Enrichissement du profil utilisateur à partir de son réseau social dans un contexte dynamique : application d'une méthode de pondération temporelle

Marie-Françoise Canut, Sirinya On-at, André Péninou, Florence Sèdes

Institut de Recherche en Informatique de Toulouse (IRIT), Université de Toulouse, CNRS, INPT, UPS, UT1, UT2J, 31062 TOULOUSE Cedex 9
{Marie-Francoise.Canut, Sirinya.On-at, Andre.Peninou, Florence.Sedes}@irit.fr

RESUME. Le profil de l'utilisateur est un élément central dans les systèmes d'adaptation de l'information. Les réseaux sociaux numériques représentent une source d'informations très riche sur l'utilisateur. Nous nous intéressons au processus d'enrichissement du profil utilisateur à partir de son réseau social. Ce processus extrait les intérêts de l'utilisateur à partir des individus dans son réseau égocentrique afin de construire la dimension sociale du profil de l'utilisateur. Afin de prendre en compte le caractère dynamique des réseaux sociaux, nous proposons, dans ce travail, de construire cette dimension sociale en intégrant un critère temporel afin de pondérer les intérêts de l'utilisateur. Ce poids "temporel", qui reflète la pertinence d'un intérêt, est calculé, d'une part, à partir de la pertinence des individus du réseau égocentrique de l'utilisateur en prenant en compte la fraicheur de leurs liens avec l'utilisateur et, d'autre part, à partir de la pertinence des informations qu'ils partagent en prenant en compte la fraicheur de ces informations. Les expérimentations sur les réseaux de publications scientifiques DBLP et Mendeley ont permis de montrer que notre proposition fournit des résultats plus satisfaisants que ceux du processus existant.

ABSTRACT. The development of users' profiles is central for adaptive system to propose information that corresponds to user specific needs. Online social networks provide valuable information sources to collect and identify user's information and user's interests. This work focuses on extracting user's interests from his social egocentric network to build the social dimension of his user profile. To improve this social dimension, we take into account the dynamic of the social network by applying a time-weight strategy in order to drift the user interests. The time-weight of an interest is computed by combining the accuracy of individuals in his egocentric network (computed by taking into account the freshness of their ties) with the information accuracy (computed by taking into account its freshness). The experiments on scientific publications networks (DBLP/Mendeley) allowed us to demonstrate the effectiveness of our proposition comparing to the time-agnostic user profiling approach.

MOTS-CLES : Profil utilisateur, Réseau social, Réseau égocentrique, Pondération temporelle.

Keywords: User Profile, Social Network, Egocentric Network, Time-weight Method.

1. Introduction

Dans les systèmes d'adaptation de l'information, pour pouvoir offrir des informations qui correspondent aux besoins de l'utilisateur, les mécanismes de personnalisation/adaptation doivent disposer d'informations sur les utilisateurs telles que leurs caractéristiques personnelles, leurs préférences générales, leurs centres d'intérêt. De ce fait, le profil utilisateur construit à partir de ces informations devient central dans ce type de mécanisme. Dans nos travaux, c'est le processus de construction du profil utilisateur qui nous intéresse particulièrement. Dans la littérature, plusieurs modèles de profil utilisateur ont été proposés. Nous représentons le profil utilisateur sous forme de vecteur de termes pondérés qui décrivent les centres d'intérêt de l'utilisateur définis par un ensemble de mots clés (foot, tennis, danse,…).

Le contexte de nos travaux porte sur le processus de construction du profil utilisateur à partir de son réseau social représenté par un graphe de relations entre les individus. Nous travaillons en particulier dans le contexte des média sociaux qui s'appuient généralement sur des réseaux sociaux numériques. Ce processus de construction permet d'une part de compléter le profil d'un nouvel utilisateur ou celui d'un utilisateur peu actif pour lequel le profil serait insuffisant pour les mécanismes d'adaptation, et d'autre part d'enrichir un profil existant. Nous utilisons dans ce papier, le terme profil social pour désigner un profil construit par ce type de processus.

Nous partons du constat que les intérêts de l'utilisateur évoluent au fil du temps, en particulier, dans le cas des intérêts extraits depuis les médias sociaux. En effet, dans les média sociaux, l'information partagée évolue sans cesse du fait des interactions sociales en ligne (partage, échange d'informations) qui génèrent (plus facilement) un volume important d'informations volatiles. Pour un utilisateur, les intérêts qui sont extraits à une période donnée peuvent ne plus être significatifs ultérieurement. Par exemple, pour un utilisateur qui regarde le foot uniquement pendant la période de la coupe du monde (une fois tous les 4 ans) et qui partage des informations dans son réseau social, ses intérêts pour la coupe du monde ou le foot ne sont significatifs pour son profil que pendant cette période.

Un autre problème qui peut être rencontré est la pertinence des liens de l'utilisateur avec les membres de son réseau social ainsi que celle des informations qu'il partage. Les utilisateurs peuvent créer des contacts en ligne sans forcément connaître les personnes dans la vie réelle. Si on reprend l'exemple précédent, un utilisateur peut suivre les informations à partir des comptes de joueurs de foot qui sont alors enregistrés comme contacts sans pour autant les connaitre personnellement. Après la compétition, les liens avec ces joueurs deviennent moins importants pour l'utilisateur. Cela nous montre bien que l'on ne peut pas prendre en compte (ou donner la même importance à) toutes les informations existant dans les réseaux sociaux pour refléter les intérêts d'un utilisateur à un moment donné. Nous nous trouvons donc dans la problématique suivante : dans le contexte des média sociaux, comment construire un profil social de l'utilisateur qui soit à la fois pertinent et à jour ?

Pour répondre à ce problème, nous considérons que l'évolution des intérêts des utilisateurs pousse ces derniers à modifier leurs relations sociales et les informations

qu'ils partagent. L'évolution du réseau social reflète alors l'évolution des intérêts des utilisateurs. Nous proposons donc de prendre en compte la caractéristique dynamique (plus précisément des informations temporelles) du réseau social pour sélectionner efficacement les informations provenant de ce réseau afin de calculer un profil social pertinent et à jour. Nous envisageons donc dans ce travail, de construire la dimension sociale de l'utilisateur par l'intégration d'une mesure temporelle dans le calcul du poids (représentant la pertinence) des intérêts liés au profil social de l'utilisateur. Ce poids, appelé poids temporel, est calculé, d'une part, à partir du poids de pertinence temporelle des individus et du poids de pertinence temporelle des informations que ces individus partagent. Le poids de pertinence des individus est calculé en appliquant une méthode de prédiction de liens temporelle afin de sélectionner les individus ayant les liens les plus actifs avec l'utilisateur. Le poids de pertinence temporelle des informations partagées est calculé en prenant en compte la fraîcheur de ces informations afin de prendre en compte les informations les plus récentes et en extraire des intérêts pertinents et à jour.

Nous présentons tout d'abord un état de l'art sur la construction du profil utilisateur à partir de réseaux sociaux, puis nous insistons sur la caractéristique dynamique des intérêts de l'utilisateur et des réseaux sociaux. Nous présentons ensuite notre proposition permettant de prendre en compte cette double dynamique dans le processus de construction du profil social de l'utilisateur. Ensuite, nous décrivons nos expérimentations effectuées dans deux réseaux de publications scientifiques (DBLP et Mendeley). Nous terminons par une conclusion et les perspectives de ce travail.

2. Etat de l'art : dérivation du profil utilisateur à partir de réseaux sociaux

L'approche de construction du profil social de l'utilisateur où les intérêts de l'utilisateur sont enrichis à partir des informations partagées par les individus de son réseau social permet d'améliorer la représentation du profil utilisateur mais aussi de résoudre le problème de démarrage à froid ou le cas d'un utilisateur très peu actif dans le système (qui interagit moins avec le système et fournit donc moins d'informations pour déduire ses intérêts) (Massa et Avesani, 2007).

(Cabanac, 2011) s'appuie principalement sur des utilisateurs considérés individuellement pour proposer un système de recommandation sociale d'articles scientifiques aux chercheurs. (Carmel et al., 2009) utilise le même principe pour proposer un système de recherche d'information sociale. Ces méthodes peuvent être définies comme des méthodes de type « autoritaire» pour lesquelles les utilisateurs les plus actifs ou les plus influents dans le réseau social seront privilégiés dans la dérivation de leur profil social. Pourtant, dans le contexte des réseaux sociaux numériques, il est difficile de considérer que tous les centres d'intérêt d'un utilisateur influent dans le réseau social, peuvent représenter fidèlement cet utilisateur. Ce problème peut être mieux perçu si l'on considère des environnements de réseaux sociaux numériques comme Facebook dans lesquels un utilisateur peut être ami avec plus de 1000 individus. Parmi ces individus, très peu sont réellement ses amis proches qui permettraient de fournir des informations permettant de caractériser correctement cet utilisateur, même s'ils sont très actifs ou influents.

Les travaux de (Tchuente, 2013) portent sur l'étude de la dérivation du profil utilisateur à partir des communautés de son réseau égocentrique, un réseau largement utilisé en sociologie. Il s'agit d'un graphe composé des relations entre les individus situés à distance 1 (directement reliés) de l'utilisateur (appelé égo), l'égo étant bien entendu exclu de ce graphe. Ces travaux proposent donc une méthode plutôt de type « affinitaire», dans laquelle c'est la présence d'affinités, de liens, de relations entre les individus d'une communauté du réseau égocentrique de l'utilisateur qui permet de dériver des informations à associer à son profil. Par rapport aux méthodes autoritaires, cette approche éliminera, d'une part, les individus ayant des relations superflues avec l'utilisateur (amitiés acceptées au hasard ou pour démontrer un certain pouvoir par le nombre d'amis), et d'autre part, les éléments de profil non significatifs pour l'utilisateur. Cette étude propose tout d'abord un modèle générique du profil utilisateur suivant deux dimensions : une **dimension utilisateur** dont les centres d'intérêt sont calculés à partir des activités propres de l'utilisateur et **une dimension sociale** (profil social) dont les centres d'intérêt sont calculés à partir des activités de son réseau égocentrique. Ces deux dimensions étant complémentaires et indépendantes, elles peuvent être utilisées par les mécanismes d'adaptation, recommandation soit individuellement soit couplées. L'étude présente ensuite un processus de dérivation de la dimension sociale du profil utilisateur à partir des communautés extraites de son réseau égocentrique (détaillé dans la section 5).

Toutefois, la gestion dynamique du profil utilisateur est un problème soulevé et non traité dans le travail existant de (Tchuente, 2013). En effet, les centres d'intérêt d'une personne sont amenés à évoluer dans le temps (changement d'environnement ou de contexte du travail). Nos travaux s'appuient sur la proposition de (Tchuente, 2013), en tentant de prendre en compte la gestion dynamique du profil utilisateur dans le processus de dérivation de la dimension sociale de ce profil. Il nous parait alors important d'étudier l'évolution de ce profil en fonction de l'évolution du réseau social de cet utilisateur. Nous nous sommes alors intéressés à ces deux types d'évolution.

3. Gestion de l'évolution des intérêts dans le profil utilisateur

L'étude de l'évolution des intérêts de l'utilisateur consiste à prendre en compte le changement de ses intérêts à travers le temps (Crabtree et al., 1998). A partir des travaux dans la littérature, nous pouvons distinguer deux approches de gestion de l'évolution des intérêts dans le profil utilisateur. La première approche consiste à gérer la dynamique des intérêts de l'utilisateur après la phase d'extraction des intérêts et correspond à un processus de mise à jour du profil utilisateur. La deuxième approche consiste à prendre en compte la dynamique des centres d'intérêt pendant l'étape d'extraction des intérêts.

Nous nous intéressons dans ce travail à la deuxième approche, dans le but de construire un profil social pertinent dès sa première utilisation. Dans ce type d'approche, les modèles de profil utilisateur utilisés sont, soit des modèles à court terme où le profil utilisateur se construit et s'utilise en exploitant les informations récentes de l'utilisateur, soit des modèles à long terme où les intérêts sont extraits et

enrichis à travers le temps. Dans la recherche d'information personnalisée, on utilise l'historique à court terme de l'utilisateur lié à une seule (la dernière) session de recherche pour extraire ses intérêts (Bennett et al., 2012). Avec le même principe, plusieurs travaux proposent une approche utilisant un critère temporel pour mieux cerner l'évolution et la dynamique des informations étudiées. La plupart de ces travaux se basent sur l'approche « time-forgotten » qui ignore les informations trop anciennes (Cheng et al., 2008; Maloof et Michalski, 2000). Dans ce type d'approche, on oublie complètement les informations dépassant une date limite. Pourtant, certaines de ces informations ignorées peuvent être utiles et ne pas les prendre en compte peut entraîner une perte d'informations intéressantes. En effet, (Tan et al., 2006) ont prouvé que l'historique de recherche à long terme est très important pour améliorer la tâche de recherche d'informations dans le cas de requêtes récurrentes.

Dans l'utilisation de modèles à long terme, toutes les informations de l'utilisateur sont conservées (et peuvent contenir éventuellement des biais), il est alors difficile de sélectionner les informations pertinentes pour représenter l'utilisateur à un instant donné. Il se peut que des intérêts anciens de l'utilisateur ne soient plus significatifs à ce jour. Cette remarque peut être retrouvée dans (Kacem et al., 2014; Li et al., 2013) qui proposent d'appliquer une fonction temporelle pour pondérer les intérêts de l'utilisateur selon leur fraicheur. Cette idée peut être retrouvée également dans le contexte de la construction du profil utilisateur à partir d'un réseau d'annotations comme dans (Zheng et Li, 2011) qui utilise des fonctions temporelles pour pondérer des tags avant d'en extraire les intérêts de l'utilisateur. Dans ce type d'approche, toutes les informations existantes de l'utilisateur sont exploitées mais de manière plus restreinte. Nous nous intéressons dans ce travail à cette dernière approche.

Il est important de noter que dans nos travaux, les principales interactions que nous souhaitons étudier pour détecter un changement de centres d'intérêt ne sont pas ciblées sur l'utilisateur lui-même mais sur les éléments de son réseau social (liens entre les membres, informations qui circulent entre les membres). L'évolution de la dimension sociale du profil utilisateur est liée à l'évolution de son réseau social. Pour étudier l'évolution des intérêts de l'utilisateur dans notre contexte, il est donc important de prendre en compte des informations temporelles en s'appuyant sur les mécanismes d'évolution du réseau social de l'utilisateur.

4. Evolution du réseau social

Depuis plusieurs années, l'étude sur les propriétés et les caractéristiques des réseaux sociaux (densité, degré de distribution, classification, composants connexes, communautés, …) a été considérée comme une piste importante de recherche. Cependant, la plupart des études ont été conduites avec une vision statique du réseau alors qu'un réseau social numérique est considéré comme un réseau dynamique (qui évolue au fil du temps). Pour répondre à ce problème, l'analyse de la dynamique du réseau social ne peut s'effectuer qu'en prenant en compte un critère temporel pour comprendre les évolutions qui se produisent dans le réseau (Spiliopoulou, 2011). Nous pouvons distinguer deux types de dynamique dans un réseau social : la dynamique de la structure du réseau et la dynamique des informations partagées.

4.1 Dynamique de la structure du réseau social (dynamique dans le réseau)

La dynamique de la structure du réseau social provient de la dynamique des interactions des individus dans le réseau, liée à la création et/ou à la suppression des liens mais aussi à la persistance des liens déjà créés. L'analyse de l'évolution de la structure du réseau social porte sur la visualisation de l'évolution du réseau, sur le modèle de l'évolution de la structure du réseau pour prédire comment le réseau va évoluer à travers le temps (Kumar et al., 2006; Leskovec et al., 2008) et sur la prédiction de liens (Liben-Nowell et Kleinberg, 2003; Hasan et Zaki, 2011) qui se focalise sur la formation de liens entre les nœuds dans le réseau. Etant donné un réseau social, on détermine, pour chaque paire de nœuds, s'il y a une possibilité de formation de lien entre eux dans le futur.

4.2 Dynamique des informations dans le réseau (dynamique sur le réseau)

Le deuxième facteur important lié à l'analyse temporelle des informations dans le réseau social porte sur le partage et la diffusion d'informations. L'analyse de la dynamique des informations porte sur le modèle de diffusion d'informations, sur la recherche des nœuds influents dans le réseau mais aussi sur les techniques permettant de diffuser efficacement des informations dans le réseau social (Jiang et al., 2014). D'après plusieurs travaux de recherche, la structure du réseau a beaucoup d'impact sur la dynamique des informations dans le réseau. Les individus qui sont considérés proches dans le réseau ont une probabilité plus grande de diffuser des informations entre eux. Dans le cas inverse, la dynamique de la diffusion d'informations peut également être à son tour, un facteur pour le changement de la structure du réseau social (Stattner et al., 2013; Weng et al., 2013).

Plusieurs travaux sur l'évolution du réseau social ne considèrent pas le changement d'informations dans le réseau comme un facteur important pour l'analyse du réseau. Toutefois, dans le contexte des média sociaux dont les membres créent et partagent une masse d'informations importante, les données circulant dans le réseau deviennent volatiles et évoluent rapidement. Il existe des facteurs comme les « Buzz », les campagnes de marketing et les événements remarquables qui peuvent provoquer des mouvements sociaux. De tels facteurs peuvent augmenter la croissance de la diffusion de l'information dans les médias sociaux mais aussi la création de nouveaux liens entre les personnes qui s'intéressent à l'événement. Pourtant, les informations partagées sont souvent temporaires et disparaissent quand l'événement se termine, par exemple, la guerre civile libyenne, la révolution égyptienne de 2011 (Gomez Rodriguez et al., 2013). Les techniques existantes de traitement des informations doivent être adaptées aux caractéristiques de ces nouvelles sources d'informations afin d'obtenir des informations qui soient les plus représentatives de l'utilisateur. Il parait donc important de prendre en compte non seulement la dynamique de la structure du réseau mais également la dynamique des informations dans l'analyse de l'évolution du réseau.

Les caractéristiques dynamiques des media sociaux nous amènent à considérer qu'il n'est pas certain que les individus du réseau social de l'utilisateur ainsi que les

informations qu'ils partagent soient toutes pertinentes et à jour. Notre proposition prend en compte ces remarques dans le processus de dérivation de la dimension sociale du profil utilisateur.

5. Proposition

Notre travail utilise des critères temporels pour prendre en compte la dynamique de la structure du réseau mais aussi la dynamique des informations échangées dans le processus de construction/dérivation de la dimension sociale l'utilisateur. Nous nous appuyons sur les travaux de (Tchuente 2013). Nous expliquons brièvement le processus de dérivation de la dimension sociale (CoBSP) proposé par ces travaux qui se fait en 4 étapes successives :

1) La première étape consiste à extraire, depuis le réseau égocentrique d'un utilisateur, les communautés de ce réseau en utilisant l'algorithme iLCD proposé par (Cazabet et al., 2012).

2) La deuxième étape consiste à calculer le profil de chaque communauté détectée dans la première étape. Le profil d'une communauté peut être calculé en utilisant les informations de tous les individus qui en font partie (ici, les poids associés aux éléments présents dans la dimension utilisateur de chaque individu).

3) La troisième étape consiste à caractériser chaque communauté en se basant sur une caractérisation structurelle et/ou une caractérisation sémantique. La caractérisation structurelle se base sur la centralité de degré (degree centrality). Selon cette mesure, la communauté qui possède le plus grand nombre de connexions directes dans le réseau est caractérisée comme la plus importante par rapport aux autres communautés. La caractérisation sémantique d'une communauté consiste à rechercher sa spécificité par rapport aux autres communautés en se basant sur la mesure de pondération tf-idf. Les deux caractérisations peuvent être combinées pour obtenir une caractérisation unique (sémantico-structurelle). La caractérisation sémantico-structurelle de chaque élément e du profil d'une communauté C1 sera calculée par la formule suivante :

$$\text{Caractérisation finale}(e) = \alpha\text{Struct}(C1) + (1-\alpha)\text{Sem}(C1) \quad (1)$$

Le paramètre α (valeur comprise entre [0,1]) dans la formule permettra de juger et de faire varier l'importance des mesures de structure dans la dérivation de la dimension sociale du profil de l'utilisateur par rapport aux mesures sémantiques. La valeur de α est déterminée de manière empirique et sera donc fixée lors des expérimentations.

4) La quatrième étape permet de dériver les intérêts de la dimension sociale par combinaison des différents poids associés à un intérêt à partir de toutes les communautés en utilisant une fonction linéaire.

Nous proposons dans ce travail de prendre en compte le critère temporel dans l'étape de pondération des intérêts de l'utilisateur. Notre travail se situe après l'étape 3 de (Tchuente, 2013) où nous intégrons un poids temporel (**structuro-semantico-temporel)** dans l'étape de calcul de la pondération des intérêts de l'utilisateur.

Notre proposition se divise en 2 parties. La première partie consiste à attribuer aux informations, un poids prenant en compte le critère temporel (poids temporel) (qui correspond à la prise en compte de la dynamique de la structure du réseau et des informations dans le réseau). La deuxième partie explique comment exploiter ce poids temporel, lors de l'étape de construction de la dimension sociale de l'utilisateur.

5.1 Calcul du poids structuro-semantico-temporel

Le poids structuro-semantico-temporel est calculé en combinant d'une part, le poids structurel du réseau égocentrique qui prend en compte la dynamique de la structure du réseau (appelé poids structuro-temporel : StrTem) et d'autre part le poids de pertinence de l'information utilisée pour extraire les intérêts (appelé sémantico-temporel : SemTem). Nous détaillons dans la suite, le calcul de ces deux poids ainsi que leur combinaison.

5.1.1 Calcul du poids structuro-temporel (StrTem) avec une méthode de prédiction de liens temporelle

Il s'agit ici de la prise en compte de la dynamique de la structure du réseau. Comme les liens entre l'utilisateur et les individus dans son réseau égocentrique peuvent varier au fil de temps, nous considérons que les individus qui ont les relations les plus récentes avec l'utilisateur ont une probabilité plus grande de partager les mêmes intérêts avec l'utilisateur (plus significatifs). Nous donnons donc un poids plus important aux informations partagées par ces individus par rapport aux individus ayant des liens moins récents. Il s'agit de calculer la pertinence d'un individu par rapport à l'utilisateur central en prenant en compte les informations temporelles de ses liens avec l'utilisateur central (date de création de liens avec l'utilisateur central, durée de la relation, …).

La technique de prédiction de liens temporelle nous semble être adaptée à ce problème. Le principe de la prédiction de liens est de calculer la similarité entre les deux nœuds en utilisant des critères tels que la topologie du réseau par exemple. Selon le score de similarité obtenu, on obtient la probabilité que ces nœuds se connectent dans le futur. Contrairement à son utilisation initiale, nous utilisons dans notre travail, la prédiction de lien pour calculer le poids de similarité entre deux nœuds déjà connectés afin de représenter la persistance du lien entre ces nœuds dans le futur. La plupart des travaux s'intéressent à la prédiction de liens sans tenir compte de l'évolution du réseau. Cependant, comme introduit précédemment, la prise en compte du temps est essentielle pour un réseau social numérique. Nous nous intéressons donc à la prédiction de liens temporelle qui prend en compte les informations structurelles et temporelles lors du calcul du score de similarité entre deux nœuds.

Nous nous appuyons sur le travail de (Tylenda et al., 2009) qui proposent d'appliquer des critères temporels (durée d'existence de la dernière collaboration entre deux nœuds étudiés.) aux méthodes de base de prédiction de liens comme Adamic/Adar, introduit dans (Liben-Nowell et Kleinberg, 2003). Adamic/Adar se base sur le nombre de voisins communs. Pour une paire de nœuds (x,y), la formule du calcul de score de similarité de Adamic/Adar est :

$$AA(x,y) = \sum_{z \epsilon \Gamma x \cap \Gamma y} \frac{1}{log|\Gamma(z)|} \quad (2)$$

où Γ(x), Γ(y) sont les ensembles des voisins de x et y respectivement, z est un noeud voisin commun de x et y et Γ(z) l'ensemble de ses voisins. La formule de Adamic/Adar qui prend en compte les informations temporelles (Tylenda et al., 2009) est :

$$AA(x,y) = \sum_{z \epsilon \Gamma x \cap \Gamma y} \frac{w(x,z) \cdot w(z,y)}{log|\Gamma(z)|} \quad (3)$$

où w(x,y) représente le poids temporel entre deux nœuds x et y (selon la date de leur dernière interaction). Pour calculer ce poids w(x,y), nous utilisons le travail de (Ding et Li, 2005) qui pondère une information selon sa date de publication (plus l'information est récente, plus elle est importante). Mais nous appliquons cette fonction temporelle à la dernière interaction entre les deux nœuds.

$$f(t) = e^{-\lambda t} \quad (4)$$

Pour chaque t_i ($i \in N$), t_0 est considéré comme l'instant le plus récent et ainsi de suite (ex : t_0 pour l'année 2014, t_1 pour l'année 2013, …). La valeur λ, fixée de manière empirique lors des expérimentations, est le taux de décroissance des valeurs. Plus λ est petit, plus les informations récentes sont importantes. La valeur de w(x,y) est calculée en fonction de f(t), t représentant l'estampille de la dernière interaction entre w et y.

$$w(x,y) = f(t) \quad (5)$$

5.1.2 Poids sémantico temporel (SemTem)

Nous donnons également de l'importance à la fraicheur des informations. Comme nous considérons que les intérêts de l'utilisateur doivent évoluer selon les informations partagées dans son réseau, les plus récentes auront plus d'importance que les plus anciennes. A partir des informations que partage chaque individu du réseau égocentrique de l'utilisateur, nous proposons de pondérer leur poids de pertinence en intégrant des informations temporelles. Nous pondérons les intérêts extraits de cette information avec la mesure de pondération tf-idf qui est une mesure de fréquence de terme et nous appliquons la même fonction temporelle exponentielle (5) au poids des informations. Nous appliquons cette fonction au poids de chaque information selon sa fraîcheur.

$$\text{SemTem(i)} = \text{score\_tf-idf} * \text{score\_temporel(i)} \quad (6)$$

5.1.3 Poids final (Poids structuro-sémantico temporel)

Pour obtenir le poids final d'une information, le poids de la caractérisation structurelle et le poids de la caractérisation sémantique sont combinés. Le poids d'un intérêt i extrait depuis un individu **ind** du réseau égocentrique d'un utilisateur sera calculé par la formule suivante :

$$\text{Poids final (i)} = \gamma(\text{StrTem(ind)}) + (1-\gamma)(\text{SemTem(i)}) \quad (7)$$

Le paramètre γ (compris entre [0,1]) dans la formule permettra de juger et de faire varier l'importance du poids structuro-temporel dans la dérivation de la dimension sociale du profil de l'utilisateur par rapport au poids sémantico-temporel. Plus la valeur de γ est importante plus on donne de l'importance au poids structuro- temporel. La valeur de γ sera fixée de manière empirique lors des expérimentations.

5.2 Construction de la dimension sociale de l'utilisateur selon le poids de pertinence temporelle des sources et des informations

Pour construire le profil social de l'utilisateur en appliquant les techniques de prédiction de liens temporelle et de pondération d'informations décrites précédemment, nous utilisons le même principe (CoSPK) proposé par (Tchuente, 2013) mais dans l'étape d'extraction des intérêts de l'utilisateur, nous pondérons un terme extrait avec le poids temporel associé calculé par l'équation (7). Cette méthode consiste à ajouter une nouvelle signification aux poids des centres d'intérêt du profil de l'utilisateur avant de les exploiter dans la construction/enrichissement du profil de l'utilisateur (il s'agit d'augmenter ou de diminuer l'importance des centres d'intérêt). Pour ce faire, premièrement, nous calculons le poids de pertinence comme indiqué dans 5.1, deuxièmement, nous classifions les intérêts de manière croissante par rapport à ce poids, enfin nous utilisons la formule de combinaison linaire pour pondérer chaque intérêt comme indiqué dans l'étape 4 du processus CoSPK.

Pour étudier l'impact de la prise en compte de la dynamique des informations par rapport à la dynamique de la structure du réseau, nous proposons deux algorithmes différents associés à cette démarche de construction du profil social de l'utilisateur. Nous appelons CoSPKTP, l'algorithme qui utilise seulement le poids structuro-temporel pour pondérer les informations traitées et nous appelons CoSPKTPI, l'algorithme qui utilise le poids structuro-sémantico temporel pour pondérer les informations traitées.

6. Expérimentation

Pour valider notre proposition, nous avons comparé la pertinence de notre approche par rapport à l'approche existante. Pour cela, nous comparons la pertinence des dimensions sociales construites par ces approches. Le domaine d'expérimentation choisi concerne les réseaux d'auteurs de publications scientifiques (DBLP et Mendeley). Dans ces réseaux, les noeuds représentent les auteurs. Deux noeuds peuvent être reliés par un lien s'ils publient ensemble. Nous exploitons le réseau des auteurs DBLP pour lesquels on peut calculer les centres d'intérêt à partir des titres de leurs publications pour construire le réseau égocentrique de chaque auteur et dériver la dimension sociale de son profil. Pour calculer la pertinence de la **dimension sociale** créée, nous la comparons avec la **dimension utilisateur** qui contient les intérêts réels de l'utilisateur que nous créons à partir des intérêts indiqués explicitement par l'utilisateur sur son profil Mendeley. L'objectif d'une telle expérimentation est de

montrer que la dimension sociale construite indépendamment des informations propres à l'utilisateur est suffisamment représentative de l'utilisateur et pourra donc être utilisée en particulier lorsque la dimension utilisateur n'existe pas ou est incomplète.

Nous partons des travaux existants pour construire les dimensions sociales et la dimension utilisateur du profil de l'utilisateur. Comme présenté dans la figure1, nous recherchons dans un premier temps, les auteurs de publications scientifiques dans Mendeley ayant plus de 6 centres d'intérêt. Ensuite, nous partons de ces auteurs pour rechercher leurs profils dans DBLP à condition qu'ils aient au moins 50 co-auteurs dans ce dernier. Notre échantillon de test porte sur 79 auteurs. Nous construisons ensuite pour un auteur, les différentes dimensions sociales à partir de ses données DBLP (issues des différentes approches de construction de la dimension sociale) et la dimension utilisateur à partir des intérêts indiqués dans son profil Mendeley. Enfin nous évaluons les dimensions sociales construites par les mesures de rappel et de précision.

6.1 Construction de la dimension sociale du profil utilisateur

Dans cette étape, nous construisons 3 dimensions sociales avec les différents algorithmes (CoSPK pour l'approche existante, CoSPKTP et CoSPKTPI pour l'approche proposée).

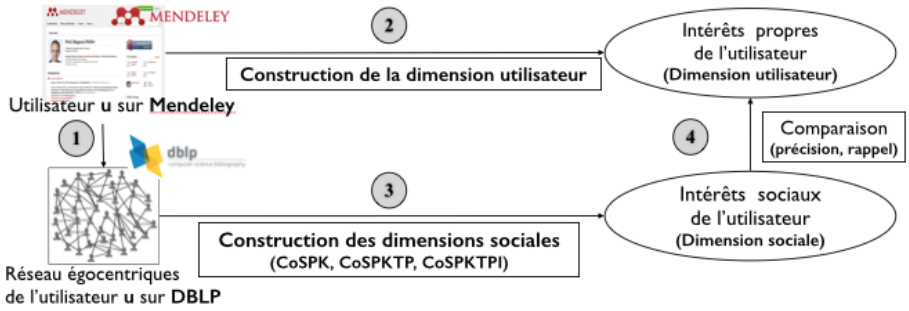

Figure1. Illustration de la construction de la dimension sociale et de la dimension utilisateur depuis des données DBLP et Mendeley.

6.1.1 Construction de la dimension sociale avec l'algorithme CoSPK

Comme décrit précédemment, le réseau égocentrique d'un auteur (ego) est construit dans DBLP à partir des relations entre ses co-auteurs. La première étape consiste donc à récupérer les co-auteurs de cet ego. Nous stockons ainsi la liste des publications de tous les auteurs qui font partie de son réseau égocentrique. Ceci permet de récupérer le titre des publications des communautés. Nous analysons ensuite les titres de publications pour extraire des termes significatifs qui seront considérés comme les centres d'intérêt de l'utilisateur. Les dictionnaires utilisés dans ce

processus sont les dictionnaires de synonymes regroupant tous les synonymes d'un mot. Ces synonymes seront ensuite considérés comme des occurrences de ce mot. Les filtres permettent de sélectionner les mots les plus importants et d'en exclure certains moins importants et moins significatifs tels que des mots vides. Seuls les mots retenus seront considérés comme les centres d'intérêt de l'utilisateur. Ensuite, nous exploitons les mesures sémantiques et les mesures structurelles et représentons les centres d'intérêt de cette dimension sociale par un vecteur de termes pondérés (Tchuente 2013).

6.1.2 Construction de la dimension sociale en prenant en compte les informations temporelles (CoSPKTP et CoSPKTPI)

Pour construire les deux dimensions sociales, nous utilisons le même principe que celui de la construction de la dimension sociale CoSPK. Par contre, dans l'étape d'extraction des intérêts de l'utilisateur, le poids d'un terme extrait correspond pour CoSPKTP, à son poids sémantico-structurel calculé par la formule (3) et pour CoSPKTPI, à son poids temporel final calculé par la formule (7).

6.2 Construction de la dimension utilisateur du profil utilisateur à partir de Mendeley

Pour construire la dimension utilisateur, nous utilisons les intérêts indiqués explicitement par l'utilisateur dans son profil Mendeley. En appliquant le même traitement que celui de la construction de la dimension sociale, nous utilisons les dictionnaires et les filtres pour extraire des termes significatifs qui seront considérés comme les centres d'intérêt dans la dimension utilisateur. A la différence de la dimension sociale, la pondération de chaque mot est définie uniquement par la mesure de fréquence tf dans le texte entier analysé. Les centres d'intérêt de cette dimension seront également représentés par un vecteur de termes pondérés.

6.3 Evaluation

Après l'étape de construction de la dimension sociale et de la dimension utilisateur du profil utilisateur, nous obtenons 3 dimensions sociales différentes construites par les algorithmes CoSPK, CoSPKTP, CoSPKTPI. Les dimensions sociales construites seront comparées à la dimension utilisateur via les mesures de précision et de rappel. Dans notre contexte d'évaluation, la précision d'un algorithme de dérivation de la dimension sociale est évaluée par le nombre de centres d'intérêt prédits qui sont trouvés dans la dimension utilisateur par rapport au nombre total des centres d'intérêt calculés dans la dimension sociale.

$$Précision = \frac{nombre\ de\ centres\ d'intérêt\ prédits\ dans\ la\ dimension\ utilisateur}{nombre\ de\ centres\ d'intérêt\ calcutés\ dans\ la\ dimension\ sociale} \quad (8)$$

Le rappel d'un algorithme de dérivation de la dimension sociale quant à lui est évalué par le nombre de centres d'intérêt trouvés dans la dimension utilisateur par rapport au nombre de centres d'intérêt de la dimension utilisateur.

$$Rappel = \frac{nombre\ de\ centres\ d'intérêt\ prédits\ dans\ la\ dimension\ utilisateur}{nombre\ total\ de\ centres\ d'intérêt\ dans\ la\ dimension\ utilisateur} \quad (9)$$

Pour calculer la précision et le rappel, nous nous intéressons uniquement aux centres d'intérêt les plus pertinents renvoyés par chaque algorithme de dérivation de la dimension sociale. Si la dimension utilisateur d'un profil possède n centres d'intérêt, la précision et le rappel de chaque algorithme de dérivation de la dimension sociale seront calculés à partir du top n+m premiers centres d'intérêt, donc un vecteur de taille n+m pour représenter la dimension sociale.

6.4. Résultats

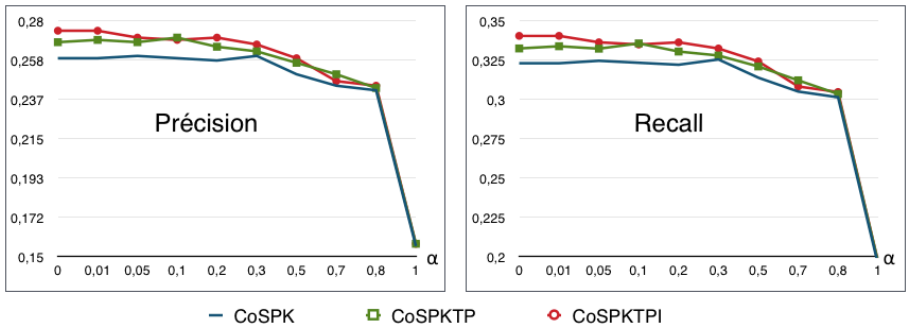

Figure2 : Comparatif de la pertinence des dimensions sociales construites avec CoSPK, CoSPKTP et CoSPKTPI avec la dimension utilisateur

La figure 2 représente les résultats permettant de comparer la pertinence des différentes dimensions sociales avec le top 10 des centres d'intérêt de l'utilisateur. Dans cette figure, la valeur α représente la proportion entre la mesure sémantique et la mesure structurelle comme présentée dans la formule (1). Pour obtenir les résultats de la figure 2, après plusieurs expérimentations avec différentes valeurs de γ et λ, nous avons calculé les poids temporels dans la formule (7) avec la valeur de $\gamma = 0,75$ et $\lambda = 0,2$ calculée avec la formule $(1/((0,5)*f(0))$ où $f(0) = 10$. Nous pouvons observer que l'algorithme CoSPKTP produit de meilleurs résultats que l'algorithme CoSPK en termes de précision et de rappel. Le meilleur résultat peut être observé quand $\alpha = 0,1$ avec successivement 4,3 % et 3,8% de taux de gain en terme de précision et rappel par rapport à l'algorithme CoSPK. Cela nous montre dans un premier temps, l'intérêt d'utiliser une méthode de prédiction de liens temporelle dans le processus de construction de la dimension sociale du profil utilisateur à partir de son réseau social.

Quand nous appliquons l'algorithme CoSPKTPI, il produit de meilleurs résultats par rapport à l'algorithme CoSPKTP et donne en conséquence un écart plus important comparé à l'algorithme CoSPK, en particulier quand $\alpha \in [0 ; 0,01]$, avec successivement 5,8 % et 5,4% de taux de gain en terme de précision et de rappel par

rapport à l'algorithme CoSPK et 1,9% et 2,0 % par rapport à l'algorithme CoSPKTP. Cela nous montre l'importance de la prise en compte de critères temporels non seulement dans la sélection de sources d'informations liées à la structure du réseau social de l'utilisateur mais aussi dans le traitement des informations partagées dans le réseau. Ces résultats nous permettent également de renforcer l'importance de la prise en compte de la dynamique du partage et de la diffusion d'informations tout comme de la dynamique de la structure du réseau dans l'étude de l'évolution du réseau social de l'utilisateur.

Algorithme	CoSPK		CoSPKTP	
	Précision	Rappel	Précision	Rappel
CoSPKTP ($\alpha = 0,1$)	4,3%	3,8%	-	-
CoSPKTPI ($\alpha \in [0 ; 0,01]$)	5,8%	5,4%	1,9%	2,0%

Tableau 1. Taux de gain (écarts entre les courbes) des algorithmes proposés (CoSPKTP et CoSPKTPI) par rapport l'algorithme existant (CoSPK)

6. Conclusion et perspectives

Dans ce travail, nous proposons de prendre en compte la caractéristique dynamique des réseaux sociaux dans l'étape d'extraction des intérêts du processus de construction de la dimension sociale du profil de l'utilisateur. Dans ce processus, nous intégrons des critères temporels (poids temporel) pour pondérer les individus et les informations du réseau social de l'utilisateur afin de pouvoir extraire des intérêts de l'utilisateur pertinents et à jour. Ce poids temporel est calculé, à partir du poids de pertinence temporelle des individus et du poids de pertinence temporelle des informations que ces individus partagent. Le poids de pertinence des individus est calculé en appliquant une méthode de prédiction de liens temporelle afin de sélectionner les individus ayant les liens les plus actifs avec l'utilisateur. Le poids de pertinence temporelle des informations partagées est calculé en prenant en compte la fraîcheur des informations. Les résultats des expérimentations nous ont permis de montrer l'efficacité de cette approche par rapport au processus de construction de la dimension sociale qui ne tient pas compte de critères temporels.

A court terme, nous envisageons d'appliquer cette approche dans d'autres types de réseaux sociaux où la caractéristique dynamique est très importante (Facebook, Twitter, …) pour effectuer une évaluation de notre proposition à plus grande échelle. Nous souhaitons également étudier d'autres algorithmes de prédiction de liens temporelle ainsi que d'autres fonctions temporelles afin de tenter d'améliorer la performance de notre approche. A plus long terme, nous envisageons la mise à jour continue du profil utilisateur afin d'avoir un profil pertinent et à jour à tout moment de son utilisation par des mécanismes d'adaptation de l'information par exemple.

Remerciements

Projet Subventionné par la Communauté de travail pyrénéenne

Bibliographie

Bennett, P. N., R. W. White, W. Chu, S. T. Dumais, P. Bailey, F. Borisyuk, et X. Cui. (2012). Modeling the Impact of Short- and Long-term Behavior on Search Personalization. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 185–194. New York, NY, USA: ACM.

Cabanac, G. (2011). Accuracy of inter-researcher similarity measures based on topical and social clues. *Scientometrics 87*, 597-620.

Carmel, D., N. Zwerdling, I. Guy, S. Ofek-Koifman, N. Har'el, I. Ronen, et al. (2009). Personalized Social Search Based on the User's Social Network. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, p. 1227–1236. New York, NY, USA: ACM.

Cheng, Y., G. Qiu, J. Bu, K. Liu, Y. Han, C. Wang, et C. Chen. (2008). Model Bloggers' Interests Based on Forgetting Mechanism. In *Proceedings of the 17th International Conference on World Wide Web*, p. 1129–1130. New York, NY, USA: ACM.

Crabtree, B., S. Soltysiak, M. Pp, et I. Re. (1998). Identifying and tracking changing interests. *International Journal on Digital Libraries 2*, 38–53.

Ding, Y. et X. Li. (2005). Time Weight Collaborative Filtering. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, p. 485–492. New York, NY, USA: ACM.

Gomez Rodriguez, M., J. Leskovec, et B. Schölkopf. (2013). Structure and Dynamics of Information Pathways in Online Media. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, p. 23–32. New York, NY, USA: ACM.

Hasan, M. A. et M. J. Zaki. (2011). A Survey of Link Prediction in Social Networks. In C. C. Aggarwal (Ed.), *Social Network Data Analytics*, p. 243-275. Springer US.

Jiang, C., Y. Chen, et K. J. R. Liu. (2014). Modeling information diffusion dynamics over social networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 1095-1099.

Kacem, A., M. Boughanem, et R. Faiz. (2014). Time-Sensitive User Profile for Optimizing Search Personlization. In V. Dimitrova, T. Kuflik, D. Chin, F. Ricci, P. Dolog, et G.-J.

Houben (Eds.), *User Modeling, Adaptation, and Personalization*, p. 111-121. Springer International Publishing.

Kumar, R., J. Novak, et A. Tomkins. (2006). Structure and Evolution of Online Social Networks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 611–617. New York, NY, USA: ACM.

Leskovec, J., L. Backstrom, R. Kumar, et A. Tomkins. (2008). Microscopic Evolution of Social Networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 462–470. New York, NY, USA: ACM.

Liben-Nowell, D. et J. Kleinberg. (2003). The Link Prediction Problem for Social Networks. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, p. 556–559. New York, NY, USA: ACM.

Li, D., P. Cao, Y. Guo, et M. Lei. (2013). Time Weight Update Model Based on the Memory Principle in Collaborative Filtering. *Journal of Computers 8*.

Maloof, M. A. et R. S. Michalski. (2000). Selecting Examples for Partial Memory Learning. *Machine Learning 41*, 27-52.

Massa, P. et P. Avesani. (2007). Trust-aware Recommender Systems. In *Proceedings of the 2007 ACM Conference on Recommender Systems*, p. 17–24. New York, NY, USA: ACM.

Spiliopoulou, M. (2011). Evolution in Social Networks: A Survey. In C. C. Aggarwal (Ed.), *Social Network Data Analytics*, p. 149-175. Springer US.

Stattner, E., M. Collard, et N. Vidot. (2013). D2SNet: Dynamics of diffusion and dynamic human behaviour in social networks. *Computers in Human Behavior 29*, 496-509.

Tan, B., X. Shen, et C. Zhai. (2006). Mining Long-term Search History to Improve Search Accuracy. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 718–723. New York, NY, USA: ACM.

Tchuente, D. (2013, janvier 28). *Modélisation et dérivation de profils utilisateurs à partir de réseaux sociaux : approche à partir de communautés de réseaux k-égocentriques* (phd). Université de Toulouse, Université Toulouse III - Paul Sabatier.

Tylenda, T., R. Angelova, et S. Bedathur. (2009). Towards Time-aware Link Prediction in Evolving Social Networks. In *Proceedings of the 3rd Workshop on Social Network Mining and Analysis*, p. 9:1–9:10. New York, NY, USA: ACM.

Weng, L., J. Ratkiewicz, N. Perra, B. Gonçalves, C. Castillo, F. Bonchi, et al. (2013). The Role of Information Diffusion in the Evolution of Social Networks. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 356–364. New York, NY, USA: ACM.

Zheng, N. et Q. Li. (2011). A recommender system based on tag and time information for social tagging systems. *Expert Systems with Applications 38*, 4575-4587.