Réconciliation des profils dans les réseaux sociaux

Nacéra Bennacer*, Coriane Nana Jipmo*, Antonio Penta**, Gianluca Quercini*

*Supélec E3S
{nacera.bennacer,coriane.nana-jipmo,gianluca.quercini}@supelec.fr
**Università di Torino
penta@di.unito.it

Résumé. Avec l'arrivée du Web 2.0, on assiste à un foisonnement de services de réseautage social, qui mettent l'utilisateur au centre des préoccupations. Ces services permettent de partager des ressources (YouTube, Flickr, Del.icio.us), d'échanger des informations et de construire des relations personnelles ou professionnelles (Facebook, LinkedIn) ou encore de diffuser des news (Twitter, blogs). Les utilisateurs disposent ainsi de plusieurs espaces d'informations sur différents réseaux sociaux. Ces espaces permettent souvent l'accès à des informations complémentaires sur l'identité de l'utilisateur, ses relations avec les autres et les ressources qu'ils partagent, qui pourraient surprendre l'utilisateur lui-même s'il en connaissait réellement toute l'étendue. Réconcilier les différents profils d'un utilisateur à travers ses réseaux sociaux permet de construire un espace d'informations global pour l'utilisateur pour mieux gérer et protéger ses données. Le principal problème réside dans le fait que ces informations sont souvent incomplètes, non mises à jour voire fausses. L'objectif de cet article est de présenter une approche de réconciliation qui exploite à la fois la topologie des réseaux interconnectés et un ensemble de règles portant sur les différentes informations publiques fournies par les utilisateurs. L'évaluation des premières expérimentations menées sur une collection d'environ 2 millions d'utilisateurs issus de Flickr et LiveJournal a permis d'identifier les règles pertinentes et d'obtenir des résultats très prometteurs.

1 Introduction

Avec l'arrivée du Web 2.0, on assiste à un foisonnement de services de réseautage social, qui mettent l'utilisateur, en tant que créateur de contenu, au centre des préoccupations. Ces services permettent de partager des ressources comme des vidéos (YouTube), des photos (Flickr) ou des ressources annotées (Del.icio.us), d'échanger des informations et de construire des relations personnelles ou professionnelles (Facebook, LinkedIn) ou encore de diffuser des news (Twitter, blogs). Les utilisateurs disposent ainsi de plusieurs espaces d'informations sur différents réseaux sociaux, où ils partagent des informations personnelles telles que leurs noms et prénoms, leur lieu géographique, leur âge, leurs adresses électroniques, leurs numéros de téléphone, leurs relations, les institutions fréquentées, etc. (Gross et Acquisti (2005); Little et al. (2011); Stutzman (2006)). Selon la politique d'accès aux données définie par le réseau

social, certaines informations sont par construction publiques ou semi-publiques (accessibles aux autres utilisateurs du réseaux social), pour d'autres c'est l'utilisateur qui choisit de restreindre l'accès à ses contacts ou d'ouvrir l'accès à tous.

Au vu de la diversité des réseaux sociaux, ces espaces permettent l'accès à une grande quantité d'informations en puisant dans des ressources très variées, complémentaires et parfois insoupçonnées, qui pourraient surprendre l'utilisateur lui-même s'il en connaissait réellement toute l'étendue. Réconcilier les différentes profils d'un utilisateur à travers ses divers réseaux sociaux lui permet de construire un espace d'informations global afin de mieux gérer et protéger ses données. Le principal problème réside dans le fait que ces informations sont souvent incomplètes, ambiguës, non mises à jour voire fausses. Il existe déjà, sur le Web, des outils agrégateurs comme *FriendFeed* [1] or *Plaxo* [2] et *Spokeo* [3], qui brassent un large éventail d'informations concernant une personne et fournissent des liens vers ses profils ou ses ressources Web classées comme les vidéos, les photos, les blogs ou les tweets. Cependant, ces informations sont souvent mêlées avec celles d'autres personnes, comme des personnes portant le même nom.

Dans cet article, nous proposons une approche de réconciliation de profils des utilisateurs à travers ses différents réseaux sociaux en exploitant la topologie de l'ensemble des réseaux sociaux interconnectés et les informations publiques fournies par les utilisateurs dans leurs profils. Notre approche repose sur l'observation que, d'une part, les réseaux sociaux sont interconnectés grâce aux liens transversaux que certains utilisateurs déclarent sur leurs différents profils (Golbeck et Rothstein (2008)). D'autre part, les utilisateurs fournissent un ensemble d'informations publiques et accessibles dans les 12 réseaux sociaux les plus répandus (Krishnamurthy et Wills (2009)). Les principales contributions de notre approche sont :

- Notre approche exploite la topologie de l'ensemble des réseaux sociaux interconnectés, afin de sélectionner un sous-ensemble de paires de profils candidats à réconcilier. Le point crucial sur lequel repose cette sélection est qu'un utilisateur déclare des liens d'amitié avec certains utilisateurs (au moins un) sur plusieurs profils.
- Nous avons défini un ensemble de règles combinant un ensemble d'attributs. Chaque règle exprime la contribution d'un ou plusieurs attributs pour mettre en correspondance deux profils. Plus le nombre d'attributs impliqués par une règle est important, plus la probabilité que les deux profils désignent la même personne est forte. Ces règles sont ré-exécutées de manière itérative afin de propager les paires de profils réconciliés et de découvrir de nouvelles paires.
- Nous avons évalué notre approche sur une collection de données d'environ 2 millions d'utilisateurs et 17 millions de liens, issus notamment de Flickr et LiveJournal les résultats obtenus ont atteint une précision de l'ordre de 94%.

La présentation de l'article est organisée comme suit. Dans la section 2, nous présentons un aperçu sur les travaux de recherche portant sur le problème de réconciliation dans les réseaux sociaux. Nous introduisons, ensuite, dans la section 3, les notations permettant de formaliser le problème de réconciliation des profils. Dans la section 4, nous présentons l'approche de réconciliation que nous avons définie. Nous poursuivons, dans la section 5, par les expérimen-

1. www.friendfeed.com
2. www.plaxo.com
3. www.spokeo.com

tations et les évaluations menées sur une collection de données provenant de différents réseaux sociaux. Enfin, nous concluons et présentons nos perspectives.

2 Aperçu de l'état de l'art

De nombreuses solutions à notre problème sont proposées dans la littérature. Perito et al. (2011); Zafarani et Liu (2009) utilisent seulement le pseudonyme d'un utilisateur pour réconcilier deux profils. Les résultats que nous avons obtenus confirment que le pseudonyme est un attribut pertinent. Cependant, notre approche prend en considération plusieurs attributs. L'utilisation des attributs du profil, tels que le nom, le prénom et le pseudonyme est largement étudiée dans (Carmagnola et Cena (2009); Cortis et al. (2012); Golbeck et Rothstein (2008); Malhotra et al. (2012); Motoyama et Varghese (2009); Raad et al. (2010); Rowe (2009)). Malhotra et al. (2012); Motoyama et Varghese (2009) décrivent toute paire de profils comme une vecteur de scores, où chaque score représente la similarité entre les valeurs d'un attribut. Ils utilisent des techniques d'apprentissage automatique et supervisé pour déterminer si une paire de profils référencent un même utilisateur. Bien que les résultats qu'ils ont obtenus soient prometteurs, ce type d'apprentissage nécessite de construire manuellement une base d'exemples suffisamment représentatifs, pour toute paire de réseaux sociaux. Golbeck et Rothstein (2008); Rowe (2009) se focalisent sur la réconciliation de profils décrits en *Friend of a friend* (FOAF). Ces techniques sont appliquées à un ensemble limité d'attributs, comme les adresses électroniques, qui sont susceptibles d'identifier un individu. Carmagnola et Cena (2009) ont introduit la notion de *facteur d'importance* afin de pondérer la contribution de chaque attribut pour déterminer la similarité de deux profils. Dans notre approche, les attributs sont considérés de la même manière dans les règles. Néanmoins, il y a un ordre d'exécution des règles, de la plus contraignante à la moins contraignante. De plus, notre approche découvre de nouveaux profils par propagation de ceux découverts et réconciliés à l'itération précédente. L'évaluation est basée sur des données réelles provenant de quatre réseaux sociaux, tandis que leur expérience est limitée à de petits systèmes fermés. La différence est importante. En effet, dans les réseaux sociaux ouverts, certains utilisateurs ne renseignent pas leur identité réelle, contrairement aux réseaux fermés, où ils pensent que leur identité n'est pas menacée. Par conséquent, les données que nous considérons sont susceptibles d'être fortement bruitées ou erronées ; ce qui constitue un défi non négligeable dans notre contexte. Cortis et al. (2012); Raad et al. (2010) ont proposé le calcul d'une similarité sémantique entre les attributs des profils ; bien que ces approches soient originales, elles ne sont évaluées que sur des petits ensembles de données (par ex., 50 profils (Raad et al. (2010))).

Bartunov et al. (2012); Buccafurri et al. (2012); Jain et al. (2013); Narayanan et Shmatikov (2009) ont étudié les propriétés topologiques des réseaux. Buccafurri et al. (2012) adoptent un raisonnement récursif et considèrent que deux profils sont similaires et donc susceptibles de référencer un même utilisateur, s'ils ont des pseudonymes similaires et les utilisateurs auxquels ils sont liés sont similaires. Cette approche a deux inconvénients que notre approche élimine. Le premier est que les profils ayant des pseudonymes différents sont ignorés, même s'ils pourraient identifier un même individu. Le second est que les profils réconciliés découverts ne sont pas propagés pour découvrir des nouveaux. Bartunov et al. (2012) proposent une approche combinant les attributs d'un profil et la topologie du réseau en utilisant les champs aléatoires conditionnels. Cette approche est robuste car elle peut être aussi utilisée quand les valeurs des

attributs ne sont pas disponibles, ce qui est le cas des réseaux anonymisés. Cependant, comme elle s'appuie sur un modèle probabiliste, les valeurs des paramètres nécessitent l'utilisation de techniques d'apprentissage supervisé et de disposer d'une base d'exemples d'apprentissage.

3 Notations et Formalisation du Problème

Dans notre approche, nous représentons un ensemble de réseaux sociaux interconnectés par un graphe orienté étiqueté, où les nœuds représentent les profils des utilisateurs, et les arcs représentent les liens existant entre eux. Chaque profil possède une uri identifiant sa page sur le Web et un ensemble d'attributs décrits sur cette page. Chaque arc possède une étiquette décrivant le type de lien.[4] Nous considérons deux types de liens, les liens d'amitié entre les utilisateurs d'un même réseau social et les liens transversaux reliant deux profils d'un même utilisateur appartenant à deux réseaux différents. Plus formellement, un ensemble de n réseaux sociaux est un graphe défini comme suit :

$\mathcal{G} =< \bigcup_{i=1}^{n} V^i, \bigcup_{i=1}^{n} E^i \bigcup_{1,i\neq j}^{n} E^{i,j} >$, où :

- V^i est l'ensemble de nœuds, chaque nœud, noté v^i, représente le profils d'un utilisateur sur un réseau social i, avec $\forall i, j,\ i \neq j,\ V^i \cap V^j = \phi$. A représente l'ensemble des attributs définis dans un profil, $P_a(v)$ représente la(les) valeur(s) associée(s) à l'attribut $a \in A$ dans le profil v.
- E^i est l'ensemble des arcs d'étiquette $friend$, chaque arc, noté $(v_1^i, friend, v_2^i)$, représente le lien d'amitié entre un utilisateur de profil v_1^i vers un utilisateur de profil v_2^i, sur le réseau social i. Dans les réseaux sociaux Twitter et YouTube, le lien $friend$ est orienté, alors que, dans Facebook et LinkedIn, ce lien est symétrique par construction puisque les utilisateurs doivent s'accepter mutuellement. La représentation de ce lien tient compte de sa sémantique telle qu'elle est définie dans le réseau social.
- $E^{i,j}$ est l'ensemble d'arcs d'étiquette me, chaque arc, noté (v^i, me, v^j), représente un lien transversal entre les profils v^i et v^j d'un même utilisateur, appartenant, respectivement, aux réseaux sociaux i et j, avec $i \neq j$. Par définition, ce type de lien est symétrique et transitif. Par exemple, Bob déclare dans la page de son profil Flickr, représenté par le nœud v^f, l'uri de la page de son profil LiveJournal, représenté par le nœud v^l, et sur cette page il déclare l'uri de la page de son profil Twitter, représenté par le nœud v^t. Dans ce cas, $E^{f,l} = \{(v^f, me, v^l), (v^l, me, v^f)\}$, $E^{t,l} = \{(v^t, me, v^l), (v^l, me, v^t)\}$ et $E^{t,f} = \{(v^t, me, v^f), (v^f, me, v^t)\}$.

Le problème de réconciliation des profils d'un même utilisateur à travers différents réseaux sociaux peut être réduit au problème de détermination des liens transversaux me manquants noté $mirror$ et formalisé comme suit :

Entrée : $\mathcal{G} =< \bigcup_{i=1}^{n} V^i, \bigcup_{i=1}^{n} E^i \bigcup_{1,i\neq j}^{n} E^{i,j} >$

Sortie : $\{(v^i, mirror, v^j) / v^i \in V^i, v^j \in V^j, 1 \leq i \neq j \leq n, (v^i, me, v^j) \notin E^{i,j}\}$

4. Par abus de langage, un nœud et un arc sont confondus avec le profil et le lien qu'ils représentent, resp.

4 Notre approche de réconciliation des profils

Une solution intuitive au problème posé consiste à comparer chaque paire de profils (v^i, v^j), non reliée par le lien me, pour chaque paire de réseaux (i, j), avec $1 \leq i \neq j \leq n$, c'est-à-dire $\sum_{i,j} |V^i| \times |V^j| - |E^{i,j}|$ paires à comparer. Au vu de la nature combinatoire du problème, nous avons défini une approche de réconciliation de profils qui procède en deux étapes :

- La première étape consiste à sélectionner un sous-ensemble de paires de profils candidats à la réconciliation. Cette étape exploite la topologie du graphe, plus précisément le lien d'amitié $friend$ reliant deux profils dans un même réseau social et le lien transversal me connectant les profils d'un même utilisateur sur différents réseaux sociaux.
- La seconde étape consiste à déterminer parmi ces candidats les paires de profils qui correspondent au même utilisateur. Cette étape exploite les valeurs des attributs de chaque paire de profils et est basée sur un ensemble de règles permettant de découvrir les liens $mirror$ et de les propager. Les deux étapes sont réitérées tant que de nouveaux liens $mirror$ sont découverts.

Nous détaillons, dans ce qui suit, les deux étapes.

4.1 Sélection des paires de profils candidats

A partir de l'analyse des différents profils d'un même utilisateur sur ses réseaux sociaux, nous avons constaté qu'un utilisateur a souvent tendance à déclarer en tant qu'ami au moins un même utilisateur dans ses différents profils. De plus, deux utilisateurs ayant différents profils déclarent qu'ils sont amis sur plusieurs d'entre eux (Golbeck et Rothstein (2008)). Notre hypothèse de base de sélection des paires de profils candidats repose sur ce constat. Par ailleurs, dans un réseau social, il existe des utilisateurs qui déclarent explicitement les liens entre leurs différents profils, ce qui permet de déterminer les amis communs déclarés dans deux profils appartenant à deux réseaux différents.

En d'autres termes, si $\exists(v^i, me, v^j) \in E^{i,j}$, $vr^i \in friend(v^i), vr^j \in friend(v^j)$ alors (vr^i, vr^j) est une paire de profils candidats, où :
$friend(v^i) = \{vr^i/(vr^i, friend, v^i) \vee (v^i, friend, vr^i) \in E^i\}$ représente l'ensemble des profils des amis d'un utilisateur de profil v^i.

L'ensemble des paires de profils candidats, pour les réseaux sociaux i et j, est formellement défini comme suit : $\mathcal{S} = \{(vr^i, vr^j)/\exists vr^i \in friend(v^i) \wedge vr^j \in friend(v^j) \wedge (v^j, me, v^i) \in E^{i,j} \wedge (vr^i, me, vr^j) \notin E^{i,j}\}$

4.2 Détermination des paires de profils *mirror*

Une fois l'ensemble des paires de profils candidats, $\mathcal{S}$, construit, il s'agit de déterminer parmi ces candidats les profils qui référencent un même utilisateur en exploitant les valeurs des attributs définis dans ces profils. Dans ce qui suit, nous présentons les attributs considérés par notre approche. Nous détaillons, ensuite, les règles permettant de déterminer les paires de profils $mirror$ ainsi que l'algorithme de réconciliation que nous avons définis.

4.2.1 Les attributs

Dans la majorité des réseaux sociaux, certains attributs sont rendus publics soit par défaut soit c'est l'utilisateur qui choisit de les publier. Krishnamurthy et Wills (2009) ont identifié un ensemble d'attributs souvent publics dans les 12 plus répandus réseaux sociaux. Notre approche de réconciliation vise à exploiter un grand nombre de réseaux sociaux, nous avons donc besoin d'isoler un ensemble d'attributs accessibles, publics et permettant de déterminer si deux profils référencent le même utilisateur. Dans cet article, nous nous focalisons sur les attributs suivants : le pseudonyme, les noms et prénoms, les adresses électroniques et les liens vers des pages Web.

L'attribut pseudonyme, noté p, est un attribut public dont la valeur est toujours accessible. C'est l'identifiant du profil sur un réseau social donné. Il fait généralement partie de l'uri de la page du profil sur le Web. Des études comme dans (Buccafurri et al. (2012); Perito et al. (2011); Zafarani et Liu (2009)) ont montré que les utilisateurs des réseaux sociaux ont tendance à utiliser des pseudonymes identiques ou comportant des sous-chaînes identiques sur leurs différents profils.

Pour évaluer la similarité de deux pseudonymes, nous avons choisi d'utiliser la distance de Levenshtein. En effet, cette distance permet de capturer les variations des sous-chaînes de caractères en calculant le nombre suppressions ou d'insertions de caractères d'un mot p_1 pour qu'il soit identique à un mot p_2. Prenons l'exemple de deux profils référençant le même utilisateur dont le pseudonyme sur la page `www.flickr.com/photos/cospics` est *"cospics"* et le pseudonyme sur la page `www.livejournal.com/users/cos/profile` est *"cos"*, le nombre de suppressions est de 4. Le seuil de similarité est défini dans les expérimentations.

L'attribut nom, noté n, est un attribut dont la valeur représente les noms et/ou prénoms renseignés par l'utilisateur. En effet, ces valeurs ne correspondent généralement pas à des champs distincts et bien identifiés dans un réseau social et ne sont pas renseignés avec le même niveau de détail d'un réseau social à un autre. Le nom est un attribut ambigu, en particulier, lorsque les valeurs correspondent à des noms et prénoms communs. Il est également sensible dans le sens où l'utilisateur préfère rester anonyme, même s'il renseigne cette information, souvent il ne révèle pas son véritable nom. Il est donc clair que l'utilisation seule de cet attribut est insuffisante pour réconcilier deux profils.

Pour mesurer la similarité de deux valeurs de l'attribut n, nous utilisons la mesure de Jaccard qui permet de calculer le nombre de mots communs sans prendre en compte leur ordre d'occurrence dans la chaîne de caractères. Par exemple, la mesure de Jaccard pour les chaînes "Barack, Obama" et "Obama Barack" est égale à 1 et à $\frac{2}{3}$ pour les chaînes "Barack, Hussein, Obama" et "Obama Barack" .

L'attribut adresse électronique, noté m, est un attribut multivalué dont les valeurs correspondent aux différentes adresses électroniques d'un utilisateur. Si renseignée, il s'agit d'un attribut qui permet d'identifier de manière unique un utilisateur. Cependant, cette information n'est pas souvent renseigné, et l'utilisation seule de cet attribut est insuffisante pour réconcilier deux profils.

Pour comparer les valeurs de l'attribut m de deux profils, il s'agit de déterminer si l'une des adresses électroniques d'un profil est identique à l'une des adresses électroniques de l'autre profil. Dans le cas positif la similarité est égale 1. Sinon, elle est égale à 0.

L'attribut liens vers d'autres pages Web est un attribut multivalué dont les différentes valeurs correspondent à différentes url référençant des liens vers des pages Web. Nous distinguons deux types de liens, ceux qui référencent des liens vers des profils de réseaux sociaux, noté s, de ceux qui référencent des liens vers d'autres pages, noté w. Le but étant de pouvoir analyser séparément la contribution de chacun des types. En effet, les liens vers les réseaux sociaux, figurant sur le profil d'un utilisateur, pourraient être des liens vers ses autres profils. Ce qui pourrait correspondre à un lien transversal. Par ailleurs, les liens vers d'autres pages Web sont des liens vers des ressources que l'utilisateur souhaite partager avec les autres et qui pourraient correspondre à ses pages personnelles.

Dans cet article, nous nous limitons dans cette première étude à rechercher si les valeurs de l'attribut w ou s, pour deux profils, ont au moins une url commune sans analyser le contenu de ces pages, en particulier les liens vers les réseaux sociaux, qui pourraient être pertinents pour la réconciliation des profils.

L'attribut lieu, noté l, est un attribut dont la valeur correspond aux différents lieux renseignés par l'utilisateur. Il s'agit d'un attribut public souvent renseigné par les utilisateurs. Néanmoins, cet attribut est souvent ambigu lorsqu'on est amené à comparer les valeurs de deux profils différents, et ce pour plusieurs raisons :

- le lieu peut être incomplet, ce qui ne permet pas de l'identifier de manière unique en tant qu'entité géographique. Même si les deux valeurs sont identiques, par exemple pour la valeur *"Paris"* il existe plusieurs entités géographiques avec ce même label *"Paris"* au *"Texas"* ou *"Paris"* en *"France"*.
- les champs pour renseigner le type de découpage administratif (ville, pays, région ou état, par ex.) ne sont pas distincts et ne sont pas du même niveau de détail d'un réseau social à un autre. Un nom d'état peut, par exemple, correspondre à un nom de ville (e.g., *New York, state* ou *New York, city*).
- la sémantique qu'on pourrait associer à un lieu n'est pas toujours très claire. S'agit-il du lieu d'origine ou de résidence temporaire ou régulière de l'utilisateur. De plus, l'utilisateur peut avoir renseigné son lieu d'origine sur un profil et son lieu actuel sur un autre profil. Sur certains réseaux sociaux, comme LiveJournal deux types de lieux peuvent être renseignés d'origine et actuelle. De plus, vu qu'il s'agit d'une information qui peut évoluer dans le temps, elle n'est pas mis à jour de la même manière sur les différents profils d'un même utilisateur.

Les valeurs de cet attribut nécessitent avant tout d'être désambiguïsées. L'utilisation d'une base de connaissances comme GeoNames permettrait d'associer la sémantique aux entités extraites, leurs labels alternatifs et aussi les relations que ces entités ont entre elles. Exploiter cet attribut ne rentre pas dans le cadre du travail de recherche que nous présentons dans cet article.

4.2.2 Les règles

Pour déterminer si deux profils v^i et v^j référencent un même utilisateur, nous avons défini un ensemble de règles exploitant les attributs présentés ci-dessus à appliquer à chaque paire de $\mathcal{S}$. Chaque règle prend en considération la contribution d'un ou plusieurs attributs. Étant donné qu'aucun attribut ne constitue l'identité de l'utilisateur, nous supposons que plus le nombre d'attributs qui correspondent, selon la mesure de similarité définie, est grand pour deux profils plus la probabilité qu'ils référencent un même utilisateur est forte. Ainsi, les règles définies sont classées par ordre de pertinence noté k. La règle la plus pertinente, d'ordre maximale,

est celle qui détermine que tous les attributs correspondent ; dans ce cas $k = |A|$. La règle la moins pertinente est celle qui détermine que un seul attribut correspond ; dans ce cas $k = 1$.

Soit le prédicat noté $match(P_a(v^i), P_a(v^j))$ qui retourne vrai si les valeurs de l'attribut a pour les profils v^i et v^j correspondent selon la mesure de similarité définie pour l'attribut a. Une règle d'ordre k, noté $\mathcal{R}^k$ est définie comme suit :

$$\mathcal{R}^k(v^i, v^j) = \begin{cases} \bigwedge_{a \in A} match(P_a(v^i), P_a(v^j)) & \text{si } k = |A| \\ \bigvee_{B \in [A]^k} \bigwedge_{a \in A \setminus B} match(P_a(v^i), P_a(v^j)) & \text{si } 1 \leq k < |A| \end{cases}$$

où $[A]^k$ représente l'ensemble des tous les sous-ensemble de A de taille k.

Ainsi, si $\bigvee_{1 \leq k \leq |A|} \mathcal{R}^k(v^i, v^j)$ est vrai, v^i et v^j sont deux profils référençant le même utilisateur. Si pour un couple de profils candidats (v^i, v^j), au moins une règle $\mathcal{R}^k(v^i, v^j)$ retourne $vrai$, alors aucune règle d'ordre $k - 1$ n'est appliquée. Au pire, pour (v^i, v^j) aucune règle d'ordre $\mathcal{R}^1(v^i, v^j)$ ne retourne vrai. Dans le cas où il existe une règle $\mathcal{R}^k(v^i, v^j)$ vraie, les nœuds v^i et v^j sont reliés par l'arc étiqueté $mirror$.

Ces liens $mirror$ découverts sont exploités et considérés comme des liens me pour sélectionner de nouveaux couples de profils candidats $C\prime$. Ainsi, les règles sont ré-exécutées pour $C\prime$, et ainsi de suite, jusqu'à ce qu'il n y ait plus de nouvelles paires de profils $mirror$.

		Arcs		
Réseau	Nœuds	$friend$	me	Total
Flickr	1 814 405	15 415 083	189	154 152 72
LiveJournal	211 045	2 093 737	161	2 093,898
Twitter	8 842	19 008	312	19 320
YouTube	1 210	1 367	286	1 653
Total	**2 035 502**	**17 529 195**	**474**	**17 529 669**

TAB. 1 – *Caractéristiques du graphe de réseaux sociaux interconnectés*

Réseau	Flickr	LiveJournal	Twitter	YouTube
Flickr	–	148	29	12
LiveJournal	148	–	11	2
Twitter	29	11	–	272
YouTube	12	2	272	–

TAB. 2 – *Liens transversaux entre chaque paire de réseaux sociaux*

5 Expérimentations et évaluations des résultats

Pour évaluer notre approche, nous avons utilisé une collection de données provenant de 4 réseaux sociaux LiveJournal, Flickr, Twitter et YouTube [5].

5. http://www.ursino.unirc.it/pkdd-12.html

Le graphe ainsi construit est composé de 93 177 nœuds et 146 075 liens, dont 462 sont des liens transversaux. Les données provenant de Flickr et LiveJournal sont de loin plus nombreuses, avec 55 117 (83 993) et 28 008 (41 244) nœuds (liens) respectivement ; celles provenant de Twitter et YouTube comportent seulement 8842 (19 008) et 1210 (1367) nœuds (liens) respectivement. Cependant l'analyse du réseau a révélé un grand nombre de liens manquants. De plus, le pseudonyme est la seule information disponible dans cette collection. Nous avons donc procédé à l'enrichissement des deux réseaux les plus importants Flickr et LiveJournal pour chaque nœud existant, on extrait en utilisant l'API appropriée du réseau tous les nœuds voisins par le lien $friend$. Ces nœuds sont rajoutés au graphe ainsi que les liens qu'ils ont avec les autres nœuds existants. Après enrichissement, nous avons obtenu, au total, plus de 2 millions de nœuds et plus de 17 millions de liens. Les caractéristiques du graphe des réseaux sociaux interconnectés sont détaillées dans la table 1. Le nombre de liens me est passé de 462 à 474 suite à la fermeture transitive impliquant les nouveaux nœuds ajoutés. La table 2 détaille le nombre de liens transversaux pour chaque paire de réseaux sociaux. Pour l'implémentation de notre approche, nous avons utilisé Neo4j [6], une base de données puissante dédiée pour représenter et manipuler des graphes de grande taille.

Pour réaliser une première évaluation, nous avons appliqué notre approche de réconciliation aux 2 réseaux sociaux Flickr et LiveJournal. Tout d'abord, les paires de profils candidats sont obtenues en exploitant les liens transversaux comme expliqué dans la section 4.1. L'ensemble des paires de profils $mirror$, noté $\mathcal{M}$, est ensuite généré en appliquant les règles présentées dans la section 4.2.2.

Les premiers résultats que nous avons obtenus, à l'issue de la 1ère itération, pour un seuil fixé arbitrairement $\theta = 0.7$, et pour chacune des règles vérifiée, sont décrits dans la table 3. Au total 3424 couples vérifient les règles sur environ 16000 couples candidats dont 0.03% vérifie la règle d'ordre 4, 0.91% vérifie les règles d'ordre 3, 7.65% vérifie les règles d'ordre 2 et 91.41% pour les règles d'ordre 1. Comme l'attribut adresse électronique n'est pas renseigné, dans la collection considérée, pour au moins un profil d'un couple candidat, l'attribut m n'intervient dans aucune règle. Ces résultats montrent qu'un faible pourcentage de paires de profils vérifient les règles d'ordre $k \geq 2$, et le plus grand pourcentage vérifie la règle d'ordre 1. Ce qui confirme le fait que les utilisateurs préfèrent généralement ne pas divulguer trop d'informations personnelles. Les résultats montrent aussi que l'attribut pseudonyme est l'attribut intervenant dans les règles vérifiées par le plus grand nombre de couples de profils candidats.

Pour l'évaluation des résultats, nous avons défini trois sous-ensembles de l'ensemble de couples $mirror$ $\mathcal{M}$: (i) corrects référençant le même utilisateur, noté C ; (ii) faux référençant deux utilisateurs différents, noté F et (iii) indéterminés pour lesquels nous ne disposons pas d'informations suffisantes pour décider. La précision est donc calculée en utilisant la formule suivante : précision= $\frac{|C|}{|\mathcal{M}|}$. Il faut noter que nous avons choisi dans ce calcul de tenir compte de l'ensemble des couples indéterminés pour évaluer au mieux notre approche. L'évaluation est faite manuellement et ce sur la totalité des 3424 couples vérifiant les règles en utilisant les uri des pages des profils.

Les résultats des règles d'ordre 1 montrent que les attributs générant le plus grand nombre de paires de profils $mirror$ faux sont le nom (54.55%) et ensuite le pseudonyme (30.19%). Par ailleurs, avec un nombre de paires de profils $mirror$ plus faible (2.92% pour l'attribut s et 2.19% pour l'attribut w), les attributs liens vers d'autres pages Web génèrent le plus grand

6. www.neo4j.org/

règle	attributs	proportion de $\lvert\mathcal{M}\rvert\%$	$\frac{\lvert C\rvert}{\lvert\mathcal{M}\rvert}$ %	$\frac{\lvert F\rvert}{\lvert\mathcal{M}\rvert}\%$	$\frac{\lvert I\rvert}{\lvert\mathcal{M}\rvert}\%$
$\mathcal{R}^1$	$\{p\}$	83.09	60.39	30.19	9.42
	$\{n\}$	3.21	42.72	54.55	2.73
	$\{s\}$	2.92	100.00	0.00	0.00
	$\{w\}$	2.19	96.00	2.00	2.00
	Total	91.41	**61.85**	**29.42**	**8.73**
$\mathcal{R}^2$	$\{p, s\}$	3.21	100	0	0
	$\{p, w\}$	2.07	99.09	0	0.91
	$\{p, n\}$	1.93	92.42	1.52	6.06
	$\{n, s\}$	0.26	100	0	0
	$\{n, w\}$	0.18	100	0	0
	Total	7.65	**97.71**	**0.38**	**1.91**
$\mathcal{R}^3$	$\{p, n, s\}$	0.55	100	0	0
	$\{p, w, n\}$	0.32	100	0	0
	$\{p, w, s\}$	0.03	100	0	0
	Total	0.91	**100**	**0**	**0**
$\mathcal{R}^4$	$\{p, n, w, s\}$	0.03	100	0	0
	Total	0.03	**100**	**0**	**0**
	Grand total	100	**64.95**	**26.93**	**8.12**

TAB. 3 – *Résultats de la 1ère itération, pour chaque règle, Flickr et LiveJournal,* $\theta = 0.7$

nombre de couples corrects (100% pour l'attribut s et 96% pour l'attribut w), ce qui confirme notre intuition sur le fait que l'attribut lien vers d'autres profils est un attribut pertinent. Nous observons que les règles combinant au moins deux attributs, $k \geq 2$, y compris celles utilisant les attributs nom et pseudonyme, atteignent une précision entre 92% et 100%. Ce qui confirme notre hypothèse initiale, plus le nombre d'attributs qui correspondent est grand plus les profils correspondent.

Ces résultats soulignent qu'une similarité forte de deux noms n'est pas à elle seule un élément pertinent. En effet, l'attribut nom est un attribut non seulement ambigu mais aussi sensible, souvent l'utilisateur fournit une information erronée de manière délibérée afin de ne pas dévoiler son identité. Par exemple, en analysant précisément les pages de profils appartenant au même utilisateur, souvent les noms sont partiellement ou complètement différents.

Pour mieux comprendre les résultats obtenus pour les attributs nom et pseudonyme, nous avons procédé à d'autres tests en utilisant un seuil de similarité plus grand et nous avons constaté que la précision augmente significativement pour les règles d'ordre 1 utilisant le pseudonyme, mais l'impact reste non significatif pour les règles utilisant le nom. Comme expliqué dans la section 4.2.1, dans notre approche, deux pseudonymes très similaires diffèrent très peu en termes de nombre et de séquencement de caractères ; et deux noms sont similaires s'ils possèdent un grand nombre de mots en communs.

Afin de propager la découverte des couples $mirror$ pour les réseaux Flickr et LiveJournal sans dégrader la précision, nous avons supprimé la règle d'ordre 1 portant sur l'attribut nom

n et nous avons fixé le seuil de similarité θ à 0.9. Nous avons obtenu 2768 paires de profils au bout de 4 itérations (1053 à la 1ère, 1005 à la 2ème, 654 à la 3ème, 56 à la 4ème itération). La précision est passée à 94% avec 2% de faux et le reste indéterminés.

Nous avons appliqué notre algorithme de réconciliation à d'autres paires de réseaux sociaux en utilisant Twitter et YouTube. Le nombre de paires de profils découverts est très faible en raison du peu d'informations disponibles dans ces réseaux sociaux. En effet, nous disposons seulement de 8842 nœuds dans Twitter et 1210 nœuds dans YouTube pour un nombre de liens transversaux plus élevé entre les deux réseaux sociaux (voir les tables 2 et 1).

6 Conclusions et perspectives

Dans cet article, nous avons présenté notre approche de réconciliation de profils dans les réseaux sociaux qui exploite la topologie du graphe et les attributs publics définis dans les profils et accessibles dans de nombreux réseaux sociaux. Les résultats obtenus sur la collection de données issues des réseaux LiveJournal et Flickr ont montré la pertinence des attributs considérés et l'efficacité des règles que nous avons définies. La précision a atteint 94% et le nombre de liens découverts est passé de 148 liens transversaux me initialement renseignés entre LiveJournal et Flickr à 2768 au bout de 4 itérations. De plus, cette précision peut être contrôlée en appliquant les règles dont l'ordre est supérieure à une valeur k fixée, contraignant ainsi la similarité d'au moins k attributs. Elle peut également être contrôlée en triant, de manière croissante, les paires de profils $mirror$ découverts (v^i, v^j), pour chaque nœud v^i, selon l'ordre de la règle et le nombre de liens $friend$ communs entre (v^i, v^j). Le nombre de liens découverts sur les paires de réseaux sociaux Twitter et YouTube comportant peu de nœuds et peu d'informations renseignées sur les profils est faible. Dans ce cas, d'autres éléments peuvent être prises en compte : (i) la topologie du graphe peut être davantage exploitée en distinguant la nature des liens $friend$ entrant ou sortant et en considérant leur nombre ; (ii) analyser le contenu ou les tags associés aux ressources sont des éléments qui pourraient être pertinents pour déterminer la similarité de deux profils ; (iii) exploiter l'attribut lieu en désambiguïsant les valeurs associées et en définissant une mesure de similarité entre deux valeurs de lieu définies dans deux profils différents. Ces derniers aspects font l'objet de nos travaux actuels.

Références

Bartunov, S., A. Korshunov, S. Park, W. Ryu, et H. Lee (2012). Joint Link-attribute User Identity Resolution in Online Social Networks. In *SNA-KDD Workshop*.

Buccafurri, F., G. Lax, A. Nocera, et D. Ursino (2012). Discovering Links among Social Networks. In *Machine Learning and Knowledge Discovery in Databases*, Volume 7524 of *Lecture Notes in Computer Science*, pp. 467–482. Springer Berlin Heidelberg.

Carmagnola, F. et F. Cena (2009). User Identification for Cross-system Personalisation. *Inf. Sci. 179*(1-2), 16–32.

Cortis, K., S. Scerri, I. Rivera, et S. Handschuh (2012). Discovering Semantic Equivalence of People Behind Online Profiles. In *In Proceedings of the Resource Discovery (RED) Workshop, ser. ESWC*.

Golbeck, J. et M. Rothstein (2008). Linking Social Networks on the Web with FOAF : A Semantic Web Case Study. In *AAAI*, Volume 8, pp. 1138–1143.

Gross, R. et A. Acquisti (2005). Information Revelation and Privacy in Online Social Networks. In *Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society*, WPES '05, New York, NY, USA, pp. 71–80. ACM.

Jain, P., P. Kumaraguru, et A. Joshi (2013). @i Seek 'fb.me' : Identifying Users Across Multiple Online Social Networks. In *WWW (Companion Volume)*, pp. 1259–1268.

Krishnamurthy, B. et C. E. Wills (2009). On the Leakage of Personally Identifiable Information via Online Social Networks. In *Proceedings of the 2nd ACM Workshop on Online Social Networks*, pp. 7–12. ACM.

Little, L., P. Briggs, et L. Coventry (2011). Who Knows about Me ? : An Analysis of Age-related Disclosure Preferences. In *Proceedings of the 25th BCS Conference on Human-Computer Interaction*, BCS-HCI '11, Swinton, UK, UK, pp. 84–87. British Computer Society.

Malhotra, A., L. Totti, W. Meira, P. Kumaraguru, et V. Almeida (2012). Studying User Footprints in Different Online Social Networks. In *International Workshop on Cybersecurity of Online Social Network (ACM ASONAM 2012)*.

Motoyama, M. et G. Varghese (2009). I Seek You : Searching and Matching Individuals in Social Networks. In *Proceedings of the Eleventh International Workshop on Web Information and Data Management*, pp. 67–75. ACM.

Narayanan, A. et V. Shmatikov (2009). De-anonymizing Social Networks. In *30th IEEE Symposium on Security and Privacy*, pp. 173–187. IEEE.

Perito, D., C. Castelluccia, M. A. Kaafar, et P. Manils (2011). How Unique and Traceable are Usernames ? In *Privacy Enhancing Technologies*, pp. 1–17. Springer.

Raad, E., R. Chbeir, et A. Dipanda (2010). User Profile Matching in Social Networks. In *Network-Based Information Systems (NBiS), 2010 13th International Conference on*, pp. 297–304. IEEE.

Rowe, M. (2009). Interlinking Distributed Social Graphs. In *Linked Data on the Web Workshop, WWW2009*.

Stutzman, F. (2006). An Evaluation of Identity-Sharing Behavior in Social Network Communities. *iDMAa Journal 3*(1).

Zafarani, R. et H. Liu (2009). Connecting Corresponding Identities across Communities. In *Third International AAAI Conference on Weblogs and Social Media*.

Summary

It is not uncommon that individuals create multiple profiles across several SNSs, each containing partially overlapping sets of personal information. As a result, the creation of a global profile that gives an holistic view of the information of an individual requires methods that automatically match, or *reconciliates*, profiles across SNSs. In this paper, we focus on the problem of identifying, or matching, the profiles of any individual across social networks.