Prévision de liens dans les graphes bipartites avec attributs

Vanessa Kamga[**,*], Maurice Tchuente[**], Emmanuel Viennet[*]

[*]Université de Paris 13, Sorbonne Paris Cité, L2TI
F-93430, Villetaneuse, France
[**]Université de Yaoundé 1, UMMISCO-LIRIMA, Equipe IDASCO
BP 812 Yaoundé, Cameroun
http://www.lirima.uninet.cm

Résumé. Les réseaux sociaux se modélisent fréquemment par des graphes exprimant les relations explicites ou implicites entre les entités considérées. Ces graphes sont très dynamiques: de nouveaux liens, de nouvelles entités apparaissent et disparaissent rapidement. Ce travail porte sur la prévision de liens dans les graphes bipartites dynamiques, en particulier dans le cas où des données (attributs) sont associées aux entités. Ce cas est important en pratique, notamment pour les systèmes de recommandation: la prévision de liens dans un réseau Clients/Produits revient en effet à prévoir les produits qu'un client est susceptible d'acquérir dans un avenir proche.

Le problème de prévision de liens peut s'aborder en considérant les propriétés structurelles du graphes (approches topologiques) ou via les systèmes de recommandation (eg filtrage collaboratif). Nous proposons des méthodes qui prennent en compte simultanément les attributs des nœuds et des liens. Nous illustrons ces méthodes sur le cas bien connu des graphes de collaborations scientifiques, où nos modèles utilisent à la fois les relations de co-publications et les résumés des articles.

Pour les graphes bipartites, une proposons approche basée sur les règles d'association liées au voisinage des nœuds. L'évaluation sur 4 sections d'arXiv montre que ces méthodes permettent d'obtenir, par rapport aux approches topologiques et le filtrage collaboratif, une amélioration d'AUC située entre 6% et 16%.

1 Introduction

Les réseaux sociaux sont des systèmes composés d'entités ayant entre elles des relations. Par exemple, un réseau de collaborations scientifiques peut se modéliser par un graphe où les nœuds représentent les auteurs et un lien entre deux nœuds représente une collaboration entre ces auteurs. Un réseau de citations peut être, quant-à-lui, représenté par un graphe où les nœuds sont des articles et un arc d'un nœud **u** vers un nœud **v** indique que l'article **u** cite l'article **v**.

Le plus souvent, les réseaux sociaux sont des structures très dynamiques : avec le temps, de nouveaux nœuds apparaissent, d'autres disparaissent et des liens se font ou se défont. Comprendre les mécanismes d'évolution des réseaux sociaux est une question fondamentale qui

n'est pas encore bien résolue. La prévision des liens, qui fait l'objet de notre travail, peut se formuler ainsi qu'il suit :

Peut-on à partir de l'observation d'un réseau social, durant une période donnée [t, t'], déterminer les liens qui apparaîtront dans le réseau pendant une période ultérieure]t', t''] ?

Le problème de prévision de liens peut être utilisé pour concevoir des systèmes de recommandation d'ouvrages. En effet prévoir un lien entre un client C et un ouvrage O permet de recommander le livre O à C. Le problème de prévision de liens peut également être utilisé pour détecter des relations cachées. En effet, la prévision d'un lien entre deux acteurs d'un réseaux social pourrait suggérer l'existence d'une relation entre ces deux individus bien que la topologie du réseau ne la montre pas de manière explicite. Ceci peut avoir des applications par exemple, dans la lutte contre le terrorisme(Krebs, 2002; Macskassy et Provost, 2005).

Dans l'approche topologique, certaines méthodes utilisent la structure du graphe pour calculer des indices de similarité entre nœuds. Le système prévoit alors les liens entre les paires de nœuds les plus similaires (Liben-Nowell et Kleinberg, 2003). Plus récemment, une approche topologique par projection pondérée et liens internes a été proposée par (Allali et al., 2011). Elle consiste à proposer, dans la prévision de liens, les arêtes dont l'ajout ne modifie pas le graphe projeté.

Parmi les approches qui exploitent les attributs, certaines utilisent les méthodes d'apprentissage pour classer les liens. Par exemple (Hasan et al., 2006), pour les graphes de collaborations, se ramènent à un problème de classification des liens dans lequel les mots clefs communs aux publications de deux auteurs u et v sont un attribut du lien (u, v). Par ailleurs, le filtrage collaboratif (Rajaraman et Ullman, 2011), pour un graphe Clients/Produits, recommande à un client U, les produits que les clients qui lui sont similaires ont acquis ; il utilise alors les attributs tels que la résidence et la profession des clients.

Dans certains réseaux, les entités mises en relation ne sont pas du même type. C'est notamment le cas pour les réseaux représentant des activités : les entités sont soit les personnes qui mènent l'activité, soit les objets sur lesquels porte l'activité. C'est notamment le cas pour les acteurs de cinéma et les films, les internautes et les pages web, les chercheurs et les publications, les clients et les produits. On parle dans ce cas de graphe bipartite, biparti ou 2-partite. De manière plus générale, lorsqu'un réseau comporte k types de nœuds avec $k > 1$, on parle de graphe k-partite . Un exemple de réseau 3-partite est obtenu du Web où des internautes consultent des pages web qui contiennent des étiquettes/tags. L'expression *graphe multipartite* se réfère aux graphes k-partites avec $k > 1$. En revanche, pour $k = 1$, c'est-à-dire lorsque le réseau ne comporte qu'un seul type de nœud, on parle de graphe unipartite.

Cet article s'intéresse aux graphes unipartites et bipartites. Formellement, un graphe bipartite sera quelque fois noté $G =< \bot, \top, E >$ où $\bot$ est l'ensemble des nœuds inférieurs, $\top$ est l'ensemble des nœuds supérieurs et $E \subseteq \bot \times \top$ désigne l'ensemble des liens. La figure 1 présente un graphe bipartite comportant 6 nœuds $\top$ et 4 nœuds $\bot$.

Définition 1. ***Projection non pondérée***
Étant donné un graphe bipartite $G = (\bot, \top, E)$, la projection non pondérée permet de construire les graphes $G_\bot$ et $G_\top$ pour chaque type de nœuds comme suit : un lien existe entre deux nœuds $x, y \in \bot$ dans $G_\bot$ s'ils sont reliés à au moins un même nœud $z \in \top$, c'est-à-dire si $\Gamma(x) \cap \Gamma(y) \neq \emptyset$. $G_\top$ est construit de manière analogue.

Avec cette définition, le graphe projeté $G_\top$ n'apporte pas d'information sur les nœuds communs auxquels deux nœuds x et y de $G_\bot$ sont connectés. Par exemple, dans la partie

supérieure de la figure 1, les nœuds 3 et 4 qui ont pour unique voisin b sont relié de la même manière que les nœuds 4 et 5 qui ont pour voisins b et c.

Pour résoudre ce problème, on peut définir une fonction de pondération permettant de représenter la force de la relation entre deux nœuds dans le graphe projeté. Des exemples de fonctions de pondération disponibles sont :

- le common neighbors qui définit la force entre deux nœuds par le nombre de voisins qu'ils ont en commun (Newman, 2001).
- le cœfficient de Jaccard qui est la proportion des voisins communs par rapport au nombre total de voisins (Jaccard, 1901).
- la mesure d'Adamic et Adar qui compte les voisins communs tout en pénalisant ceux ayant un grand degré, car ils sont considérés comme banals (Adamic et Adar, 2003).

La figure 1 présente un graphe bipartite sur lequel ont été effectuées deux projections, l'une pondérée, l'autre non.

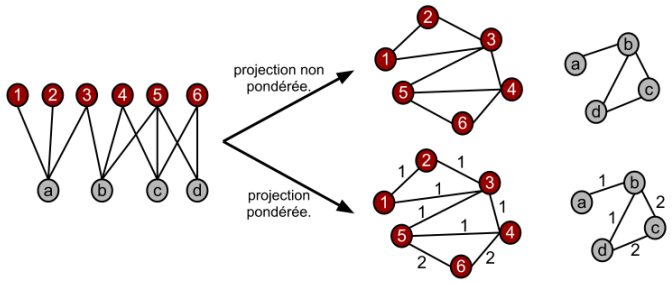

FIG. 1: Projections non pondérée et pondérée. La fonction de pondération utilisée est common neighbors.

Nous proposons dans cet article de nouvelles approches de prévision de liens basées sur les propriétés structurelles ainsi que sur les attributs des nœuds et ceux des liens : voisinage, résumé d'articles dans les graphes de collaborations scientifiques.

La suite de l'article est organisée ainsi qu'il suit : la section 2 est consacrée au rappel des méthodes classiques développées dans la littérature pour résoudre le problème de prévision de liens dans les réseaux sociaux. Cette analyse révèle les limites des modèles connus, ce qui permet de justifier les approches présentées dans la section 3. Dans la section 3, nous proposons de nouvelles approches qui tiennent compte de la structure du réseau et des règles d'association associées au voisinage. Dans le cas particulier des réseaux de collaborations, nous introduisons une méthode qui tient compte des résumés d'articles. La section 4 présente la mise en œuvre de ces approches sur les données obtenues à partir d'Arxiv. La section 5 est consacrée à la conclusion et aux perspectives.

2 Approches topologiques, supervisées et pondérées

Dans cette section, nous définissons d'abord le problème de prévision de liens puis nous présentons les principaux algorithmes connus pour la prévision de liens dans les réseaux sociaux.

Définition 2. ***Prévision de liens***
Étant donnée l'observation $G_{[0,t']}$ d'un réseau social pendant une période $[0, t']$, une des définitions formelles du problème de la prévision de liens est de prédire les liens susceptibles d'apparaître dans ce réseau dans un avenir proche correspondant à un intervalle $]t', t"]$. A cet effet, on commence par subdiviser l'intervalle d'observation $[0, t']$ en une période $[0, t], t < t'$ et une période $]t, t']$. Soit $G_{]t,t']}$ le sous-graphe correspondant à l'ensemble des arêtes qui apparaissent dans l'intervalle $]t, t']$. Comme c'est l'usage dans les méthodes d'apprentissage, on construit un ensemble d'apprentissage constitué ici par $G_{[0,t]}$ et une partie A de $G_{]t,t']}$, puis un ensemble de test constitué par $G_{]t,t']} \setminus A$. La méthode de prévision de liens, une fois conçue, doit produire pour l'intervalle de test $]t, t']$, un ensemble de lien E^* aussi proche que possible de l'ensemble $E^{'}$ des liens de $G_{]t,t']} \setminus A$.

2.1 Cas des graphes unipartites

Une première approche, proposée en 2003 par David Liben et Jon Kleinberg (Liben-Nowell et Kleinberg, 2003), consiste à définir des indicateurs topologiques de similarité entre deux nœuds et de prévoir les liens entre les paires de nœuds les plus similaires. Cette approche est considérée comme topologique car ne tient compte que de la structure du réseau et ignore les informations que peuvent apporter les nœuds et les liens. La similarité entre deux nœuds x et y est notée dans la suite Score(x,y).

La mesure de similarité la plus simple entre deux nœuds x et y, encore appelée *common neighbors* (Newman, 2001), est définie par :

$$Score(x, y) = |\Gamma(x) \cap \Gamma(y)|$$

Pour cette mesure, si x et y ont un grand nombre de voisins en commun, alors la similarité entre ces nœuds est grande, ce qui correspond à l'effet souhaité. En revanche, si x et y ont très peu de voisins en commun, mais avec $\Gamma(x) = \Gamma(y)$, alors cette fonction n'exprimera pas le fait que x et y sont très similaires. Un autre problème avec cette mesure est que la valeur de $Score(x, x)$ dépend de x, ce qui n'est pas normal car le degré de similarité entre un nœud et lui-même est maximal et indépendant de x.

Une solution à ces inconvénients est le cœfficient de Jaccard (Jaccard, 1901) :

$$Score(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$

Pour cette mesure, la plus grande valeur est 1 et correspond au cas où $\Gamma(x) = \Gamma(y)$, ce qui résout un problème posé par la précédente mesure. Par contre, si l'un des nœuds, x, a beaucoup de voisins, tandis que l'autre, y, en a très peu, mais tous voisins de x, alors ce cœfficient sera très petit et n'exprimera donc pas le fait que y est fortement similaire à x.

Une solution à cet inconvénient est la mesure d'Adamic et Adar (Adamic et Adar, 2003) :

$$Score(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log(|\Gamma(z)|)}$$

Dans cette mesure, un sommet voisin de x et y et ayant de nombreux autres voisins n'apporte pas d'information car il est banal. Par contre si un nœud z n'a que x et y comme voisins,

alors z apporte une information importante sur le comportement de x et y. Pour cette mesure, on attribue alors à chaque nœud z une force égale à 1 (d'où le numérateur de l'expression à sommer) ; cette force est ensuite équitablement répartie aux couples de voisins de z. Ainsi, la contribution au couple (x, y) d'un voisin z commun à x et y et ayant de nombreux voisins sera pénalisée par la taille du voisinage de z. Pour que la décroissance de cette contribution par rapport au nombre de voisins de z ne soit pas trop brutale, on utilise $\frac{1}{log(|\Gamma(z)|)}$ au lieu de $\frac{1}{|\Gamma(z)|}$.

Une autre mesure bien connue est l'attachement préférentiel (Mitzenmacher, 2001) :

$$Score(x, y) = |\Gamma(x)| \times |\Gamma(y)|$$

L'idée ici est que plus les nœuds x et y ont des voisins, plus grande sont leurs chances d'en avoir davantage et en particulier, la probabilité d'apparition d'un lien entre x et y est forte. Cette notion, également connue sous le nom d'avantage cumulatif, a été initialement conçue pour les graphes orientés par Dereck de Sola Price (Price, 1976) puis a été rebaptisée par Barabasi et al. (Mitzenmacher, 2001). Comme indiqué dans (Newman, 2003), son principe nous rappelle une parabole biblique : " Car on donnera à celui qui a, et il sera dans l'abondance, mais à celui qui n'a pas on ôtera même ce qu'il a. " (Mathieu 25 :29). Autrement dit, la probabilité qu'un nœud y s'attache à un sommet x est proportionnelle à la taille du voisinage de x, c'est-à-dire $|\Gamma(x)|$. Partant de ce principe, la probabilité qu'une arête ait pour extrémités x et y est donc, avec une hypothèse d'indépendance, la probabilité qu'une extrémité soit x et que l'autre extrémité soit y, d'où le score $Score(x, y) = |\Gamma(x)| \times |\Gamma(y)|$.

La mesure Katz (Katz, 1953) s'écrit :

$$Score(x, y) = \sum_{\ell=1}^{\infty} \beta^{\ell}.|path_{x,y}^{\ell}|$$

L'idée repose sur le fait que deux nœuds ayant de nombreux chemins les reliant sont similaires, et le sont encore plus si ces chemins sont de petites longueurs. $path_{x,y}^{\ell}$ représente l'ensemble des chemins de longueur ℓ entre x et y. Par ailleurs, les chemins courts doivent être avantagés et les chemins longs pénalisés. C'est pourquoi, est utilisé un facteur $\beta < 1$. Selon que les arêtes du graphe sont pondérées ou pas, on obtient une version non-pondérée ou pondérée. A noter que, avec cette définition, les poids des arêtes doivent être des entiers positifs.

Hasan et al. (Hasan et al., 2006) proposent une approche par apprentissage supervisé (classification) pour la prévision de liens, qui permet d'exploiter les attributs portés par les nœuds. Dans cette approche, le jeu de données est constitué de toutes les paires de nœuds non reliées dans $[0, t]$, pour $t < t'$, auxquelles on ajoute un attribut de classe 1 si le lien correspondant apparaît dans la période $]t, t']$ et 0 sinon. Ce jeu de données est ensuite subdivisé en deux ensembles : un pour le training et un pour le test. Pour cela, ils proposent de déterminer pour chaque lien (u, v) non encore présent dans le réseau, un ensemble d'attributs basés sur la topologie du réseau ainsi que les informations portées par les nœuds. Pour ne pas favoriser les attributs de grandes valeurs par rapport aux autres, chaque attribut est normalisé afin d'avoir une moyenne de 0 et une variance de 1. Finalement, on ajoute à chaque couple, un attribut classe qui prend la valeur 1 lorsque le lien correspondant apparaît durant la période de test, et 0 dans le cas contraire. On est ainsi ramené à un problème de classification.

2.2 Cas des graphes bipartites

Pour les graphes bipartites Clients/Produits, le filtrage collaboratif (Rajaraman et Ullman, 2011) consiste à déterminer pour chaque utilisateur U l'ensemble des produits utilisés par les utilisateurs qui lui sont similaires puis à les lui recommander. Cette méthode a connu de nombreux succès (Ricci et al., 2011).
(Allali et al., 2011) ont proposé une autre approche basée sur la notion de liens internes. Le principe des liens internes stipule que deux nœuds ayant au moins un voisin en commun pourront en acquérir davantage dans le futur tandis que deux nœuds qui n'en ont pas n'en auront jamais dans le futur. Cette dernière approche n'est malheureusement pas applicable aux graphes bipartites de publications car un lien interne ne peut exister qu'entre deux nœuds existants tandis que dans les graphes de publications, la création d'un lien se fait à travers une nouvelle publication (création d'un nœud de type "article").

3 Nouvelles approches pour la prévision de liens

Dans cette section, nous présentons trois nouvelles approches pour la prévision de liens dans les graphes bipartites. La première, que nous noterons Common*, s'applique pour la prévision de nouvelles collaborations entre auteurs dans les réseaux de publications et revient à introduire des pondérations dans l'approche *common neighbors* (Liben-Nowell et Kleinberg, 2003). On utilise un critère de similarité de type common neighbors, et on peut ainsi prévoir les 3 ou 5 liens qui maximisent la similarité pour chaque nœud. La deuxième s'applique aussi aux graphes de collaborations scientifiques et prend en compte des résumés d'articles. Enfin, la troisième est applicable à tous les graphes bipartites et est basée sur la construction de règles d'association appliquées aux voisinages des nœuds.

3.1 Common*

La mesure *common neighbors* appliquée aux graphes de collaborations scientifiques, utilise comme mesure de similarité entre deux auteurs x et y, la mesure $Score(x, y) = |\Gamma(x) \cap \Gamma(y)|$ qui est le nombre de co-auteurs communs. Dans cette approche, on ne tient pas compte du nombre d'articles publiés par deux co-auteurs. La mesure Common* considère cette information en considérant le graphe projeté pondéré et en prenant comme mesure de similarité :

$$Score^*(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} g(x, z) \times g(y, z)$$

où $g(e)$ représente le poids de l'arête e dans le graphe projeté. Cela revient à dire que si deux nœuds x et y ont chacun collaboré plusieurs fois avec un voisin commun z alors, x et y ont plus de chance de collaborer dans un futur proche. Pour retrouver la forme initiale de la mesure common neighbors, il suffit de prendre $g(x, z) = 1 \; \forall z \in \Gamma(x)$.

3.2 Utilisation des résumés des documents

Soit $G = (\perp, \top, E)$ un graphe bipartite de publications où $\perp$ représente l'ensemble des auteurs et $\top$ l'ensemble des articles. Les articles ayant fait l'objet de collaborations livrent des

informations sur les sujets d'intérêt commun aux auteurs. Il est naturel de supposer que les auteurs qui travaillent dans le même domaine ont plus tendance à collaborer et de co-publier un article sur ces mêmes thématiques.
Pour formaliser cette intuition, nous proposons de calculer la proximité entre deux auteurs à partir des contenus des articles qu'ils ont publiés. Ces articles peuvent être représentés par :

1. le titre et les mots clés : facile d'accès mais fournissant peu d'informations ;
2. le résumé : extraction possible à partir des bases bibliographiques, et assez informatif ;
3. l'article complet : beaucoup d'informations (mais souvent redondantes) et accès restreint pour certains articles par exemple lorsqu'ils sont payants.

Une façon raisonnable de représenter les articles serait alors d'utiliser les résumés. Toutefois, ces résumés nécessitent un pré-traitement. En effet, certains des mots constituant le résumé n'apportent pas d'information et constituent ce qu'on appelle *les mots vides* (Manning et Schütze, 1999). Les exemples de mots vides sont : les nombres, les mots de longueur 1 ou 2, les articles, les pronoms, les conjonctions, etc.
Par ailleurs, les mots d'un résumé peuvent être remplacés par leurs radicaux, afin d'ignorer les formes fléchies (conjugaisons, accords). Ce procédé appelé *racinisation* [1] (ou *stemming* en anglais) vise à rassembler les différentes variantes flexionnelles et dérivationnelles d'un mot autour d'un stemme ou radical. Ce procédé génère alors pour chaque résumé un ensemble de radicaux. Pour la mise en œuvre de cette méthode nous utilisons l'algorithme de Porter (Porter, 1980). Dans la suite, on suppose que l'on connaît l'ensemble S de tous les radicaux des mots utilisés dans le domaine de spécialité.
Comme l'objectif est de définir une nouvelle mesure de similarité entre auteurs et non entre résumés, il reste à montrer comment utiliser les résumés pour décrire les auteurs. La manière la plus simple de procéder consiste à « concaténer » tous les radicaux des résumés des articles qu'il a publiés : on obtient ce qui est appelé par la suite un descripteur de x. Le descripteur d'un auteur x est alors un vecteur $V(x) \in \mathbb{N}^n$ où $n = |S|$ et $V_i(x)$ est le nombre d'occurrences du radical numéro i dans l'ensemble des résumés des articles publiés par x.
Afin que les radicaux fréquents ne créent pas de fausses similarités, on normalise en divisant $V_i(x)$ par le nombre total d'occurrences du radical i dans la base. Ceci revient à calculer le *TF-IDF* (Manning et Schütze, 1999) qui est une mesure statistique permettant d'évaluer l'importance d'un terme contenu dans un document, relativement à une collection ou un corpus. Le poids d'un radical est proportionnel au nombre d'occurrences du radical dans les articles produits par un auteur (Tf). Il est inversement proportionnel à la fréquence du mot dans le corpus (Idf).
Après cette opération, chaque auteur x est représenté par un vecteur $v(x) \in [0,1]^n$. La proximité entre deux auteurs x et y est mesurée par le cosinus de l'angle formé par $v(x)$ et $v(y)$. En effet, pour deux descripteurs proportionnels $v(x)$ et $v(y)$, l'angle $(v(x), v(y))$ est nul et le cosinus est maximum. Pour deux auteurs n'ayant pas de mots communs dans leurs descripteurs $v(x)$ et $v(y)$, on a un produit scalaire $v(x).v(y) = 0$ et le cosinus est nul.

$$Score(x,y) = cos(v(x), v(y))$$

1. http ://fr.wikipedia.org/wiki/Racinisation, 24 Septembre 2012.

3.3 Prévision des liens basée sur les règles d'association

Dans cette section, nous présentons une approche pour la prévision de liens dans les graphes bipartites basée sur les règles d'association. L'idée sous-jacente est que si un groupe d'auteurs A publie généralement avec un autre groupe d'auteurs B et que B publie généralement avec un autre groupe C, alors les groupes A et C vont probablement publier ensemble dans un futur proche.

Pour mettre en œuvre cette approche, il faut rechercher les groupes d'auteurs fréquents puis extraire les règles d'association à partir de ces groupes. La méthode fonctionne comme suit :

1. Soit un graphe bipartite $G =< \bot, \top, E >$, où $\top$ représente les auteurs et $\bot$ les articles.
2. Construire une base de données transactionnelle où chaque enregistrement est l'ensemble des co-auteurs d'un article donné.
3. Rechercher dans la base de transactions, les groupes d'auteurs fréquents c'est-à-dire qui ont co-publié au moins s articles, s étant un seuil de fréquence fixé.
4. Extraire à partir de chaque transaction fréquente A, les règles d'association $P_1 \rightarrow Q_1$ définies par :
   (a) $P_1 \cup Q_1 = A$
   (b) $P_1 \cap Q_1 = \emptyset$
   (c) Une confiance c_1 qui indique que c_1 % des articles co-publiés par P_1 ont aussi Q_1 comme co-auteurs.
5. Utiliser les règles pour la prévision des liens comme indiqué dans l'algorithme 1.

Dans la suite, une règle d'association $X \rightarrow Y$ ayant pour confiance c est notée $X \rightarrow Y :: c$.

Définition 3. Règles compatibles
Une règle r_2 est dite compatible avec une règle r_1 si la condition de r_2 est incluse dans la conséquence de r_1.

Soient $r_1 : P_1 \rightarrow Q_1 :: c_1$ et $r_2 : P_2 \rightarrow Q_2 :: c_2$, deux règles compatibles. La probabilité d'observer Q_1 sachant P_1 est c_1 tandis que la probabilité d'observer Q_2 sachant P_2 est c_2. Sachant que la probabilité d'observer $P_2|Q_1$ est 1 puisque $P_2 \subseteq Q_1$, la probabilité d'observer Q_2 sachant P_1 est alors $c_1 \times c_2$.

Algorithme 1. Algorithme de prévision de liens à partir des règles d'association.

Entrée : Un ensemble de règles d'association R

Sortie : Un ensemble de liens prédits $E_{predict}$ avec le seuil de confiance *minconf.*

fonction PredireLien(R, *minconf*) :

début

$E_{predict} = \emptyset$

pour chaque règle $r = P_1 \rightarrow Q_1 : c_1$ dans R faire :

si $c_1 = 1$ alors :

$E_{predict} \leftarrow E_{predict} \cup \{(\min(u,v), \max(u,v)) |\, u \in P_1 \; et \; v \in Q_1\}$

Calculer l'ensemble H des règles qui complètent r

pour chaque règle $P_2 \rightarrow Q_2 : c_2$ dans H faire :

si $c_1 \times c_2 \geq$ *minconf* alors :

$E_{predict} \leftarrow E_{predict} \cup \{(\min(u,v), \max(u,v)) |\, u \in P_1 \; et \; v \in Q_2\}$

retourner $E_{predict}$

fin

4 Expérimentations

Nous étudions expérimentalement les modèles présentés dans la section 3 et les comparons aussi aux approches plus classiques mentionnées en introduction. Nous utilisons des graphes extraits de la base bibliographique arXiv.org pour une période d'apprentissage allant de 1994 à 1996 et une période de test allant de 1997 à 1999. Il s'agit des sections : Condensed Matter (Cond-mat), General Relativity et Quantum Cosmology (Gr-qc), High energy physics phenomenology (Hep-ph) et (Quant-ph). Le tableau 1 décrit les graphes qui ont été obtenus. $G =< \perp, \top, E_{pub} >$ désigne le graphe de publications. $G_\perp =< \perp, E >$ est le graphe simple de collaborations (il y a au plus un lien entre deux auteurs). $G^+_\perp =< \perp, E^+ >$ est le graphe pondéré des collaborations (le poids d'un lien (x, y) est le nombre de co-publications des auteurs x et y). *Core* est l'ensemble des auteurs ayant publié au moins k_1 articles pendant la période d'apprentissage et au moins k_2 articles pendant la période de test. Durant la période de test, on observe $|E'|$ liens apparaissant dans le graphe $G_\perp$ et $E^* = E' \cap (Core \times Core)$.

	$\lvert\perp\rvert$	$\lvert\top\rvert$	$\lvert E_{pub}\rvert$	$\lvert E^+\rvert$	$\lvert E\rvert$	k_1	k_2	$\lvert core\rvert$	$\lvert E^{'}\rvert$	$\lvert E^*\rvert$	$\lvert E^* \setminus E\rvert$
Quant-ph	1077	1062	582	1443	1044	1	1	416	4759	543	241
Gr-qc	3291	2641	1850	4417	3086	2	2	1153	6945	1461	409
Cond-mat	6697	6502	5322	17137	12114	3	3	1233	36692	2360	1356
Hep-ph	10277	6537	6721	33407	22489	4	4	1268	29129	3870	2163

TAB. 1: Récapitulatifs des données extraites.

La première méthode d'évaluation utilisée ici est basée sur la courbe ROC et l'AUC (Hanley et Mcneil, 1983). Une courbe ROC représente le taux de Vrais Positifs comme fonction du taux de Faux Positifs. L'AUC (*Area Under Curve*) est la surface comprise entre la courbe ROC et l'axe des abscisses. Dans cette approche, un modèle A est dit meilleur qu'un modèle B si l'AUC du modèle A est plus grande que celle du modèle B. En général, la courbe ROC et l'AUC permettent une bonne évaluation des performances d'un modèle (Hastie et al., 2009).

Pour la prévision de liens, il est souvent intéressant de minimiser le nombre de Faux Positifs. Par exemple, dans un système où chaque recommandation occupe un espace publicitaire, il est préférable d'avoir peu de Faux Positifs même si cela entraîne l'augmentation des Faux Négatifs. C'est pour cette raison que nous utilisons dans un deuxième temps, comme paramètre

d'évaluation, le *point de fonctionnement* γ qui se définit comme le taux de Faux Positifs que l'on tolère dans un système.

Dans un troisième temps, nous nous intéresserons au cas où l'on voudrait déterminer pour chaque auteur u, la liste des 3 ou 5 auteurs avec lesquels il est susceptible de publier dans le futur.

4.1 Approches implémentées

Nous avons implémenté dix méthodes :

1. Common neighbors (Newman, 2001) (*common*)
2. Jaccard (Jaccard, 1901) (*Jaccard*)
3. Adamic et Adar (Adamic et Adar, 2003) (*Adamic*)
4. Attachement préférentiel (Mitzenmacher, 2001) (*Attachement*)
5. Katz (Katz, 1953) (*Katz*)
6. Prévision aléatoire (*Random*) : ce modèle (en anglais *random predictor*) prévoit les liens de manière aléatoire. Plus précisément, la valeur de similarité entre deux nœuds u et v est obtenue en effectuant un tirage aléatoire dans $[0, 1]$. Cette mesure ne tient pas compte de la topologie du graphe ni des informations portées par les nœuds et/ou les liens. Ce modèle constitue évidemment une référence minimale, tout modèle pertinent devrait être au moins aussi bon que lui.
7. Filtrage collaboratif (*Filtering*) : regroupe l'ensemble des méthodes qui visent à construire des systèmes de recommandation utilisant les opinions et évaluations d'un groupe d'utilisateurs pour aider un autre utilisateur.
8. Méthode *Common** : utilise le nombre de collaborations entre deux auteurs pour calculer un indice de similarité.
9. Méthode *Abstract* : prévoit les liens entre auteurs travaillant sur les mêmes domaines en utilisant uniquement les résumés des articles.
10. Méthode *Rules* : utilise les règles d'association sur co-auteurs. Le minimum de support utilisé pour cette méthode est 1, 1, 2 et 3 respectivement pour les sections Quant-ph, Gr-qc, Cond-mat et Hep-ph.

4.2 Évaluation basée sur la courbe ROC et l'AUC

Les résultats obtenus sont présentés dans le tableau 2. Les nouvelles méthodes proposées sont en gras. Le facteur d'amélioration μ désigne la différence entre la valeur maximale des AUC des méthodes proposées (en gras) et le maximum des AUC des méthodes existantes.

De ces résultats, il ressort que les valeurs maximales d'AUC pour toutes les sections sont atteintes soit par la méthode *Rules* soit par la méthode *Abstract*.

4.3 Utilisation d'un point de fonctionnement

Nous rappelons que dans certains systèmes de prévision, il est important de minimiser le taux de Faux Positifs γ. A cet effet, on utilise un point de fonctionnement qui correspond au taux de Vrais positifs obtenu lorsque l'on tolère au plus γ de fausses alertes.

	Quant-ph	**Gr-qc**	**Cond-mat**	**Hep-ph**
Common	62%	58 %	58 %	66 %
Jaccard	62 %	58 %	58 %	66 %
Adamic	62 %	58 %	58 %	66 %
Attachement	*72 %*	*72 %*	*61 %*	*67 %*
Katz ($\beta = 0.05$)	63 %	62 %	59 %	59 %
Random	49 %	48 %	48 %	49 %
Filtering	62 %	57 %	57 %	62 %
Abstract	78 %	**79 %**	**77 %**	**82 %**
Common*	62 %	58 %	58 %	66 %
Rules	**82 %**	74 %	64 %	60 %
μ	***10 %***	***7 %***	***16 %***	***15 %***

TAB. 2: Les valeurs d'AUC par section et par méthode.

Dans cette section, nous considérons qu'une valeur de $\gamma = 5\ \%$ est raisonnable. L'idée ici consiste à déterminer l'ensemble T de paires de nœuds les plus similaires dans lequel on recense un taux $\gamma = 5\ \%$ de Faux positifs.

Le tableau 3 montre le taux de Vrais positifs obtenus sur les quatre sections de ArXiv. La méthode *Adamic* donne les meilleurs résultats sur les données générées à partir de la section **Hep-ph** suivie de près par la méthode *Abstract*. Pour la section **Cond-mat**, la méthode *Abstract*, basée sur les résumés d'articles, fournit les meilleurs résultats. Pour les sections **Quant-ph** et **Gr-qc**, la méthode *Rules* est de loin la meilleure suivie par la méthode *Abstract*. On en déduit que les nouvelles méthodes présentées dans cet article sont les plus adaptées lorsque l'on a une borne sur le taux de Faux Positifs.

	Quant-ph	**Gr-qc**	**Cond-mat**	**Hep-ph**
Common	12,60	12,54	12,54	12,78
Jaccard	15,72	12,54	12,54	18,73
Adamic	13,11	12,54	12,54	**25,51**
Attachement	17,68	18,34	13,65	15,42
Katz	12,66	12,54	12,54	14,26
Random	12,57	12,29	12,54	12,54
Filtering	12,64	12,62	12,54	13,36
Abstract	24,35	27,97	**22,04**	24,99
Common*	12,60	12,54	12,54	13,36
Rules	**54,75**	**39,89**	20,65	12,68

TAB. 3: Expérimentation du point fonctionnement pour $\gamma = 5\%$.

4.4 Recommandation de co-auteurs

Nous avons indiqué en introduction que le problème de prévision de liens dans les réseaux sociaux représentés par des graphes bipartites a de nombreuses applications telles que la recommandation de produits, la recommandation d'amis, la recommandation de co-auteurs et la détection des liens cachés. Les données collectées dans le cadre de ce travail nous permettent de mettre en œuvre l'une de ces applications : la recommandation de co-auteurs.

Un système de recommandation d'auteurs est un système où on détermine pour chaque auteur u un ensemble fini $M(u)$ d'auteurs susceptibles de collaborer avec u. Ici, nous avons choisi de déterminer pour chaque auteur la liste des 3 puis celle des 5 co-auteurs les plus similaires, les chiffres 3 et 5 étant choisis pour des raisons d'ergonomie afin que la recommandation puisse tenir sur un écran.

	Quant-ph		Gr-qc		Cond-mat		Hep-ph	
n	3	5	3	5	3	5	3	5
Common	913	1447	*1317*	2025	2863	4577	3208	5273
	3,61	2,83	**3,42**	2,62	4,23	3,45	5,58	4,68
Jaccard	920	1452	1307	*2017*	2773	4455	2977	4785
	2,93	2,69	3,06	**2,83**	3,25	2,76	3,16	2,88
Adamic	912	1458	1317	2029	2896	4622	3168	5165
	2,63	2,19	2,96	2,56	3,76	3,12	3,60	3,35
Attachement	1226	2039	1797	2991	3663	6096	3736	6210
	1,88	1,28	2,23	1,64	0,63	0,56	1,31	1,08
Katz	1007	1679	1309	2102	2519	3893	2811	4344
	2,68	1,97	2,98	2,24	3,61	2,95	3,27	2,85
Random	826	1304	1183	1880	2426	3853	2501	3929
	0,00	0,15	0,34	0,27	0,04	0,18	0,32	0,33
Filtering	1105	1836	1578	2576	3394	5299	3208	5273
	0,18	0,11	0,51	0,43	1,38	1,66	5,58	4,68
Abstract	858	1385	1236	1943	2810	4544	2728	4341
	0.58	0.36	0.00	0.05	0.60	0.40	0.15	0.23
Common*	*929*	*1455*	1317	2026	*2873*	*4613*	*3136*	*5103*
	3.77	**3.16**	3.19	2.42	**4.98**	**3.84**	**6.03**	**5.06**

TAB. 4: Résultats de la recommandation de 3 et 5 auteurs.

Le tableau 4 indique pour $n = 3$ et $n = 5$, la proportion de Vrais Positifs obtenus par le système de recommandation pour chaque méthode. L'utilisation de la proportion se justifie par le fait que deux modèles ne produisent pas forcement le même nombre de liens. Cette proportion est égale au nombre de liens correctement prédits sur le nombre total de liens prédits (entier sur la ligne précédant la proportion). Les valeurs en gras indiquent les valeurs de qualité maximales pour chaque section. Par exemple, pour la section *Quant-ph*, avec $n = 5$, la méthode *random* détermine 1304 liens parmi lesquels *0,15 %* sont des Vrais Positifs. La performance maximale est obtenue par la mesure *Common** qui détermine 1455 liens pour *3,16 %* de Vrais Positifs.

On en déduit que la mesure *Common** est meilleure sur les jeux de données obtenus des sections Quant-ph, Cond-mat et Hep-ph respectivement pour $n = 3$ et $n = 5$. Les mesures *Common* et *Jaccard* sont meilleures sur le jeu de données obtenu de la section Gr-qc. La méthode *Common** semble globalement être la mieux adaptée pour la recommandation même si, pour la section Gr-qc, elle est légèrement moins bonne de *Common* et *Jaccard*.

5 Conclusion

De nombreuses approches ont été développées dans la littérature pour résoudre le problème de prévision de liens dans les réseaux sociaux ayant une structure unipartite (Liben-Nowell et Kleinberg, 2003), (Hasan et al., 2006), mais très peu ont été proposées lorsque la structure utilisée pour représenter le réseau social est un graphe bipartite (Allali et al., 2011). Dans cet article, nous avons étudié le problème de prévision de liens dans les réseaux sociaux ayant une structure bipartite.

Nous avons proposé deux nouvelles méthodes topologiques, basées uniquement sur la structure du graphe. La première, *Common**, est une version pondérée de la méthode *Common* proposée par (Newman, 2001). La seconde, *Rules*, est basée sur les règles d'association liées aux voisinages des nœuds. Ces deux méthodes s'appliquent à tout graphe bipartite. La troisième méthode que nous avons présentée s'applique aux réseaux de collaborations scientifiques (ou autres réseaux associant auteurs et textes) et prend en compte les attributs des nœuds (calcul d'une similarité entre textes).

Ces trois nouvelles méthodes ont été implémentées et comparées aux meilleurs algorithmes connus, en utilisant les données extraites d'ArXiv. Les résultats montrent qu'elles permettent une amélioration d'AUC entre 6% et 16%.

Nous travaillons actuellement sur les techniques permettant de combiner ces méthodes pour obtenir des algorithmes encore plus performants.

Remerciements

Ce travail a été partiellement financé par les projets ANR ExDEUSS et FUI AMMICO.

Références

Adamic, L. A. et E. Adar (2003). Friends and neighbors on the web. *Social Networks 25*, 211–230.

Allali, O., C. Magnien, et M. Latapy (2011). Link prediction in bipartite graphs using internal links and weighted projection. *Proceedings of the third international workshop on Network Science for Communication Networks (NetSci- Com)*.

Hanley, J. A. et B. J. Mcneil (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology 148*(3), 839–843.

Hasan, M. A., V. Chaoji, S. Salem, et M. Zaki (2006). Link prediction using supervised learning. *Proceedings of SDM 06 workshop on Link Analysis, Counterterrorism and Security*.

Hastie, T., R. Tibshirani, et J. Friedman (2009). *The elements of statistical learning: data mining, inference and prediction* (2 ed.). Springer.

Jaccard, P. (1901). Etude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Societe Vaudoise des Science Naturelles 37*, 547.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychmetrika 18*, 39.

Krebs, V. (2002). Mapping networks of terrorist cells. *J. Amer. Soc. In- form. Sci. 27 24(3)*, 43–52.

Liben-Nowell, D. et J. Kleinberg (2003). The link prediction problem for social networks. *Proceedings of the twelfth international conference on Information and knowledge management, CIKM '03, New York, NY, USA,*, 556–559.

Macskassy, S. A. et F. Provost (2005). Suspicion scoring based on guilt-by-association, collective inference, and focused data access. *International conference on intelligence analysis*.

Manning, C. D. et H. Schütze (1999). *Foundations of statistical natural language processing*. Cambridge, MA, USA: MIT Press.

Mitzenmacher, M. (2001). A brief history of lognormal and power law distributions. *Proceedings of the Allerton Conference on Communication, Control, and Computing*, 182–191.

Newman, M. E. J. (2001). Clustering and preferential attachment in growing networks. *Phys. Rev. E 64*, 025102.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review 45*, 167–256.

Porter, M. (1980). An algorithm for suffix stripping. *Program 14*(3), 130–137. (We used the Python implementation avaible at https://pypi.python.org/pypi/stemming/1.0).

Price, D. J. d. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *J. Amer. Soc. In- form. Sci. 27*, 292–306.

Rajaraman, A. et J. D. Ullman (2010-2011). *Mining of Massive Datasets*.

Ricci, F., L. Rokach, B. Shapira, et P. B. Kantor (Eds.) (2011). *Recommender Systems Handbook*. Springer.

Summary

Social networks are very dynamic structures. A direction of research related to this dynamics is the links prediction problem, which, for example, for a clients/products network, is to predict the products that a customer is likely to purchase in the near future. In the so-called topological approach, the links prediction methods use only the structure of the graph. We are interested on bipartite graphs. So far, this problem has been solved by recommendation approaches such as collaborative filtering or using the projected graph. We propose methods that take into account the attributes of nodes and links. Specifically, we propose, for graphs of collaborations, a method that takes into account the abstracts. Then we introduce, for bipartite graphs, one approach based on association rules linked to the neighborhood of nodes. The evaluation on four sections of arXiv shows that these methods yield, compared to topological approaches and collaborative filtering, AUC improvement between 6% to 16%.