gamma \rangle$ et est égal au degré moyen de tous ses nœuds. Une boucle est une arête ou un arc reliant un nœud à lui-même. Un graphe simple est défini comme un graphe sans boucles et avec pas plus d'une arête ou d'un arc entre chaque paire de sommets.

Un chemin est une séquence de liens qui connecte une séquence de nœuds dans le graphe. Dans les graphes dirigés, les étapes du chemin sont limitées pour se déplacer du nœud source vers le nœud de destination du même arc. La longueur du chemin est le nombre de liens dans le chemin. Le chemin

le plus court entre deux sommets est le chemin avec la plus petite longueur entre ces sommets.

Un graphe est appelé connecté s'il existe un chemin entre chaque paire de nœuds $x, y \in V$. Si le graphe n'est pas connecté, il est composé de composants. Un composant est un sous-graphe connecté. Un graphe connecté n'a qu'un seul composant. Si l'un des composants a un nombre de nœuds significativement plus grand que les autres composants, il est généralement appelé composant principal ou géant.

II.3 Les méthodes de prédictions de liens

.Pour résoudre le problème de prédiction des liens, nous déterminons les possibilités de formation ou de décomposition des liens entre les paires de nœuds. On peut mesurées ces possibilités par des similitudes *Figure 5* ou des rangs relatifs (Relative Ranks) entre les paires de nœuds. Nous illustrons les solutions du problème de prédiction de liens dans la Figure 6 [12]

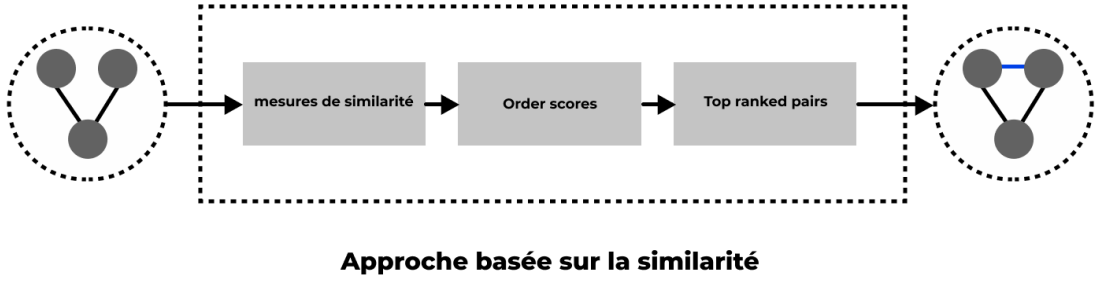

Figure 5 : Illustration des solutions du problème de prédiction de liens on utilisant une approche basée sur la similarité

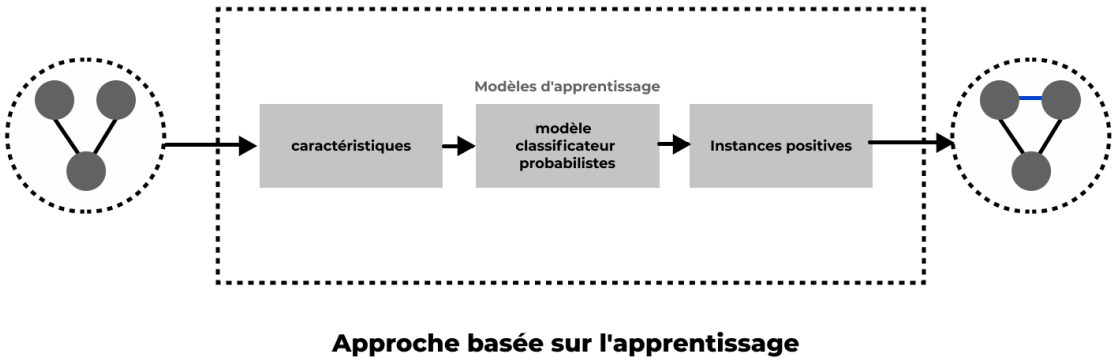

Figure 6 : Illustration des solutions du problème de prédiction de liens en utilisant une approche basée sur la l'apprentissage

Il existe plusieurs approches de prédiction des liens, notamment, les approches basées sur la similarité et les approches basées sur l'apprentissage.

Une approche basée sur la similarité consiste à calculer la similarité sur des paires de nœuds non connectés dans un réseau. Elle est basée sur des mesures pour analyser la proximité des nœuds. Pour chaque paire de nœuds potentiels (x, y) on attribue un score, où le score le plus élevé signifie une probabilité plus élevée que x et y soient liés à l'avenir, et vice versa. Ensuite, une liste classée par ordre décroissant de scores est obtenue et les liens en haut de la liste sont les plus susceptibles d'apparaître.

Une approche basée sur l'apprentissage traite le problème de prédiction de lien comme une tâche de classification binaire. Certains modèles d'apprentissage automatique tels que le classificateur et le modèle probabiliste peuvent être utilisés pour résoudre ce problème. Chaque paire de nœuds non connectés correspond à une instance avec des caractéristiques décrivant les nœuds et l'étiquette de classe. S'il existe un lien reliant une paire de nœuds, cette paire est étiquetée comme positive, sinon elle est négative. Pour les approches basées sur l'apprentissage, les fonctionnalités se composent de deux parties : l'une est les fonctionnalités de similarité des approches basées sur la similarité, l'autre est les fonctionnalités dérivées du réseau complexe. [12]

II.3.1 Mesures de similarités

Il existe de nombreuses métriques de prédiction de liens simples et basiques, qui utilisent les informations des nœuds, la topologie et la théorie sociale pour calculer les similitudes des paires de nœuds. Les métriques complexes sont établies sur des fonctionnalités fournies par les métriques de base et les informations externes

II.3.1.1 Métriques basées sur les nœuds

Le calcul de la similarité est basé sur une idée simple : plus la paire est similaire, plus il y aura de lien entre elles et vice versa. Par exemple, les utilisateurs des réseaux sociaux ont tendance à créer des relations avec des personnes qui sont similaires en termes d'éducation, de religions, d'intérêts et de lieux.

Pour chaque paire de nœuds non connectés (x, y) on l'attribue un score signifiant la similarité entre x et y. Un score élevé indique une forte probabilité que x et y seront liés au futur et vice versa, puis, on utilise le *rang des scores de similarité*, pour prédire les liens.

Les métriques basées sur les nœuds utilisent les attributs (informations) et les actions des nœuds, qui peuvent refléter les intérêts personnels et les comportements sociaux dans un réseau social, pour calculer les similitudes entre les paires de nœuds.

II.3.1.2 Métriques basées sur les typologies

Dans un réseau simple sans attributs de nœud ou des liens, il existe de nombreuses métriques pour calculer la similitude entre deux nœuds, La plupart de ces métriques sont basées sur les informations.

Selon les caractéristiques de ces métriques basées sur la typologie, elles peuvent être divisées en métriques basées sur le voisin (neighbor), métriques basées sur le chemin (path) et métriques basées sur la marche aléatoire (random walk).

D'abord, soit la matrice A la matrice de contiguïté d'un réseau social donné. Soit $\Gamma(x)$ l'ensemble des voisins du nœud x, et soit $|\Gamma(x)|$ être le nombre de voisins du nœud x.

II.3.1.2.1 Métriques basées sur les voisins

Les gens se joignent les réseaux sociaux pour crée de nouvelles relations avec des personnes ou maintenir une relation avec ses proches. Les voisins sont les plus proches d'un utilisateur donné. On utilise les métriques basées sur les voisins pour la prédiction de liens.

Voisins communs (Commun Neighbors) :

La similitude entre deux nœuds est le nombre de voisins partagés entre les deux nœuds [13]. Il est logique de supposer que, si deux personnes partagent de nombreuses connaissances, elles sont plus susceptibles de se rencontrer que deux personnes sans contacts communs. Cette méthode est formalisée par :

$$CN\,(x,y) = |\,\Gamma x \cap \Gamma y\,| \quad \text{.....(II.1)}$$

Malgré sa simplicité, cette mesure fonctionne étonnamment bien sur la plupart des réseaux du monde réel et bat des approches très complexes.

L'indice Jaccard (JA) :

JA mesure le rapport des voisins partagés dans l'ensemble complet des voisins pour deux nœuds. Cette méthode est encore une autre variante de la méthode des voisins communs (CN) où il y a une pénalisation pour chaque voisin non-partagé. Cette fonction de similarité est définie comme

$$JA\,(x,y) = \frac{|\Gamma x \cap \Gamma y\,|}{|\,\Gamma x \cup \Gamma y\,|} \quad \text{.....(II.2)}$$

L'indice Adamic-Adar (AA) :

La mesure AA est formulée en fonction du coefficient de Jaccard. Mais ici, les voisins communs qui ont moins de voisins sont plus pondérés, cela peut être écrit comme

$$AA\ (x,y)\ =\ \sum_{z\ \in\ \Gamma x\ \cap\ \Gamma y}\frac{1}{log\ |\ \Gamma z\ |} \quad \ldots\ldots\ldots\ldots\ldots\ldots\text{(II.3)}$$

L'indice d'attachement préférentiel (PA) :

La métrique PA indique que les nouveaux liens seront plus susceptibles de se connecter des nœuds de degré supérieur que ceux inférieurs La similitude entre deux nœuds est formulée par :

$$PA\ (x,y)\ =\ |\ \Gamma x\ ||\ \Gamma y\ | \quad \ldots\ldots\ldots\ldots\ldots\ldots\text{(II.4)}$$

L'indice d'allocation des ressources (RA) :

Cet indice modélise la transmission d'unités de ressources entre deux nœuds non connectés x et y via des nœuds de voisinage. Chaque nœud de voisinage obtient une unité de ressource de x, puis il la distribue d'une manière égale à ses voisins. La quantité de ressources obtenue par le nœud y peut être considérée comme la similitude entre les deux nœuds. Cette fonction de similarité est formulée comme

$$RA\ (x,y)\ =\ \sum_{z\ \in\ \Gamma x\ \cap\ \Gamma y}\frac{1}{\Gamma z} \quad \ldots\ldots\ldots\ldots\ldots\ldots\text{(II.5)}$$

Naïve Bayes local (LNB) :

Cette méthode suppose que chaque voisin partagé a un rôle ou un degré d'influence différent, qui peut être estimé à l'aide de la théorie des probabilités [14]. Cette méthode estime la similitude de deux nœuds comme

$$LNB\ (x,y)\ =\ \sum_{z\ \in\ \Gamma x\ \cap\ \Gamma y}\ f(z)\ log(oRz) \quad \ldots\ldots\ldots\ldots\ldots\ldots\text{(II.6)}$$

o est une constante pour le réseau : $o\ =\ \dfrac{Punconnected}{Pconnected}$

$$=\ \frac{\frac{1}{2}\ |\ V\ |\ (|\ V\ |\ -\ 1)}{|\ E\ |}\ -\ 1$$

Rz est le rôle ou l'influence du nœud: $Rz\ =\ \dfrac{2\ \left|\left\{e_{\mathrm{x,y}}\colon\ \mathrm{x,y}\ \in\ \Gamma_z, e_{\mathrm{x,y}}\ \in\ \mathrm{E}\right\}\right|\ +\ 1}{2\ \left|\left\{e_{\mathrm{x,y}}\colon\ \mathrm{x,y}\ \in\ \Gamma_z, e_{\mathrm{x,y}}\ \notin\ \mathrm{E}\right\}\right|\ +\ 1}$

$f(z)$ est une fonction qui mesure l'influence du nœud. $f(z) = 1$ à partir de voisins communs (CN), $f(z) = \frac{1}{\log|\Gamma z|}$ à partir de l'indice Adamic-Adar (AA), ou $f(z) = \frac{1}{|\Gamma z|}$ de la méthode d'allocation des ressources.

L'indice Sørensen (SO) :

Malgré sa similitude avec l'indice de Jaccard (JA), il est moins sensible aux valeurs aberrantes [15]. La similitude de Sørensen est définie comme

$$SO(x,y) = 2|\Gamma x \cap \Gamma y|\,|\Gamma x| + |\Gamma y| \quad \text{.....(II.7)}$$

L'index promu par le hub (Hub Promoted Index | HPI) :

HPI définit le chevauchement topologique des nœuds x et y [16], la valeur HPI est déterminée par le degré inférieur de nœuds. L'objectif, cette mesure de similarité évite la formation de liens entre les nœuds de hub et de promouvoir la formation de liens entre les nœuds de bas degré et les hubs [17]. Cet indice définit la similitude comme

$$HPI(x,y) = \frac{|\Gamma x \cap \Gamma y|}{min(|\Gamma x|,|\Gamma y|)} \quad \text{.....(II.8)}$$

L'indice déprimé du hub (Hub Depressed Index | HDI) :

Zhou et coll. [32] propose une mesure similaire à HP, mais la valeur est déterminée par les plus hauts degrés de nœuds. Cette fonction de similarité peut être définie comme

$$HDI(x,y) = \frac{|\Gamma x \cap \Gamma y|}{max(|\Gamma x|,|\Gamma y|)} \quad \text{.....(II.9)}$$

Cet indice est basé sur l'indice promu par le hub mais a un objectif opposé [17].

Dans le $tableau$ 1, nous comparons les métriques populaires basées sur les voisins en termes de complexité temporelle et de caractéristiques. La complexité temporelle est un facteur important pour la sélection des métriques, en particulier pour les réseaux sociaux à grande échelle. Supposons que le nombre moyen de voisins dans un réseau est n, pour deux nœuds x et y,

Tableau 1: Comparaison des métriques populaires basées sur les voisins en termes de complexité temporelle et de caractéristiques

Métrique	Complexité Temporelle	Caractéristique	Référence
CN	$O(n^2)$	Simple et intuitive	[13]
JA	$O(2n^2)$	Proportion de voisins communs par rapport au nombre total de voisins	[18]
AA	$O(2n^2)$	Les voisins communs ayant moins de voisins sont pondérés plus lourdement	[19]
PA	$O(2n)$	Les liens simples et nouveaux seront plus susceptibles de connecter des nœuds de niveau supérieur	[20]
RA	$O(2n^2)$	Similaire aux AA, mais punit plus sévèrement les voisins communs de haut niveau	[21]
LNB	$O(O(f(z)) + 2n^2)$	Chaque voisin partagé a un rôle ou un degré d'influence différent, qui peut être estimé à l'aide de la théorie des probabilités	[22]
SO	$O(2n^2)$	Des degrés inférieurs de nœuds auraient une plus grande probabilité de liaison	[23]
HPI	$O(2n^2)$	La probabilité de lien est déterminée par le degré inférieur de nœuds	[17]
HDI	$O(2n^2)$	La probabilité de lien est déterminée par le degré le plus élevé de nœuds	[17]

Cette comparaison peut aider les gens à choisir des mesures appropriées pour les réseaux sociaux pratiques.

II.3.1.2.2 Métriques basées sur le chemin

Les chemins entre deux nœuds peuvent être utilisés pour calculer les similitudes des paires de nœuds.

Chemin local (LP - Local Path):

La métrique LP utilise les informations des chemins locaux de longueur 2 et de longueur 3. Les chemins de longueur 2 sont plus pertinents que les chemins de longueur 3, il y a un facteur d'ajustement α appliqué dans la mesure α est le petit nombre proche de 0. La métrique est formalisée par :

$$P = A^2 + \alpha A^3 \quad \text{.....(II.10)}$$

Ici, A^2 et A^3 sont des matrices d'adjacence des nœuds ayant respectivement 2 longueurs et 3 distances de longueur. Par conséquent, LP est

également une matrice d'adjacence qui décrit les paires de nœuds avec des distances de longueur 2 et 3.

Katz :

La métrique de Katz résume l'influence de tous les chemins possibles entre deux paires de nœuds, pénalisant progressivement les chemins par leur longueur [24]. Katz est définie par :

$$Katz(x,y) = \sum_{i=1}^{\infty} \beta^l . |Path_{x,y}^l| = \beta A + \beta^2 A^2 + \beta^3 A^3 + \cdots \quad \ldots\ldots\text{(II.11)}$$

$Path_{x,y}^l$ est l'ensemble de tous les chemins de x à y avec une longueur l.

$$Non - Pondéré : \begin{cases} |Path_{x,y}^l| = 1, & si\ il\ ya\ une\ collaboration\ x\ et\ y \\ 0, & sinon \end{cases}$$

$$Pondéré : \{Path_{x,y}^l \ le\ nombre\ de\ fois\ de\ la\ collaboration\ entre\ x\ et\ y$$

$\beta > 0$ et $\beta \leq 1$ β causera la métrique Katz d'agir un peu comme la métrique CN, c'est un paramètre de régularisation, il punis la contribution des chemins longs dans le calcul de similarité.

Friend Link (FL) :

La métrique FL est une similitude entre les nœuds x et y, en parcourant tous les chemins d'une longueur bornée. Elle fournit une prédiction de lien plus précise et plus rapide. FL suppose que les personnes d'un réseau social, par exemple, peuvent utiliser tous les chemins entre eux. La similitude entre x et y est définie comme le nombre de chemins de longueur variable l de x à y:

$$FL(x,y) = \sum_{i=1}^{l} \frac{1}{i-1} \cdot \frac{|paths_{x,y}^i|}{\prod_{j=2}^{i}(n-j)} \quad \ldots\ldots\ldots\ldots\ldots\text{(II.12)}$$

Ici, n est le nombre de sommets dans le réseau. l est la longueur d'un chemin entre x et y. Les chemins $paths_{x,y}^i$ est l'ensemble de toutes les $length - 1$ chemins de x à y.

II.3.1.2.3 Métriques basées sur la marche aléatoire (Random walk based metrics)

Les interactions entre les nœuds des réseaux complexes peuvent également être modélisées par une marche aléatoire, qui utilise des probabilités de transition d'un nœud à ses voisins pour désigner la destination d'un marcheur aléatoire à partir du nœud actuel.

SimRank :

Deux nœuds sont similaires s'ils sont connectés ou similaires aux nœuds similaires. Il existe un paramètre γ qui contrôle la vitesse à laquelle le poids des nœuds connectés diminue à mesure qu'ils s'éloignent des nœuds d'origine. L'équation de SimRank :

$$SimRank\ (x,y) = \begin{cases} 1, & x = y \\ \gamma.\frac{\sum_{a \in \Gamma(x)} \sum_{b \in \Gamma(y))} \text{SimRank(a,b)}}{|\Gamma(x)|\cdot|\Gamma(y)|}, & sinon \end{cases} \quad \text{.....(II.13)}$$

PageRank Enraciné (Rooted PageRank - RPR):

RPR est une modification du PageRank. Le rang d'un nœud dans le graph est proportionnel à la probabilité que le nœud soit atteint par une marche aléatoire sur le graphique.
Le facteur ϵ qui spécifie la probabilité que l'algorithme visite les voisins du nœud au lieu de recommencer. Soit D une matrice diagonale avec $Di,i = \sum_j A_{i,j}$. La mesure est définie comme :

$$RPR = (1 - \epsilon)(I - \epsilon D^{-1}A)^{-1} \quad \text{.....(II.14)}$$

II.3.1.3 Métriques basées sur la théorie sociale

On peut prédire les liens en utilisant la théorie social, telles que la communauté, la fermeture triadique, les liens forts et faibles, l'homophilie et l'équilibre structurel, pour résoudre les problèmes d'exploitation et d'analyse des réseaux sociaux. Les métriques basées sur la théorie sociale peuvent améliorer les performances en capturant des informations d'interaction sociale, en particulier pour les réseaux sociaux à grande échelle. [25]

II.3.2 Approches discriminantes

Les approches discriminantes apprend une fonction de décision a associé une paire de nœuds une étiquette qui indique l'absence ou la présence de lien entre ces deux noeuds.

II.3.2.1 Apprentissage supervisé

M est un ensemble de nœuds, pour chaque nœud x est défini par un vecteur $\vartheta(x)\epsilon R^p$, on note E l'ensemble des liens $e_1, \dots, e_n$. Dans le cas de l'inférence supervisée, la phase d'apprentissage d'un ensemble $S = \{(e_1, l_1), \dots, (e_n, l_n)\}$ les paires de nœuds sont liée avec des étiquettes $l_i \in \{0,1\}$, qui déterminent l'existence du lien entre les nœuds de e_i.

II.3.2.2 Méthodes d'apprentissage

Pour prédire les liens manquants ou futurs, nous présentons quelques algorithmes du modèle supervisé :

Arbre de décision (Decision Trees) :

Les arbres de décision sont une méthode d'apprentissage supervisé utilisée pour la classification et la régression. Elles apprennent à partir des données pour approcher une courbe sinusoïdale avec un ensemble de règles de décision si-alors-sinon (if-then-else). Plus l'arbre est profond, plus les règles de décision sont complexes et plus le modèle est en forme.

L'arbre de décision construit des modèles de classification sous la forme d'une structure arborescente. Elle décompose un ensemble de données en sous-ensembles de plus en plus petits tout en développant en même temps un arbre de décision associé. Le résultat final est un arbre avec des nœuds de décision et des nœuds feuilles. Un nœud de décision à deux branches ou plus. Le nœud feuille représente une classification ou une décision. Le nœud de décision le plus élevé dans un arbre qui correspond au meilleur prédicteur appelé nœud racine.

KNN (K Nearest Neighbors (K le plus proche voisin)) :

L'algorithme K plus proche voisin est très simple. Il fonctionne en fonction de la distance minimale entre le lien que nous voulons prédire et la base de données d'apprentissage pour déterminer les K voisins les plus proches.

Après avoir rassemblé les K voisins les plus proches, nous prenons la majorité simple de ces K voisins les plus proches pour être les futurs liens.

Les données de l'algorithme KNN peut être constituées de plusieurs attributs multivariés X_i qui seront utilisés pour classer $Y \in [0,1]$.

Supposons que nous déterminions K = 3 (nous utiliserons 3 voisins les plus proches) comme paramètre de cet algorithme. Ensuite, nous calculons la distance entre le lien que nous voulons prédire et tous la dataset d'apprentissage.

L'étape suivante consiste à trouver les K voisins les plus proches. Nous incluons la base de données d'apprentissage en tant que plus proche voisin si la distance de cette dataset d'apprentissage à le lien que nous voulons prédire est inférieure ou égale à la K-ième plus petite distance. En d'autres termes, nous trions la distance de tous les la base de données d'apprentissage à le lien que nous voulons prédire et déterminons la K-ième distance minimale.

Si la distance de la dataset d'apprentissage est inférieure au K-ième minimum, alors nous rassemblons la catégorie Y de la dataset d'apprentissage de ce voisin le plus proche.

II.4 Évolutivité de la prédiction de liens

L'évolutivité et l'efficacité sont importantes pour les réseaux sociaux massifs du monde réel.

Une Deep Learning Framework appelé la machine Boltsmann à restriction temporelle conditionnelle, a été proposé par [26], qui prédit des liens basés sur la variance de transition individuelle et l'influence introduite par les voisins locaux.

Une prédiction de lien, a été offert par [27], non paramétrique pour les réseaux dynamiques dans lesquels leur modèle peut accueillir des régions avec des profils d'évolution très différents, autrement impossible par la métrique ou l'heuristique de prédiction de lien. Il permet également un apprentissage basé à la fois sur les fonctionnalités topologiques et sur d'autres fonctionnalités disponibles en externe. Ils ont également adapté l'algorithme de hachage sensible à la localité pour résoudre l'évolutivité de la prédiction de lien dans les grands réseaux et les séquences de longue durée.

II.5 Domaine d'application de prédiction de liens :

Dans les réseaux sociaux, la prédiction de lien peut être utilisée pour diverses applications ; ici, nous aborderons quelques applications typiques.

Système de recommandation : Recommander des partenaires, des amis et des abonnés est une application typique pour la prédiction de liens.

Trouver des relations réciproques : Dans les réseaux sociaux, une relation bidirectionnelle (également appelée réciproque), généralement est développée à partir d'une relation unidirectionnelle, représente une relation plus confiante entre les personnes. Comprendre la formation des relations bidirectionnelles peut nous donner un aperçu de la dynamique au niveau micro du réseau social, telle que la structure communautaire sous-jacente et l'influence des utilisateurs les uns sur les autres.

Trouver des experts et des collaborations dans le réseau social académique : Les réseaux sociaux universitaires contiennent des quantités massives d'experts dans divers domaines et il est difficile pour le chercheur individuel de décider quels experts correspondront le mieux à sa propre expertise.

Une méthode a été proposée par [28] pour construire des prédicteurs de lien dans les réseaux sociaux académiques, il utilise une méthode d'apprentissage supervisé pour prédire les liens à partir d'attributs structurels du réseau.

Dans un réseau de chercheurs, en cherche les attributs structurels du graph des collaborations passées ainsi que les caractéristiques sémantique et les

fonctionnalités basées sur les évènements, pour former un ensemble de prédicteurs à l'aide d'algorithmes d'apprentissage supervisé, ces prédicteurs peuvent être utilisés pour prédire les liens futurs. Les collaborations interdisciplinaires sont précieuses dans la société humaine. L'établissement de collaborations inter-domaines est difficile pour certaines raisons :

- Les collaborations inter-domaines sont rares ;
- Les collaborateurs inter-domaines ont souvent des compétences et des intérêts différents ;
- La collaboration intersectorielle se concentre sur un sous-thème.

Par conséquent, les collaborations inter-domaines ont des modèles différents par rapport aux collaborations traditionnelles dans le même domaine.
En analysant les réseaux de collaboration inter-domaines, [29] ont proposé le modèle Crossdomain Topic Learning (CTL) pour résoudre les défis ci-dessus :

- Pour la gestion de connexions clairsemées, à travers les couches du thème au lieu de couches d'auteur, CTL consolide les collaborations inter-domaines existants
- Pour gérer l'expertise complémentaire, CTL modélise séparément les distributions des thermes à partir des domaines source et cible, ainsi que la corrélation entre les domaines.
- Pour gérer l'asymétrie des thèmes, CTL ne modélise que les thèmes pertinents pour la collaboration inter-domaines. La prédiction des liens entre co-auteurs est un problème fréquemment étudié.

Des chercheurs [30] ont étudié le problème de la prédiction des relations de co-auteurs dans le réseau bibliographique hétérogène, et une nouvelle méthodologie appelée PathPredict basée sur un modèle de prédiction de relations de méta-chemins a été proposée pour résoudre ce problème, les caractéristiques topologiques basées sur les méta-chemins sont systématiquement extraites du réseau, un modèle supervisé est utilisé pour apprendre les meilleurs poids associés aux différentes caractéristiques topologiques pour trouver des relations de co-auteur.

Prédiction du lien social : Lorsqu'un réseau social change dynamiquement, les liens sociaux évoluent avec le temps. Les forces du lien social sont différentes les unes des autres même si elles font partie du même groupe. [31] ont étudié l'évolution des relations sociales de personne à personne, et la prédiction des forces des liens sociaux sur une dataset des appels des téléphones mobiles. Ils proposent un modèle d'affinité pour quantifier les forces du lien social dans lequel un indice de réciprocité est intégré pour mesurer le niveau de réciprocité entre les utilisateurs et leurs partenaires de communication. Il y a des travaux

qui concentrent sur la force des relations dans les réseaux sociaux, comme la prédiction de la force des liens avec les médias sociaux [32] et la modélisation de la force des relations dans les réseaux sociaux en ligne [33]. Bien que ces travaux ne soient pas directement liés à la prédiction de liens, leurs conclusions et résultats peuvent être étendus à de nouvelles méthodes de prédiction de liens.

II.6 Orientations et défis futurs

Il existe encore de nombreux défis futurs potentiels, certains nouveaux problèmes ouverts qui nécessitent une étude plus approfondie. Ici, nous abordons certains défis de recherche futurs possibles sur le problème de prédiction de lien.

- **Prédiction de lien disparaissant.** La plupart des travaux de prédiction de liens existants se concentrent sur les liens qui apparaîtront dans le futur, seuls quelques travaux traitent de la prédiction de liens qui vont être disparaitre dans le futur [34]. Ce problème il n'est pas facile de résoudre mais également très important. On note que la prédiction des liens qui disparaisse n'est pas l'inverse du problème la prédiction des liens manquants ou apparent, parce que le mécanisme de la dissolution du lien n'est pas le même que le mécanisme de la formation du lien. Par conséquent, nous ne pouvons pas appliquer directement les méthodes de prédiction de liens pour prédire les liens qui disparaissent.

- **Prédiction de lien sous des nœuds dynamiques.** Considérons cette hypothèse : les nœuds des réseaux sociaux sont connus et ne changeront pas à l'avenir. Mais, cette hypothèse ne peut être satisfaite dans les cas pratiques. Les réseaux sociaux sont très dynamiques, un nœud peut quitter ou rejoindre le réseau. Dans les réseaux sociaux, il y a plusieurs utilisateurs ne sont jamais actifs après un certain temps, ces utilisateurs ne doivent pas être pris en compte dans la prédiction de liens car ils ont effectivement quitté les réseaux sociaux. On considère aussi le cas des programmes malveillants qui contrôle des faux utilisateurs, ou leurs activités sociales sont similaires à celles de vrais utilisateurs, les méthodes de prédiction de liens doivent prendre en compte l'influence négative des faux utilisateurs.

II.7 Conclusion

Dans cette section, nous avons présenté le problème de prédictions d'interaction dans les réseaux et comment aborder ce problème par une formulation mathématique. Nous avons également présenté les approches et les méthodes principales de prédiction de liens les plus utilisées. Enfin, nous avons cité les domaines d'application, les orientations et les défis futurs.

CHAPITRE III

Tests et expérimentations

III.1 Introduction

Dans ce dernier chapitre, nous présentons l'implémentation de deux approches de prédiction d'interactions dans les réseaux complexe. La première est basée sur les méthodes de similarité et la deuxième est basée sur l'apprentissage supervisé. Nos expérimentations sur les trois dataset choisies ainsi que les tests et les résultats obtenus serons exposés à la fin du chapitre.

III.2 Technologies utilisées

Nous avons utilisé le langage de programmation « Python » pour faire les différentes implémentations. Ce dernier nous a permis d'effectuer facilement nos expérimentations.

III.2.1 Python

Python est un langage de programmation orienté objet, interprété et de haut niveau. Il possède une sémantique dynamique et robuste pour toutes les plateformes (Unix, MacOs, Windows). Python est un langage puissant qui permet des représentations simples et flexibles pour les graphes ainsi qu'il possède des expressions claires et concises pour le Graph Mining.

III.2.2 Bibliothèques utilisées

NetworkX

NetworkX est un package Python pour la création, la manipulation et l'étude de la structure, de la dynamique et des fonctions pour les réseaux complexes.

Matplotlib

Matplotlib est une bibliothèque pour créer et tracer des visualisations statiques, interactives et animées en Python.

NumPy

NumPy est une bibliothèque pour créer et manipuler de grands tableaux et matrices multidimensionnels, ainsi qu'une grande collection de fonctions mathématiques de haut niveau pour manipuler les matrices et les tableaux.

Pandas

Pandas est une bibliothèque permettant l'analyse et la manipulation des données. Elle offre des structures de données et des opérations de séries temporelles et des fonctions pour la manipulation de tableaux numérique.

Sklearn

Sklearn est une bibliothèque destinée au Machine Learning, la régression, l'estimation et l'apprentissage automatique. Elle contient des fonctions et outils simples et efficaces pour l'analyse prédictive des données.

III.2.3 Jupyter Notebook

Jupyter Notebook est une application Web open source permette de créer et de partager des documents contenant du code, des équations, des visualisations et des textes explicatifs.

Il nous permet de faire le nettoyage et la transformation des données, la simulation numérique, la modélisation statistique, l'apprentissage automatique et bien plus encore.

III.3 Datasets

Nous travaillons sur 3 datasets de prédiction de liens de différents réseaux complexes, leurs sources sont disponibles sur : https://noesis.ikor.org/datasets/link-prediction

- HMT: Réseau social, amitiés entre utilisateurs du site hamsterster.com (KONECT [2014])

- YST: Réseau biologique, réseau d'interactions protéine-protéine de levure bourgeonnante (Bu et al. [2003])

- SMG: réseau de co-auteurs (l'un des dataset de Pajek)

Tableau 2 : Liste des bases de données avec leurs caractéristiques

	Nombre du Nœuds	**Nombre des Liens**
YTS	2 284	6 646
SMG	1 024	4 916
HMT	2 426	16 630

III.4 Les processus de prédiction des liens

III.4.1 Processus de prédictions des liens basé sur la similarité

Tout d'abord on divise la dataset en deux sous-graphes. Le premier sous-graphe est utilisé pour l'apprentissage et le deuxième pour le test. Nous éliminons ensuite, tous les nœuds isolés et les nœuds incommuns des deux graphes.

Ensuite, nous construisons le vecteur des données réel, qui contient la liste des liens à prédire. Après appliquer les métriques de similarités de prédictions sur le graphe d'apprentissage nettoyé, nous obtenons les scores de prédictions.

Enfin, nous comparons le vecteur des scores et des données réels pour obtenir les résultats qui nous permet de faire une évaluation des performances de ce processus.

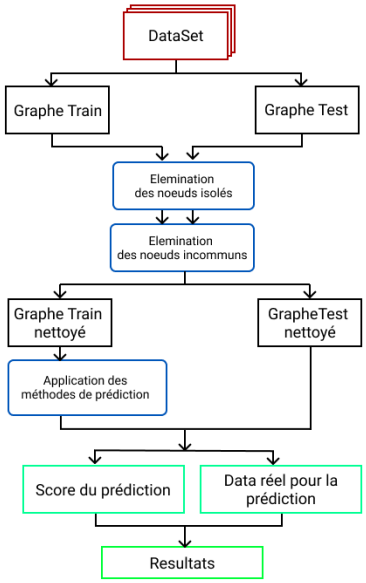

Figure 7 : Organigramme des étapes d'implémentation de l'approche basée sur la similarité

III.4.2 Processus de prédictions des liens basé sur la classification supervisé

Initialement, on divise la base de données en deux graphes, graphe d'apprentissage et graphe de test, puis nous faisons le prétraitement des données qui consiste à faire une élimination de tous les nœuds et une isolation des nœuds incommuns des deux graphes.

Ensuite, les deux graphes nettoyés sont partitionnés au hasard en cinq sous-graphes de la même taille. Parmi ses cinq sous-graphes, un seul sous-graphes est conservé comme données de validation pour tester le modèle. Les quatre sous-graphes restant sont utilisé comme données d'apprentissage. Ce processus de validation est ensuite répété cinq fois et chacun des cinq sous-graphes est utilisé exactement une fois comme données de validation. Nous refaits ensuite, les prétraitement une dernière fois.

Après, nous entrainons un modèle de classification sur les sous-graphes d'apprentissage pour générer des prédictions des liens, pour nous les comparons avec les graphes de tests.

Enfin, nous obtenons les résultats qui nous permettent de faire une évaluation des performances de ce processus.

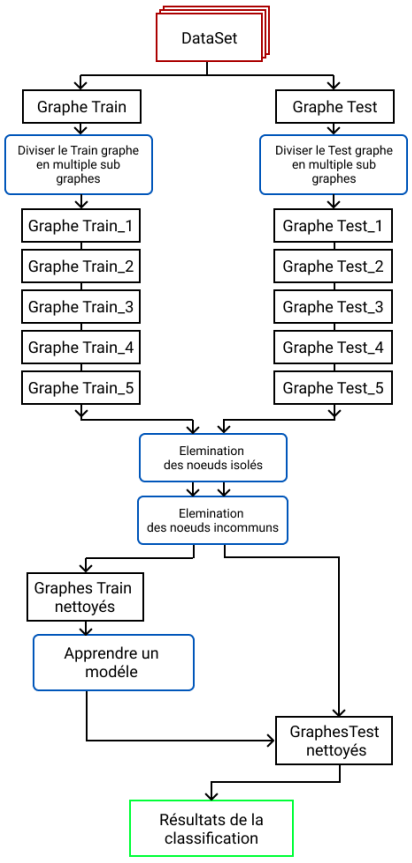

Figure 8 : Organigramme des étapes d'implémentation de l'approche basée sur la classification supervisée

III.5 Les mesures d'évaluation

Après l'implémentation, nous calculons les résultats possibles, elles peuvent être formulées par une matrice de confusion ou un tableau de contingence qui compare la classe prédite à la classe réelle. Nous avons 4 situations différentes :

- **True Positive**: Nombre de liens qui existe dans la dataset d'apprentissage et dans la dataset du test.
- **True Negative**: Nombre de liens qui n'existe pas dans la dataset d'apprentissage et existe dans la dataset du test.
- **False Positive**: Nombre de liens existe dans la dataset d'apprentissage et n'existe pas dans la dataset du test.
- **False Negative**: Nombre de liens qui n'existe pas dans la dataset d'apprentissage et n'existe pas dans la dataset du test.

Tableau 3 : Matrice de confusion

		Classe Réelle	
		Lien	**Pas de Lien**
Classe Prédite	**Lien**	**TP**	**FP**
	Pas de Lien	**FN**	**TN**

Précision : c'est la fraction de vrais liens positifs parmi l'ensemble des liens prédits comme positifs. Où, c'est la fraction de faux liens positifs parmi l'ensemble des liens prédits comme négatifs.

$$Précision = \frac{TP}{TP + FP}, ou, Précision = \frac{TN}{TN + FN}$$

Rappel (Recall) : également connu sous le nom de *sensibilité* ou taux de True Positive, qui est le rapport entre les vrais liens positifs entre les liens prédits et le nombre réel de liens positifs. Le rappel peut être considéré comme la probabilité qu'un lien réellement positif soit prédit.

$$Rappel = \frac{TP}{TP + FN} = \frac{TP}{P}$$

Le rappel est aussi connu sous le nom de taux True Negatives qui est le rapport entre les vrais liens négatifs entre les liens prédits et le nombre réel de liens négatifs.

$$Rappel = \frac{TN}{TN + FP} = \frac{TN}{N}$$

F-mesure : Le score F, également appelé F mesure, combine la précision et le rappel dans un seul score analytique.

$$F\_score = 2 \times \frac{Précision \times Rappel}{Précision + Rappel}$$

Courbe ROC (Receiver Operating Characteristic): est la variation du taux de vrais positifs (TPR: True Positives Rate) par rapport au taux de faux positifs (FPR: False Positive Rate) à divers réglages de seuil, respectivement au axes **x** et **y**.

TPR : fraction de vrais positifs sur le total des positifs, c'est-à-dire

$$T\,P\; =\; P\; =\; T\,P\; =\; (T\,P\; +\; F\,N)$$

FPR : fraction de faux positifs sur le total des négatifs, c'est-à-dire

$$F\,P\; =\; N\; =\; F\,P\; =\; (F\,P\; +\; T\,N)$$

Courbe de Rappel-Précision : est la variation de la précision par rapport au rappel à divers réglages de seuil. Ainsi, il est conçu par Rappel et Précision en tant qu'axes **x** et **y** respectivement.

AUROC (Area Under the ROC) : La courbe AUC - ROC est une mesure de performance pour un problème de classification à divers réglages de seuils. ROC est une courbe de probabilité et AUC représente le degré ou la mesure de la séparabilité. Il indique dans quelle mesure le modèle est capable de distinguer les classes. Plus l'AUC est élevée, mieux le modèle est à prédire les 0 comme 0 et 1 comme 1.

Accuracy (Exactitude) : calcule les performances globales du modèle de classification indépendamment les classes, elle est donné par le ratio entre le nombre total des liens correctement prédit sur le nombre total des liens

$$Accuracy = \frac{(TP\; +\; TN)}{(TP\; +\; TN\; +\; FP\; +\; FN)}$$

III.6 Résultats et expérimentation

III.6.1 Approche basée sur la similarité

Après avoir appliqué les méthodes de similarité et les mesures de performance sur les datasets on a obtenu les résultats suivants :

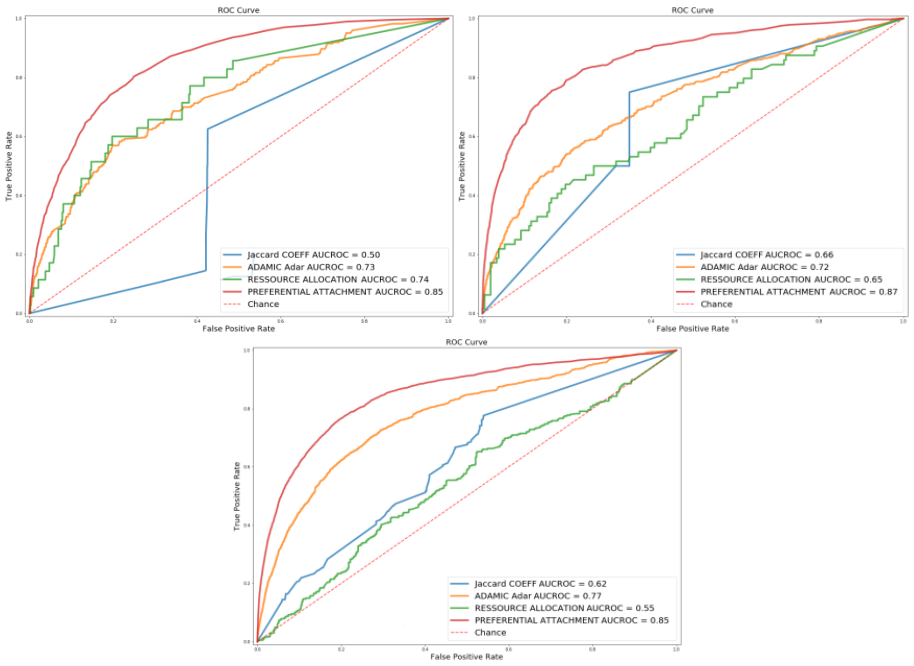

Figure 9 : Courbe ROC des bases de données YTS, SMG et HMT

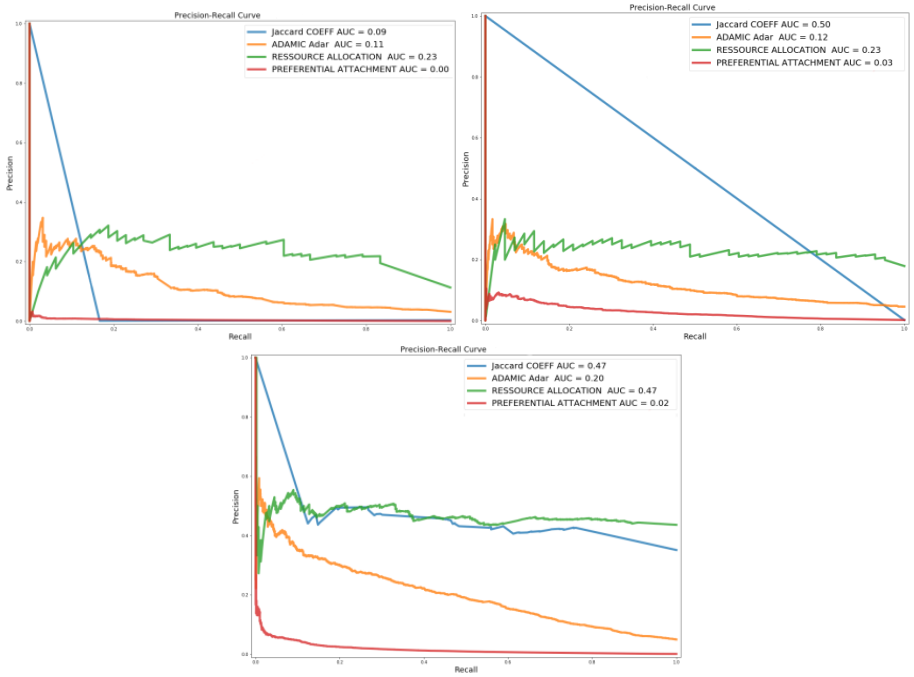

Figure 10 : Courbe AUCROC des bases de données YTS, SMG et HMT

III.6.2 Approche basée sur la classification

III.6.2.1 Classification KNN

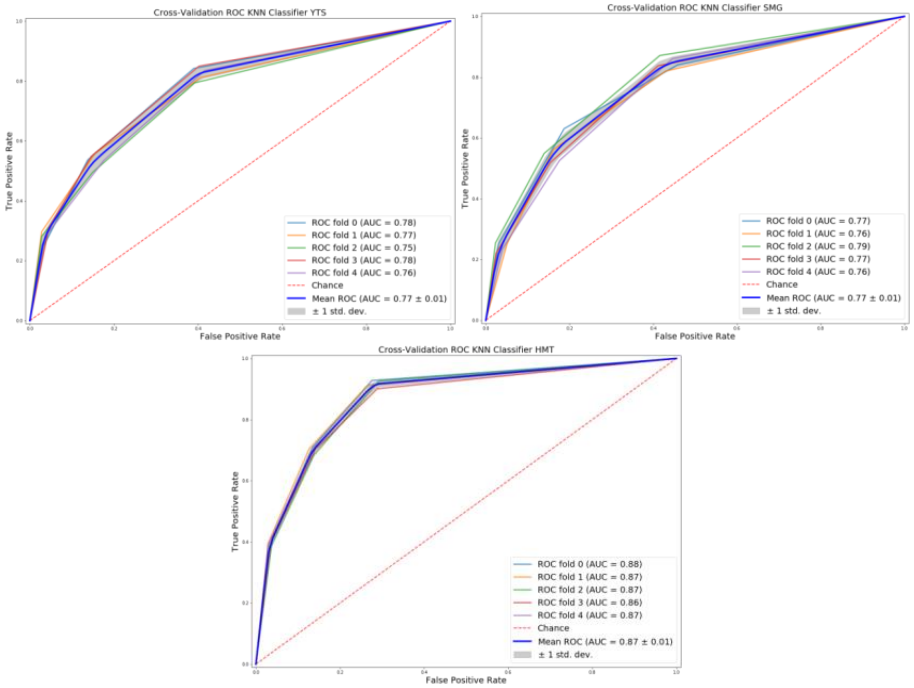

Figure 11 : Courbe ROC des bases de données YTS, SMG et HMT de classification KNN

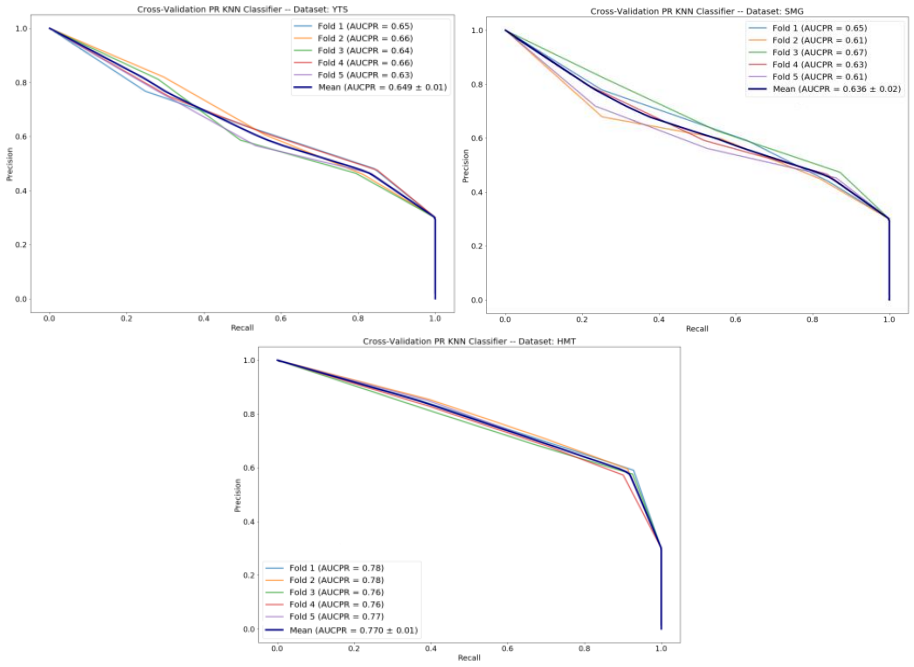

Figure 12 : Courbe AUCROC des bases de données YTS, SMG et HMT de classification KNN

III.6.2.2 Classification arbre de décision

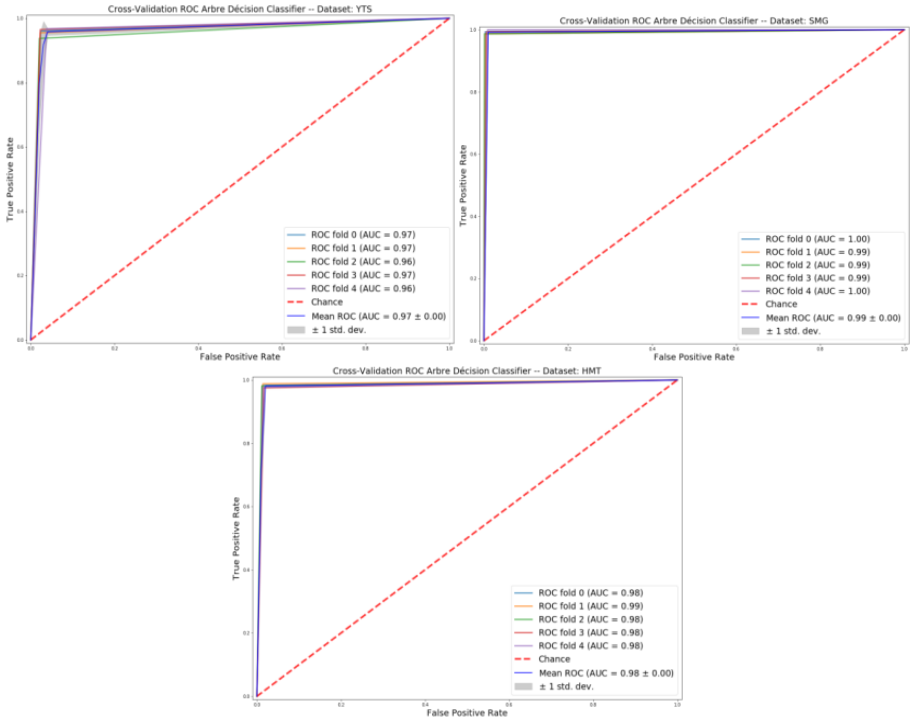

Figure 13 : Courbe ROC des bases de données YTS, SMG et HMT de classification AD

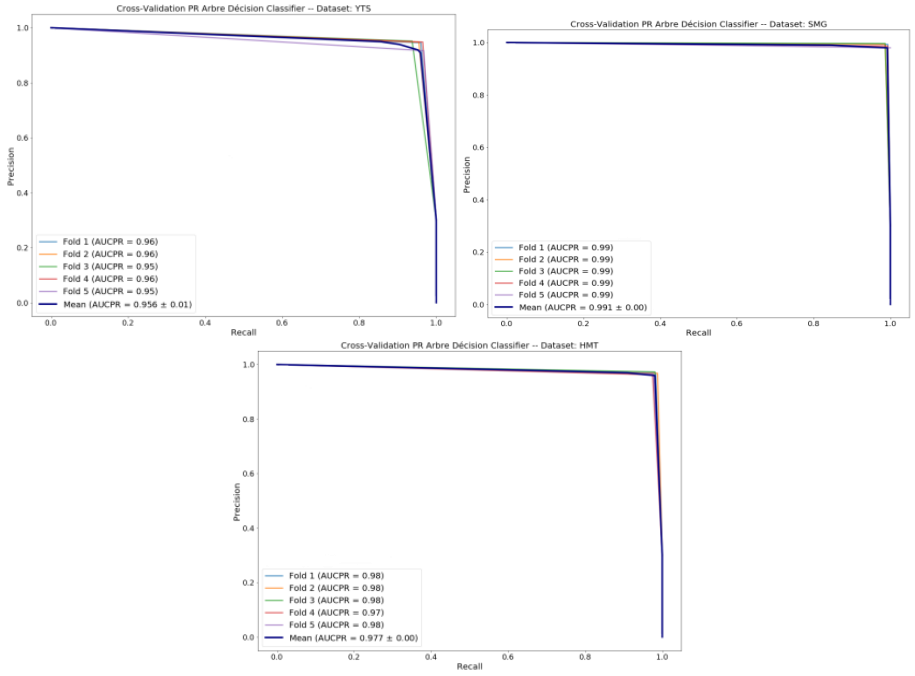

Figure 14 : Courbe AUCROC des bases de données YTS, SMG et HMT de classification AD

III.7 Discussion des résultats :

Pour évaluer chaque technique (approche), nous avons tracé la courbe ROC et la courbe AUCROC pour chaque méthode implantée. On note que, plus la courbe suit la bordure gauche, puis la bordure supérieure de l'espace ROC, plus le test est précis, et plus la courbe se rapproche de la diagonale de 45 degrés, moins le test est précis.

Tout d'abord, nous remarquons dans la courbe ROC que la métrique Préférentiel Attachement a obtenue de bons résultats dans les trois datasets. Ensuite, les métriques Adamic Adar et Allocation des Ressource ont eu des résultats moins bons que la métrique PA mais, elles sont considérées aussi bonnes. En fin, la métrique Coefficient de Jaccard a eu des résultats moyens.

Ensuite, nous observons dans la courbe AUCROC, la métrique PA a obtenue de bons résultats au début puis elle coule vite et mal dans toutes les bases de données. Puis, les métriques AA et AR ont eu des résultats un peu moins que la moyenne. En fin, la métrique CJ a obtenue de résultats mauvais pour la première dataset et moyenne pour la deuxième et la troisième.

Nous déduirons que pour les bases de données utlisées, le Coefficient de Jaccard est la moins bonne métrique de similarités puis l'allocation des ressources, ensuite Adamic Adar et en fin Préférentiel Attachement.

Pour estimer chaque algorithme de classification, nous avons effectué une validation croisée en cinq volets (folds). Tout d'abord, l'algorithme KNN a réalisé de bons résultats sur les deux courbes ROC et AUCROC, les résultats de base de données HMT est mieux qu'YTS et SMG. Ensuite, nous observons que les résultats d'algorithme arbre de décision sont parfaits dans toutes les datasets, dans les deux courbes.

Nous conclurons que l'arbre de décision est le meilleur algorithme de classification pour faire les prédictions des interactions. La variété des métriques montre que les performances de chaque technique dépendent fortement sur les propriétés structurelles du réseau. Cela met en évidence l'importance d'analyser les propriétés du réseau avant de choisir une métrique de prédiction de lien particulière. Comme nous nous observons dans nos résultats, la qualité des résultats est liée au coefficient de clustering moyen des nœuds de degré supérieur à un. Ceci est raisonnable car la plupart des techniques de prédiction de lien sont des variations de comptage des voisins partagés, et le nombre de voisins partagés augmente avec le coefficient de regroupement. Cela a du sens car, comme nous savons que plus il y a de voisins d'un nœud, plus nous avons d'informations pour prédire de nouveaux liens.

III.8 Conclusion

En arrivant à la fin de notre mémoire, nous avons appliqué quatre métriques (Préférentiel Attachement, Adamic Adar Index, Coefficient de Jaccard, Allocation des Ressources) de prédiction des interactions sur trois bases de données des réseaux complexes (YTS, SMG et HMT). Ensuite, nous avons appliqué sur ces mêmes bases de données deux algorithmes d'apprentissage de classification (Arbres de décision et KNN). Enfin, nous avons effectués une comparaison entre les méthodes et les différents algorithmes utilisés.

Conclusion générale

A l'occasion de notre projet de fin d'études, nous avons choisi un domaine de recherche très intéressent et récent qui est l'analyse des réseaux complexe.

Ce domaine constitue une réponse de mise en œuvre des techniques intelligentes de recherches et de connaissances à partir de vastes ensembles de données. Ces connaissances étés utilisées pour des applications se rattachant à plusieurs domaines. Parmi celle-ci, la prédiction des liens dans les réseaux complexe, qui est un domaine de recherche rendu très actif par multiplication du nombre de documents numérique actuellement disponibles. Vu la croissance du flux et de la masse d'informations disponibles, il est nécessaire de livrer aux personnes la compréhension des interactions entre eux, et faciliter la visualisation et la navigation dans les réseaux énormes.

La première partie de notre travail a été destiné à l'étude théorique. Nous avons présenté les réseaux complexes d'une façon général et aussi caractériser le problème de prédiction des liens et les différents approche et algorithmes existants.

La deuxième partie a été consacrée aux expérimentations. Nous avons testé quatre méthodes de prédiction des liens (Common Neighbors, Coefficient de Jaccard, Adumic/Adar et et Preferential Attachment) sur trois réseaux différents (YTS, SMG et HMT). Aussi, on a testé deux méthodes d'apprentissage supervisé (Arbre de décision et KNN (K Nearest Neighbors)) sur les réseaux mentionnés.

Dans ce projet, nous avons présenté un cadre pour caractériser le problème de prédiction des liens dans les réseaux complexe. On s'intéresse principalement aux méthodes de similarité et aux méthodes d'apprentissage supervisé pour prédire les liens.

Les buts de notre projet de fin d'étude sont d'obtenir un ensemble de connaissances dans le domaine de l'analyse des réseaux complexe, de recevoir une bonne initiation à la recherche scientifique et d'assembler plusieurs compétences tel que Python et théorie des graphes.

A la fin de notre travail nous voulons bien dire que nous avons atteint les objectifs et les buts tracés derrière cette étude.

Bibliographie

[1] https://en.wikipedia.org/wiki/Complex_network
[2] R. Albert and A.-L. Barabási (2002). "Statistical mechanics of complex networks". Reviews of Modern Physics. Article, published 30 January 2002.
[3] Mark Newman (2010). "Networks: An Introduction". Livre, 25 Mars 2010.
[4] Reuven Cohen and Shlomo Havlin (2010). "Complex Networks: Structure, Robustness and Function". Book published by Cambridge University Press, August 2010
[5] https://en.wikipedia.org/wiki/Social_network
[6] Kietzmann, Jan H.; Kristopher Hermkens (2011). "Social media? Get serious! Understanding the functional building blocks of social media', article, May 2011, project Innovation Dynamics.
[7] Obar, Jonathan A.; Wildman, Steve (2015). "Social media definition and the governance challenge: An introduction to the special issue". Paper, 26 August 2015
[8] Aichner, T.; Jacob, F. (2015). "Measuring the Degree of Corporate Social Media Use". January 2015 International Journal of Market Research Vol. 57 Issue 2
[9] Aggarwal C.C. (2011) An Introduction to Social Network Data Analytics. In: Aggarwal C. (eds) Social Network Data Analytics. Springer, Boston, MA. https://doi.org/10.1007/978-1-4419-8462-3_1
[10] https://www.sci.unich.it/~francesc/teaching/network/biological.html, link of teaching of Massimo Franceschet Professor of Computer Science, University of Udine
[11] https://www.researchgate.net/figure/An-example-metabolic-subnetwork-A-A-small-metabolic-network-with-17-reactions_fig3_225072275. Article Elementary Conversion Modes Unveil All Capabilities of Metabolic Networks January 2020SSRN Electronic Journal DOI: 10.2139/ssrn.3671940, Authors: Tom J. Clement, Erik B. BaalHuis, Bas Teusink, Frank J.Bruggeman, Rober Planqué, Daan de Groot
[12] Peng Wang, BaoWen Xu, YuRong Wu, and XiaoYu Zhou. 2015. Link prediction in social networks: The state-of-the-art. Science China Information Sciences 58, 1 (2015), 1–38.
[13] David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology 58, 7 (2007), 1019–1031.
[14] Liu Z, Zhang Q, L¨u L, et al. Link prediction in complex networks: A local na¨ve Bayes model. Europhysics Letters (EPL), 2011, 96: 48007
[15] Bruce McCune, James B. Grace, and Dean L. Urban. 2002. Analysis of Ecological Communities. Vol. 28. MjM Software Design, Gleneden Beach, Oregon.
[16] Allali O, Magnien C, Latapy M. Link prediction in bipartite graphs using internal links and weighted projection. In: INFOCOM Workshop on Network Science for Computer Communications, Shanghai, China, 2011. 936–941
[17] Erzsebet Ravasz, Anna Lisa Somera, Dale A. Mongru, Zoltan N. Oltvai, and A.-L. Barabasi. 2002. Hierarchical organization of modularity in metabolic networks. Science 297, 5586 (2002), 1551–1555.
[18] Paul Jaccard. 1901. Etude comparative de la distribution florale dans une portion des alpes et des jura. Bulletin de la Societe Vaudoise des Sciences Naturelles 37 (1901), 547579.
[19] Lada A. Adamic and Eytan Adar. 2003. Friends and neighbors on the web. Social Networks 25, 3 (2003), 211–230.

[20] Albert-Laszl ´ o Barab ´ asi and R ´ eka Albert. 1999. Emergence of scaling in random networks. Science 286, 5439 (1999), 509–512.
[21] Linyuan Lu, Ci-Hang Jin, and Tao Zhou. 2009. Similarity index based on local paths for link prediction of complex networks. Physical Review E 80, 4 (2009), 046122.
[22] Liu Z, Zhang Q, Lu L, et al. Link prediction in complex networks: A local na¨ve Bayes model. Europhysics Letters (EPL), 2011, 96: 48007
[23] Thorvald Sørensen. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biologiske Skrifter 5 (1948), 1–34.
[24] Leo Katz. 1953. A new status index derived from sociometric analysis. Psychometrika 18, 1 (1953), 39–43. Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. Computer 8 (2009), 30–37.
[25] David Easley and Jon Kleinberg Networks, Crowds, and Markets: Reasoning About a Highly Connected World, Published July 31st 2010 by Cambridge University Press (first published June 30th 2010) ISBN0521195330 (ISBN13: 9780521195331)
[26] Li X, Du N, Li H, et al. A deep learning approach to link prediction in dynamic networks. In: Proceedings of the 2014 SIAM International Conference on Data Mining, Philadelphia, Pennsylvania, USA, 2014. 289–297
[27] Sarkar P, Chakrabarti D, Jordan M. Nonparametric link prediction in dynamic networks. In: Proceedings of the 29th International Conference on Machine Learning (ICML'12), Edinburgh, Scotland, 2012.
[28] Pavlov M, Ichise R. Finding experts by link prediction in co-authorship networks. In: Proceedings of the 2nd International ISWC+ASWC Workshop on Finding Experts on the Web with Semantics (FEWS), Busan, Korea, 2007. 42–55
[29] Tang J, Wu S, Sun J M, et al. Cross-domain collaboration recommendation. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'12), Beijing, China, 2012. 1285–1293
[30] Sun Y, Barber R, Gupta M, et al. Co-author relationship prediction in heterogeneous bibliographic networks. In: Proceeding of the 2011 IEEE/ACM International Conference on Advanced in Social Networks Analysis and Mining (ASONAM'11), Kaohsiung, Taiwan, 2011. 121–128
[31] Zhang H, Dantu R. Predicting social ties in mobile phone networks. In: Proceedings of 2010 IEEE International Conference on Intelligence and Security Informatics, Vancouver, Canada, 2010. 25–30
[32] Gilbert E, Karahalios K. Predicting tie strength with social media. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Boston, USA, 2009. 211–220
[33] Xiang R, Neville J, Rogati M. Modeling relationship strength in online social networks. In: Proceedings of the 19th International Conference on World Wide Web (WWW'10), Raleigh, USA, 2010. 981–990
[34] Qiu B, He Q, Yen J. Evolution of node behavior in link prediction. In: Proceedings of the 25th AAAI Conference on Artificial Intelligence, San Francisco, California, USA, 2011. 1810–1811