Dérivation de profils utilisateurs à partir de réseaux sociaux : une approche par communautés de réseaux égocentriques

Dieudonné Tchuente, Marie-Françoise Canut, Nadine Jessel, André Péninou, Florence Sèdes

Dérivation de profils utilisateurs à partir des réseaux sociaux : une approche par communautés de réseaux égocentriques

Dieudonné Tchuente, Marie-Françoise Canut, Nadine Baptiste Jessel, André Péninou, Florence Sèdes
Université de Toulouse
Institut de Recherche en Informatique de Toulouse (IRIT)
Equipe SIG (Systèmes d'Information Généralisés)
118, Route de narbonne, 31062, Toulouse, Cedex
tchuente@irit.fr, marie-francoise.canut@irit.fr, baptiste@irit.fr, andre.peninou@irit.fr, sedes@irit.fr

RÉSUMÉ. *Lorsqu'un utilisateur est peu ou pas actif dans un système d'information, son profil est peu/mal connu et les mécanismes qui s'y réfèrent (personnalisation de contenus par exemple) sont moins efficaces. Pour améliorer les résultats de ces mécanismes, plusieurs approches récentes montrent la pertinence de l'usage du réseau social de l'utilisateur comme source d'information complémentaire au profil de l'utilisateur. Ces approches présentent cependant deux inconvénients majeurs : d'une part, chacune des approches reste très spécifique à son domaine d'application (et au mécanisme associé), et d'autre part, elles exploitent de manière unilatérale les profils des individus autour de l'utilisateur pour améliorer ces mécanismes. Pour pallier ces inconvénients, nous proposons d'une part, un modèle générique et social de profil utilisateur susceptible d'être réutilisé par plusieurs mécanismes dans différents domaines d'application, et d'autre part, un processus de dérivation des éléments du profil de l'utilisateur non pas à partir des individus autour de l'utilisateur, mais à partir des communautés autour de l'utilisateur dans une portion de son réseau social (réseau égocentrique). Une évaluation préliminaire menée dans Facebook montre l'intérêt de cette nouvelle approche.*

RÉSUMÉ. *When a user is not or slightly active in an information system, his profile is incomplete and even unknown. Thus, associated mechanisms (e.g. personalization) are less reliable. To address this problem, several studies show that using the user's profile combined with data derived from the user's social network gives more reliable results. However, these studies have two major drawbacks: each study is specific in its own field (and associated mechanisms) and all these studies unilaterally use profiles of individuals around the user in the social network. To overcome these drawbacks, we aim at defining a generic social model of users' profiles that can be reuse in many application domains and for several mechanisms. Moreover, we propose a new way to enrich the user profile by communities (rather than individuals) around the user in a part of his social network (egocentric network). A preliminary evaluation of this new approach in Facebook shows his relevance.*

MOTS CLES : PROFIL UTILISATEUR, RESEAUX SOCIAUX, RESEAUX EGOCENTRIQUES, COMMUNAUTES, FACEBOOK
KEYWORDS : USER PROFILE, SOCIAL NETWORKS, EGOCENTRIC NETWORKS, COMMUNITIES, FACEBOOK

1. Introduction

La construction des profils utilisateurs se situe au centre des problématiques identifiées lors de la mise en œuvre de mécanismes d'adaptation, de recommandation ou de personnalisation de l'information aux utilisateurs afin de prendre en compte leurs besoins spécifiques. Les performances de ces mécanismes sont très dépendantes de la qualité du profil construit pour chaque utilisateur. Ces profils sont généralement construits et enrichis de manière itérative au fur et à mesure que l'utilisateur interagit avec le système d'information. De par la nature incrémentale de ce processus, un profil utilisateur n'est jamais complètement connu à un instant donné. Il en découle deux problématiques principales :

- Premièrement, le profil d'un nouvel utilisateur du système d'information, et de ce fait, aucun mécanisme ne peut modifier son comportement en fonction de son besoin. Ce problème est connu dans la littérature comme celui du démarrage à froid ou *« cold start problem»* (Massa et al*.,* 07).
- Deuxièmement, à tout instant, le profil de l'utilisateur est a priori incomplet car toutes les informations sur l'utilisateur peuvent ne pas être connues. Ainsi, on peut à un instant t avoir des informations sur le profil de l'utilisateur dans le domaine du sport, mais pas dans le domaine de la culture. Un mécanisme qui utilise ce profil (personnalisation par exemple) ne pourra donc pas personnaliser à cet instant t l'information à renvoyer à l'utilisateur dans le domaine de la culture. Ceci pose le deuxième problème de l'enrichissement d'un profil utilisateur et de la recherche de sources d'informations sur le profil de l'utilisateur.

Pour résoudre ces deux problèmes, le profil de l'utilisateur peut être enrichi à partir d'autres sources de données que l'utilisateur lui-même. Il peut s'agir :

a) de sources de données produites par le même utilisateur dans d'autres systèmes d'information (Viviani et al., 10),

b) de sources de données produites par d'autres utilisateurs : en général les utilisateurs qui lui sont similaires (Su et al*.,* 09) ou les utilisateurs de son réseau social (Kautz et al., 97)(Carmel et al.,09).

La première approche (a) est restreinte à un cadre de coopération entre plusieurs systèmes d'information. La seconde approche (b) qui nous intéresse dans cet article dépasse cette restriction, mais peut avoir des inconvénients dans plusieurs cas :

- pour le cas de l'usage d'utilisateurs au profil similaire, il faudrait que l'utilisateur lui-même dispose déjà d'un profil non vide, afin de pouvoir être comparé à ceux d'autres utilisateurs. Donc, le problème de démarrage à froid *« cold start problem»* ne peut être résolu que très difficilement dans ce cas. De même, un profil utilisateur avec peu d'informations rend les comparaisons difficiles,

- pour le cas des utilisateurs du réseau social, les travaux qui les utilisent actuellement sont très dépendants de la nature du lien (confiance, amis, co-auteur, etc.) (Massa et al., 07)(Carmel et al., 09)(Cabanac, 11) défini pour générer le réseau social d'une part, et d'autre part du mécanisme d'usage des profils construits (personnalisation, recommandation, etc.). En d'autres termes, les travaux proposés dépendent fortement de la nature du réseau social (réseau de confiance, réseau de co-auteurs, réseau d'amis, etc.), et ces travaux sont difficilement réutilisables car elles n'établissent pas une séparation claire entre le *profil social* de l'utilisateur (profil construit à partir du réseau social), et les mécanismes qui utilisent ce profil (personnalisation, recommandation, etc.).

Dans cet article, nous nous intéressons aux réseaux sociaux dont la nature du lien entre utilisateurs correspond à celle utilisée dans les sciences sociales (les individus et leurs relations dans la vie réelle) tels que les réseaux téléphoniques (contacts entre abonnés), réseaux sociaux numériques, réseaux de co-auteurs. Pour être indépendant des mécanismes d'usage des profils, nous proposons un modèle générique de profil utilisateur qui intègre une dimension sociale. Nous intégrons dans cette dimension sociale des communautés extraites d'une partie bien définie du réseau social de l'utilisateur (son réseau égocentrique). Ceci se justifie dans la mesure où des travaux existants en sciences sociales (Gofmann, 59) démontrent que l'on peut caractériser un individu à partir des communautés sociales auxquelles il appartient. A partir de ces communautés extraites du réseau social, nous proposons une technique de dérivation du profil social de l'utilisateur. Une évaluation préliminaire de cette approche dans Facebook démontre son intérêt relativement aux approches existantes.

La suite de cet article est structurée comme suit. La section 2 propose une revue de la littérature sur les travaux connexes à notre approche. Dans la section 3 nous présentons le modèle « social » du profil utilisateur. La section 4 décrit d'une part le processus de construction du profil utilisateur à partir du modèle proposé précédemment, et d'autre part, l'approche par communautés permettant de dériver le profil social de l'utilisateur à partir de ce modèle. La section 5 présente l'évaluation préliminaire réalisée sur Facebook pour montrer l'intérêt de l'approche proposée par rapport aux approches existantes. Enfin, la section 6 présente les conclusions et perspectives de ce travail.

2. Revue de la littérature

Nos travaux se situent à la frontière de deux axes de recherche : la modélisation de profils utilisateurs et l'analyse des réseaux sociaux. Nos objectifs tentent de répondre à la question suivante : comment analyser efficacement les réseaux sociaux pour dériver des profils utilisateurs, et donc permettre de résoudre les problèmes de démarrage à froid ou d'enrichissement des profils utilisateurs ? Pour ce faire, nous étudions les travaux existants qui s'appuient sur les réseaux sociaux pour enrichir les profils et améliorer les mécanismes d'adaptation, de recommandation ou de personnalisation.

De manière formelle, un réseau social est défini comme un graphe pour lequel les nœuds sont des individus. Les liens entre individus (dont la nature doit être clairement définie) peuvent être orientés ou non, pondérés ou non. On suppose que l'on dispose de propriétés, éventuellement pondérées, attachées à chaque nœud. Ces propriétés caractérisent l'individu, il peut s'agir par exemple des données d'identification, des données démographiques, des centres d'intérêts, ... La question qui se pose ici est alors de pouvoir dériver des propriétés d'un nœud en fonction des propriétés des autres nœuds du graphe. Nous pensons que deux éléments sont fondamentaux pour répondre à cette question :

A) La nature des liens dans le graphe (exemple : amis, co-auteurs, etc.) peut avoir un impact du point de vue de la sémantique des traitements ou du point de vue de la topologie (distribution des degrés[1] des nœuds) du réseau obtenu (exemple : graphes « à effet petits mondes »[2] , graphes « sans échelle »[3]).

B) Le filtre à appliquer pour sélectionner les nœuds à partir desquels on pourra dériver les propriétés sur les individus.

Nous avons comparé des travaux de la littérature qui utilisent les réseaux sociaux pour améliorer les mécanismes de recommandation ou de personnalisation par rapport à ces deux critères (nature des liens, filtre) et par rapport à la résolution des problématiques de démarrage à froid ou d'enrichissement des profils soulevées dans l'introduction (voir tableau 1).

Pour les systèmes de recommandation, on distingue dans la littérature : (i) les systèmes de recommandation classiques qui s'appuient sur le filtrage collaboratif (Su et al., 09) pour lesquels la nature du lien entre les individus est la similarité comportementale. Pour résoudre le problème du démarrage à froid, ces systèmes peuvent être étendus avec des réseaux sociaux comme les réseaux de confiance (Massal et al., 07). (ii) Les systèmes de recommandation sociaux (*Social Recommender Systems*) qui s'appuient uniquement sur des réseaux sociaux, exemple de (Guy et al., 09)(Cabanac, 11). Pour les systèmes de personnalisation, on retrouve principalement dans la littérature les travaux en recherche d'information sociale (Bender et al., 08) (Carmel et al., 09)(Ren et al., 10)(Bouadjenek et al., 11), dans lesquels les résultats d'une requête utilisateur sont personnalisés via son profil et via le profil de son réseau social. Le tableau 1 présente un comparatif de trois exemples représentatifs de techniques exploitées dans les travaux récents. De ce tableau, il ressort deux principales remarques :

1- Nature du lien : les techniques utilisées diffèrent considérablement en fonction de la nature des liens dans le réseau social dans la mesure où les réseaux sociaux

---

1 Le degré d'un nœud est son nombre de voisins.

2 Graphe qui respecte la théorie des six degrés de séparation.

3 Graphe dans lequel la distribution des degrés est très loin d'être uniforme entre les nœuds.

générés à partir des activités des utilisateurs ont des sémantiques différentes (achats de produits similaires, co-écritures d'articles, tags de même ressources, etc.) et où ces activités sont en lien direct avec le domaine visé (achats de produits similaires pour la recommandation de produits à acheter, …).

2- Filtrage des nœuds : il existe peu de techniques avancées pour le filtrage de nœuds à exploiter pour enrichir le profil de l'utilisateur à partir de son réseau social. En général une même importance est accordée à tous les individus directement liés à l'utilisateur (Carmel et al., 09). (Cabanac, 11) propose des techniques plus élaborées dans ce sens, cependant elles restent très spécifiques au domaine de la bibliométrie.

	Système de recommandation+ réseau social Exemple de (Massa et al., 07)	**Système de recommandation social** Exemple de (Cabanac, 11)	**Système Recherche d'information sociale** Exemple de (Carmel et al., 09)
Contexte du travail	E-commerce (recommandation d'achat de produits)	Bibliométrie (recommandation d'auteurs)	Moteur de recherche (personnalisation de requêtes)
Nature des liens (A)	Achat de produits similaires sur un site de e-commerce.	2 graphes : co-auteurs d'articles scientifiques, et participation à une même conférence scientifique.	3 graphes (dans des outils collaboratifs d'un intranet d'entreprise) : similarité, familiarité, similarité et familiarité.
Filtre réalisé pour sélections des nœuds (B)	Nœuds dont la mesure de similarité (vectorielle par exemple) supérieure à un seuil défini.	Nœuds dont les mesures de proximité et de connectivité dans le graphe des co-auteurs, probabilité de rencontre dans le graphe de participation à des conférences.	Voisins directs (distance 1 dans chacun des graphes considérés).
Problème du démarrage à froid	Pas résolu : car calcul de similarité impossible pour un utilisateur n'ayant fait aucun achat.	Pas résolu, car calcul de similarité impossible pour un utilisateur n'ayant pas encore d'article publié.	Résolu, si le graphe entre utilisateurs est indépendant des activités (graphe de familiarité). Toutefois, le filtrage considérant tous les voisins n'est pas forcément optimal…
Problème d'enrichissement du profil	Résolu, mais le filtrage nécessite très souvent de maintenir d'énormes matrices creuses et n'est pas optimal.	Résolu, mais la technique reste très spécifique au domaine de la bibliométrie.	Résolu, mais comme pour le problème du démarrage à froid, le filtrage considérant tous les voisins n'est pas forcément optimal…

Tableau 1. Comparaison de différentes techniques d'usage de réseaux sociaux pour améliorer les mécanismes de recommandation ou de personnalisation.

3. Proposition d'un modèle de profil utilisateur social

Nos travaux se distinguent des travaux de la littérature par rapport aux deux éléments évoqués dans la section précédente :

1- Nature du lien : nous considérons les réseaux sociaux au sens sociologique du terme, c'est-à-dire dont la nature des liens entre nœuds correspond aux liens entretenus entre les individus dans la vie réelle (réseaux téléphoniques, réseaux

sociaux numériques, réseaux de co-auteurs, etc.). Ce choix se justifie par le fait que dans ce type de réseaux, il est possible d'exploiter la richesse et la pertinence des travaux qui ont déjà été menés dans le domaine des sciences sociales, et qui peuvent être passés à l'échelle. Ainsi, à la différence des travaux existants qui restent très spécifiques aux domaines ou mécanismes étudiés, nous pouvons définir un modèle de profil social générique et réutilisable, pour des réseaux sociaux au sens étymologique du terme.

2- Filtrage de nœuds : à la différence des travaux existants qui s'appuient sur des nœuds sélectionnés individuellement dans le réseau social de l'utilisateur, nous utilisons un filtrage basé sur les différentes communautés extraites dans le réseau égocentrique de l'utilisateur. Ce choix se justifie par la définition en sociologie d'un réseau égocentrique, que nous présentons dans la section suivante.

3.1. *Réseau égocentrique d'un utilisateur*

La notion de réseau égocentrique est principalement utilisée en sociologie (Granovetter, 73). 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. Cette notion peut être généralisée pour prendre en compte les utilisateurs situés à distance k de l'*égo* dans le réseau social. Si on considère un réseau social modélisé sous forme d'un graphe G = (V, E), V étant l'ensemble des individus, et E l'ensemble des liens entre ces individus. Le réseau égocentrique d'un individu v Є V peut être défini comme :

$$R_{ego}(v) = G' (V', E') \subset G / \forall\ u \in V', d(u, v) <= k \text{ et } \exists\ u' \in V' / (u, u') \in E' \quad \textbf{(1)}$$

d(u,v) représente la distance entre l'individu u et l'individu v, (u, u') représente un lien entre l'individu u et l'individu u'.

Pour k=1, le graphe G' correspond au réseau égocentrique tel que celui défini en sociologie. Ce réseau regroupe l'ensemble des nœuds à distance 1 de l'utilisateur ou égo (d (u, v)=1) et qui ont des relations entre eux ($\exists$ u' Є V'/ (u, u') Є E'). C'est ce réseau égocentrique que nous prenons en compte dans notre proposition. Toutefois, l'approche peut être généralisée pour des valeurs de k>1. Dans les graphes que nous étudions, les propriétés d'un nœud sont les attributs de son profil tel que défini sur le modèle de la figure 2, et nous nous intéressons particulièrement aux attributs qui représentent les centres d'intérêts de l'utilisateur.

Notre motivation pour l'usage du réseau égocentrique d'un utilisateur est liée au fait que nous considérons que dans la vie réelle, les centres d'intérêts d'un utilisateur sont directement liés aux différentes communautés dont il fait partie dans la société (Gofmann, 59). Par exemple, il est normal de considérer qu'un utilisateur qui est inscrit dans un club de tennis, possède le tennis comme centre d'intérêt. Pour retrouver ce centre d'intérêt, nous considérons le fait que les contacts (liens directs)

de l'utilisateur qui sont abonnés à ce club de tennis se connaissent également, et possèdent un nombre plus important de liens entre eux, par rapport aux autres contacts de l'utilisateur, qui ne sont pas abonnés à ce club. Pour revenir dans le contexte plus général du réseau égocentrique d'un utilisateur, l'exécution d'un algorithme de détection de communautés dans ce réseau, permettra d'extraire les différentes communautés autour de l'égo (l'utilisateur à « profiler »). La figure 1 présente une structuration manuelle en communautés d'un réseau égocentrique (Cardon, 05). Dans cette figure, on peut par exemple identifier des communautés de musiciens, internautes, etc.

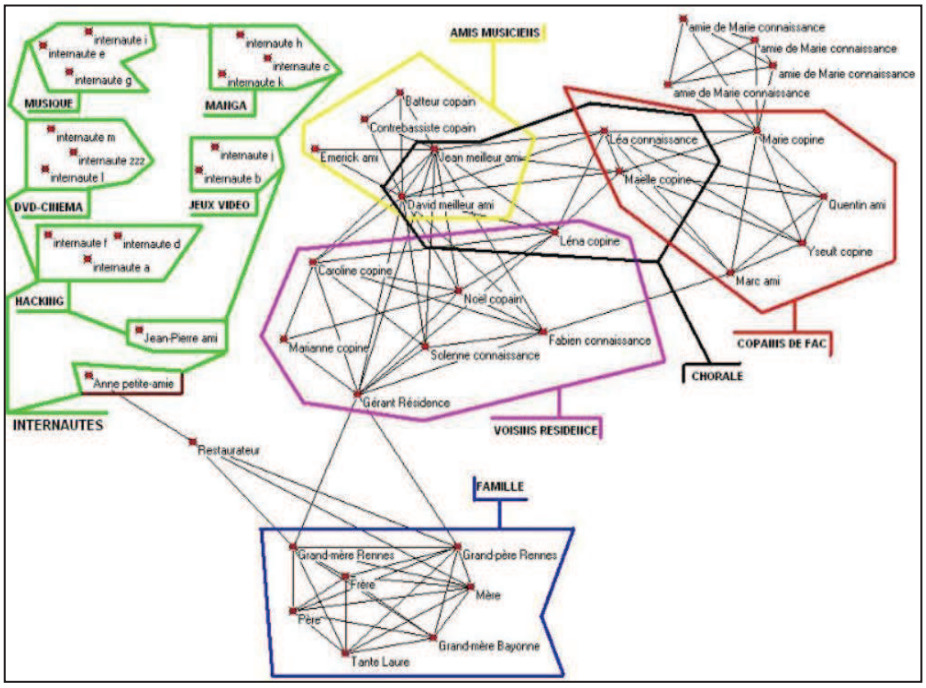

Figure 1. Exemple de structuration manuelle en communautés d'un réseau égocentrique (Cardon, 05).

3.2. *Filtrage basé sur des communautés du réseau égocentrique*

Pour réaliser les filtres de sélection des nœuds pertinents dans le réseau social de l'utilisateur, les travaux existants (tableau 1) s'appuient principalement sur des utilisateurs pris individuellement. 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 du profil social de l'utilisateur. Pourtant, il est difficile de considérer que tous les centres d'intérêts d'un utilisateur influant dans le réseau social, peuvent également se rapporter à l'égo. 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 pertinents pour l'utilisateur, même s'ils sont très actifs ou influents.

L'approche par les communautés du réseau égocentrique est 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 au profil de l'utilisateur. Par rapport aux méthodes « autoritaires », cette approche éliminera (uniquement de par le réseau égocentrique), 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 dans Facebook par exemple), et d'autre part les éléments de profil non significatifs pour l'utilisateur (même si ceux-ci émanent d'un ou plusieurs utilisateurs influents dans son réseau). Si on considère par exemple la figure 1, et qu'on suppose que l'on ne connaît pas le profil de l'utilisateur (égo) dont le réseau égocentrique est représenté, et que l'on peut identifier une communauté de musiciens dans ce réseau, il nous semble logique de déduire que l'utilisateur peut être probablement intéressé par la musique.

3.3. *Proposition d'un modèle social du profil utilisateur*

A la différence de la plupart des modèles de profils utilisateurs proposés dans la littérature (Bouzeghoub et al., 05) (Gauch et al., 07) (Abbar et al., 10), nous proposons un modèle social de profil utilisateur qui intègre des informations déduites sur l'utilisateur en provenance de son réseau social. Au lieu de s'intéresser au réseau social dans sa globalité selon une approche socio-centrée (nombre de nœuds trop élevés, complexité trop importante en temps de calcul, difficulté d'accès aux réseaux très larges, etc.) (Chung et al., 05), notre démarche se propose de ne s'intéresser qu'au voisinage de chaque utilisateur selon une approche égocentrée pour dériver son profil social (moins de difficulté à accéder aux données, communautés pertinentes, complexité de calcul réduite, etc.) (Everett et al., 99). Ainsi nous proposons un modèle de profil utilisateur composé *de deux grandes dimensions*, une *dimension utilisateur* (dont les centres d'intérêts sont calculés à partir des activités propres de l'utilisateur) et une *dimension sociale* (dont les centres d'intérêts sont calculés à partir des activités dans son réseau égocentrique). Ces dimensions sont caractérisées par les mêmes attributs (figure 2).Ce modèle intègre également les concepts existants dans les modèles proposés dans la littérature : données (attributs), sémantique (hiérarchie de domaines) et contexte.

La dimension sociale est composée des différentes communautés du réseau égocentrique de l'utilisateur. Plusieurs algorithmes dans la littérature sont dédiés à la détection de communautés dans les réseaux sociaux. Dans notre cas, il s'agira particulièrement de considérer des algorithmes permettant de générer des communautés recouvrantes (car un même individu peut appartenir à plusieurs

communautés dans le réseau égocentrique de l'utilisateur). Les communautés détectées peuvent être caractérisées par des mesures de structure qui dépendent uniquement de leurs liens vers d'autres communautés (centralité de degré, proximité ou intermédiarité par exemple) (Everett et al., 99). Il semble logique que ces attributs de structure puissent influencer l'importance de certains centres d'intérêt de l'utilisateur. Par exemple, on peut penser que l'affinité d'une communauté complètement isolée vis-à-vis des autres communautés ait une signification particulière pour l'utilisateur. Les communautés peuvent également être caractérisées par les mêmes attributs que l'on retrouve dans la dimension utilisateur du profil. Il s'agit en fait des attributs discriminants de la communauté par rapport aux autres communautés, qui vont permettre de retrouver l'affinité des individus de cette communauté avec l'utilisateur. Comme indiqué dans les paragraphes précédents, cette affinité peut être influencée par les mesures de structure de la communauté (nous y reviendrons dans la section suivante).

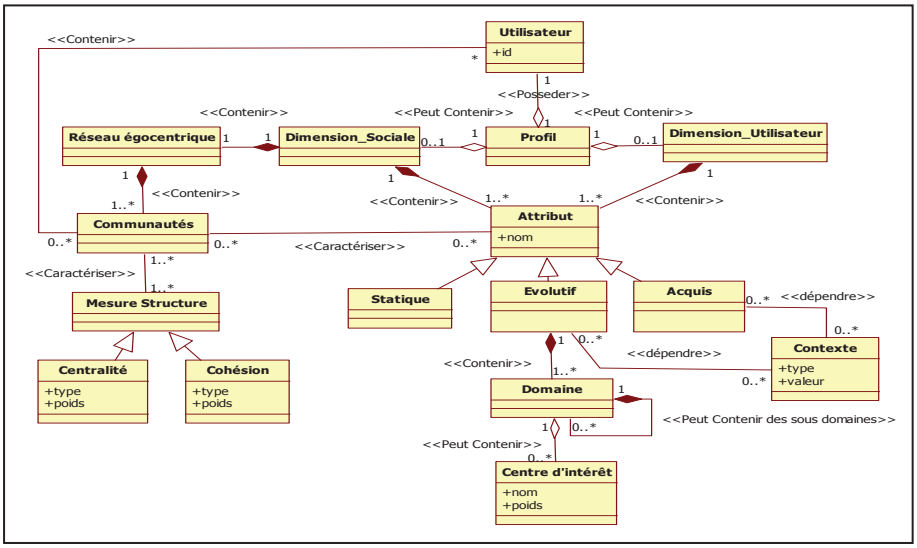

Figure 2. Proposition de modèle de profil utilisateur social

Les attributs du profil que nous considérons dans ce modèle sont liés à la nature des communautés pouvant être détectées dans le réseau égocentrique de l'utilisateur. Il s'agit des attributs que nous qualifions soit de statiques (exemple : nom, sexe) car invariants dans le temps, soit d'acquis (exemple : emplois occupés, établissements fréquentés) car acquis au cours du temps puis restant inchangés, soit d'évolutifs (les centres d'intérêts) car ils sont construits de manière itérative au fur et à mesure des interactions de l'utilisateur avec le système. Ces attributs sont structurés ainsi pour permettre de caractériser des communautés de natures différentes, par exemple des communautés statiques (comme la famille, attribut *nom*), des communautés acquises (comme les collègues, attribut *emploi occupé*), des communautés d'intérêts (comme le club de tennis, attribut *centre d'intérêt tennis*).

La sémantique des centres d'intérêts du profil (attributs évolutifs) est gérée par une taxonomie de concepts (domaines et sous domaines) qui sera identique pour tous les profils de tous les utilisateurs et dans les deux dimensions (utilisateur et sociale). Nous considérons que cette taxonomie peut être prédéfinie ou construite en fonction des domaines d'application (par un expert du domaine). La taxonomie pourra se limiter à la racine (vecteur de termes pondérés) s'il n'existe pas de taxonomie prédéfinie. Cette taxonomie pourrait être généralisée sous forme d'une ontologie de domaine. Toutefois, dans nos travaux, une taxonomie de concepts nous semble plus facile à définir qu'une ontologie (dans un cas assez simple d'expérimentation sous Facebook, nous avons réalisé manuellement une taxonomie dont un bref extrait est présenté sur la figure 3). Enfin, nous n'exploitons pas les possibilités de raisonnement offertes par les ontologies (règles d'inférences, typage de relations entre concepts, etc.). Pour un utilisateur, chaque niveau de la taxonomie se présente sous la forme d'un profil vectoriel du domaine correspondant (centres d'intérêts pondérés dans ce domaine). Chaque domaine est également pondéré par rapport aux domaines situés au même niveau dans l'arbre. Cette structuration permet d'exploiter le profil de l'utilisateur à des niveaux de granularité plus ou moins importants en fonction des besoins (un système pourrait par exemple avoir uniquement besoin du niveau d'intérêt de l'utilisateur pour le sport, alors qu'un autre pourrait vouloir plus de détails concernant les types de sport pour lesquels l'utilisateur a un intérêt).

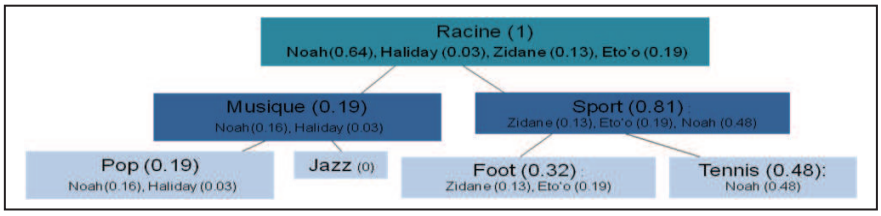

Figure 3. Exemple de profil structuré suivant une taxonomie

Sur la figure 3 par exemple, à la racine le centre d'intérêt *Noah* est vu de manière globale, pourtant à un niveau plus bas, on voit très bien la différence d'intérêt entre *Noah* chanteur et *Noah* sportif. Les centres d'intérêts sont calculés à partir des feuilles de la taxonomie. Chaque nouveau centre d'intérêt est positionné de manière adéquate sur une feuille correspondant à un chemin bien défini dans la taxonomie (ceci permettra par exemple de désambigüiser le centre d'intérêt *Noah* chanteur par rapport à *Noah* sportif, figure 3). Les calculs ont lieu au niveau de chaque feuille. Le poids d'un centre d'intérêt peut par exemple être sa mesure *tf* ou *tf-idf* dans un texte ou un ensemble de textes. Nous reviendrons sur le calcul des poids au niveau des feuilles pour la dimension sociale dans la section suivante. Ces poids sont normalisés de sorte à avoir une somme égale à 1. Les poids des centres d'intérêts sont remontés vers la racine via tous les ancêtres de la feuille concernée (figure 3). Le poids d'un domaine correspond toujours à la somme des poids des centres d'intérêts de ce domaine (ou la somme des poids des sous-domaines si le domaine n'est pas une feuille).

Enfin les attributs acquis et évolutifs du profil peuvent être regroupés suivant des contextes (figure 2) ou des ensemble de contextes bien précis (temps, espace, matériel, environnement, etc.). Ceci facilitera par exemple les usages des profils par différents mécanismes lorsque l'utilisateur se retrouve dans un contexte particulier. Nous avons traité seulement le cas du contexte temporel des attributs pour la visualisation des profils à court à terme et à long terme des utilisateurs (Tchuente et al., 10)(Tchuente et al., 12). Cet aspect n'est pas détaillé ici, nous nous focalisons dans cet article uniquement sur la dimension sociale du profil.

4. Algorithme de dérivation du profil social de l'utilisateur (CoBSP)

Si l'on considère un élément noté *e* du profil de l'utilisateur[4], cet élément sera associé à un poids noté P^{social} dans cette dimension sociale de son profil (e, P^{social}), et à un poids noté P dans la dimension utilisateur de son profil (e, P). Nous nous intéressons ici au calcul du poids P^{social} pour chaque élément de la dimension sociale du profil de l'utilisateur. Comme indiqué dans les sections précédentes, nous nous appuyons sur les communautés du réseau égocentrique pour dériver ce poids. Cette dérivation se fait suivant un processus (algorithme CoBSP) que nous décomposons en quatre phases : détection de communautés, profilage de communautés, caractérisation de communautés, et dérivation de la dimension sociale.

Pour un élément *e* de la dimension sociale du profil de l'utilisateur, des poids intermédiaires seront affectés à chacune des phases : P pour la détection de communautés, P' pour le profilage des communautés, P'', *Struct* et P''' pour la caractérisation des communautés, et enfin P^{social} pour la dérivation de la dimension sociale (figure 4). Chacune de ces phases est décrite dans les sections qui suivent.

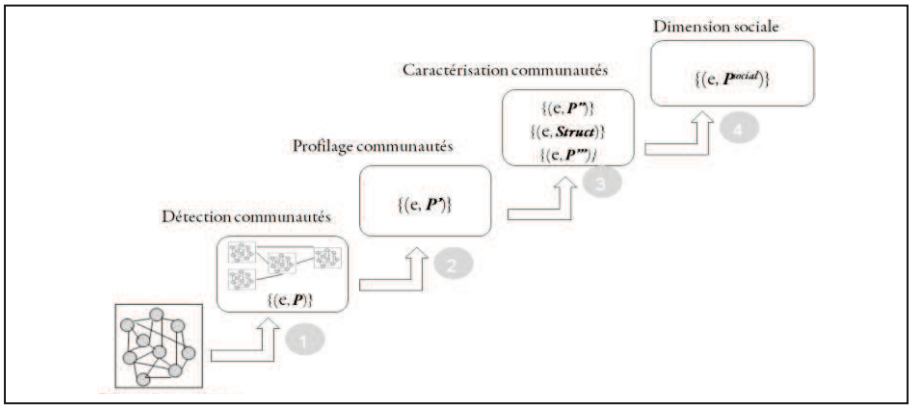

Figure 4. Etapes de l'algorithme de dérivation de la dimension sociale du profil

[4] e représente un couple (Attribut, Valeur), par exemple (Genre, Masculin) pour l'attribut statique Genre, (Employeur, IRIT) pour l'attribut acquis Employeur, ou (Sport, Tennis) pour un attribut évolutif lié au centre d'intérêt Sport

4.1. *Détection de communautés*

Dans cette première phase, seules les relations entre individus dans le réseau égocentrique sont exploitées pour la détection de communautés. Plusieurs algorithmes dans la littérature s'intéressent à la détection des communautés dans les réseaux sociaux. La qualité des communautés extraites est très souvent mesurée par des mesures telles que la modularité (Newmann, 06). Pour le choix de l'algorithme à utiliser, nous nous appuyons sur un travail d'évaluation par la perception humaine de la pertinence de plusieurs algorithmes de détection de communautés appliqués au cas spécifique de réseaux égocentriques dans Facebook (Cazabet et al., 12). D'après cette étude, les communautés extraites par l'algorithme iLCD (Cazabet et al., 10) en prenant en compte la notion de recouvrement (un individu peut appartenir à plusieurs communautés), ont été considérées comme étant les plus pertinentes par les retours d'utilisateurs. Ceci justifie notre choix de cet algorithme dans nos travaux. Toutefois, l'algorithme à utiliser dans cette phase est paramétrable, il est donc possible de le modifier à tout moment si un autre algorithme est jugé plus pertinent.

Une fois les communautés détectées, chaque communauté contient les dimensions utilisateur des profils de ses membres. Ainsi dans une communauté, à un élément de profil *e,* on associe les différentes couples *(e, P)* de cet élément de profil dans les dimensions utilisateurs des profils des membres de cette communautés.

4.2. *Profilage de communautés*

Si l'on considère une communauté, pour un élément de profil *e* d'un individu de cette communauté, on l'associe un poids P' *(e, P')*, P' désigne ici la moyenne des pondérations P associé à l'élément *e* dans les dimensions utilisateur des profils des membres de la communauté.

Considérons par exemple une communauté composé de trois individus ayant respectivement les couples (e, P) : (e, 0.12), (e, 0.15), (e, 0.29) dans la dimension utilisateur de leur profil. Le poids *P'* associé à *e* dans le profil de cette communauté va valoir (0.12+0.15+0.29)/3=**0.14**.

4.3. *Caractérisation de communautés*

Les communautés peuvent être caractérisées de deux manières : par les attributs du profil de la communauté (nous parlerons ici de caractérisation sémantique), et par ses relations dans le réseau égocentrique (nous parlerons ici de caractérisation structurelle). Ces deux caractérisations seront associées pour caractériser chaque communauté.

Au-delà du profil d'une communauté, ***la caractérisation sémantique*** d'une communauté a pour but de rechercher la spécificité de cette communauté

relativement aux autres communautés. Si l'on dispose par exemple de *m* communautés, pour un couple (e, P') du profil d'une communauté, le poids P'' associé à e par la caractérisation sémantique est donné par :

$$P'' = P' * \log \frac{m}{\sum_{j=1}^{m} presence(e, C_j)} \qquad \textbf{(2)}$$

Dans cette formule *presence (e, C_j)* vaut 1 si l'élément e existe dans le profil de la communauté C_j, et *presence (e, C_j)* vaut 0 sinon. Le principe utilisé ici est le même que celui de l'usage du *tf-idf* (Salton, 78) utilisé en recherche d'information pour caractériser un document particulier dans un ensemble de documents.

Considérons par exemple le cas de quatre communautés (C1, C2, C3, C4) dans lesquelles le profil de chacune des communautés est défini par les poids P' associés aux éléments des profils notés A, B, C, D (tableau 2). Les poids P'' calculés suivant la formule 2 sont présentés dans le tableau 3. On peut remarquer sur cet exemple que l'élément de profil C possède une pondération P'' de caractérisation sémantique très élevée, car C possède un poids P' important (0.9) dans la communauté C4, et surtout, C possède un poids P' nul (représenté par l'absence de valeur sur les figures) dans les autres communautés. Donc C est très caractéristique de la communauté C4. A l'inverse, l'élément de profil B a une caractérisation sémantique moins importante, car il n'est pas assez caractéristique d'une seule communauté.

P'	**A**	**B**	**C**	**D**
C1	**0.14**	**0.86**		**0.71**
C2	**0.1**			**0.1**
C3	**0.05**	**0.05**		
C4			**0.9**	**0.1**

Tableau 2. Exemple de profils de communautés

P''	**A**	**B**	**C**	**D**
C1	0.14*log4/3=**0.017**	0.86*log4/2=**0.258**		0.71*log4/3=**0.088**
C2	0.1*log4/3=**0.012**			0.1*log4/3=**0.012**
C3	0.05*log4/3=**0.006**	0.05* log4/2=**0.015**		
C4			0.9*log4/1=**0.541**	0.1*log4/3=**0.012**

Tableau 3. Exemple de profils de communautés caractérisés sémantiquement

En parallèle de la caractérisation sémantique, ***la caractérisation structurelle*** d'une communauté a pour but de différencier la pertinence d'une communauté pour l'utilisateur en s'appuyant sur la position ou la structuration de cette communauté dans le réseau égocentrique. On peut par exemple penser logiquement qu'une communauté centrale et une communauté isolée dans le réseau égocentrique d'un utilisateur n'ont pas la même signification pour ce dernier. Pour matérialiser la caractérisation structurelle, nous nous appuyons sur les mesures de centralité de communautés (degré, proximité, intermédiarité, etc.) dans les réseaux sociaux

proposés par (Everrett et Borgatti., 99). Par exemple, la centralité de proximité d'une communauté mesure la distance de rapprochement de la communauté par rapport à tous les individus externes à cette communauté (formule 3). La notion de distance de la communauté C vers un individu *x* externe à C peut être définie de plusieurs manières (distance moyenne ou minimale ou maximale de tous les individus de C vers cet individu, $d_f(x, C)$). Plus la proximité est élevée, plus la communauté est centrale dans le réseau égocentrique de l'utilisateur.

Cette centralité peut ainsi être exploitée pour calculer la pondération de structure *Struct* associée à une communauté noté C (et ainsi à l'ensemble des éléments du profil de cette communauté) dans la formule 3.

$$Struct = \frac{|V - C|}{\sum_{x \in V - C} d_f(x, C)} \quad \textbf{(3)}$$

V : ensemble des nœuds du graphe.
C : ensemble des nœuds de la communauté pour laquelle la mesure est calculée
$d_f(x, C)$: distance entre un individu x et une communauté C *(f indique s'il s'agit de la distance moyenne, minimale ou maximale)*

La caractérisation finale P''' de chaque élément *e* du profil d'une communauté C_i est une caractérisation sémantico-structurelle qui va prendre en compte le poids P'' de la caractérisation sémantique et le poids *Struct* de la caractérisation structurelle en fonction d'un paramètre α compris entre 0 et 1 (formule 4).

$$P''' = \alpha Struct(C_i) + (1 - \alpha) P'' \quad \textbf{(4)}$$

Le paramètre α dont la valeur optimale sera déterminée expérimentalement permettra de juger de l'importance de l'usage des mesures de structure dans la dérivation de la dimension sociale du profil de l'utilisateur.

Si l'on considère l'exemple des tableaux 2 et 3, et que l'on suppose que les poids de structure *Struct* dans le réseau égocentrique de l'utilisateur valent respectivement 0.9, 0.5, 0.3 et 0.8 pour les communautés C1, C2, C3 et C4, et si l'on fixe arbitrairement ici la valeur du paramètre α à 0.1, le tableau 4 présente les poids P''' calculés pour chacun des éléments de profil A, B, C, D dans chacune des communautés.

P'''	**A**	**B**	**C**	**D**
C1	0.1*0.9+0.9*0.017=**0.105**	0.1*0.9+0.9*0.258=**0.322**		0.1*0.9+0.9*0.088=**0.169**
C2	0.1*0.5+0.9*0.012=**0.06**			0.1*0.5+0.9*0.012=**0.06**
C3	0.1*0.3+0.9*0.006=**0.035**	0.1*0.3+0.9*0.015=**0.043**		
C4			0.1*0.8+ 0.9*0.541 =**0.566**	0.1*0.8+0.9*0.012=**0.09**

Tableau 4. Exemple de caractérisation de profils de communautés

4.4. *Dérivation de la dimension sociale*

Pour un élément de profil *e,* plusieurs communautés pourront avoir une pondération P''' de caractérisation sémantico-structurelle. La question qui se pose ici est alors de savoir quelle pondération définitive sera affectée à cet élément de profil dans la dimension sociale du profil de l'utilisateur. Dans un cas idéal, cet élément de profil peut se retrouver dans le profil d'une seule communauté. Ceci indiquera alors que seule, cette communauté est complètement discriminante pour cet élément. Dans ce cas, cet élément pourra alors être directement rapporté avec sa pondération dans le profil social de l'utilisateur (dans la mesure où il représente l'affinité entre l'utilisateur et cette communauté). Dans le cas où plusieurs communautés partagent un même élément de profil avec des poids P'''>0, le principe consiste alors à trouver une combinaison optimale de ces poids pour avoir le poids P^{social} dans la dimension sociale du profil de l'utilisateur. Nous partons de l'hypothèse selon laquelle plus le poids P''' d'un élément de profil est important dans une seule communauté, plus cet élément est important pour l'utilisateur (ego). La combinaison des poids devra donc privilégier le score des communautés ayant le(s) poids(s) le(s) plus élevé(s) pour cet élément de profil.

Cette problématique peut se rapprocher de celle de la combinaison des systèmes de classement de plusieurs moteurs de recherche, pour classer des documents en recherche d'information. Chaque système apportant un score de classement à un même document, le but consiste à déterminer le score final du document. En général, la fonction de combinaison CombMNZ (Shaw et Fox, 94) est utilisée. Une variante de cette fonction proposée dans (Hubert et al., 07) permet de réaliser la combinaison linéaire des classements, en privilégiant un document si un seul des moteurs de recherche l'a très bien classé. Nous assimilons notre cas à ce principe, en faisant les analogies suivantes : a) les documents correspondent aux éléments de profil, b) les systèmes de classement correspondent aux communautés du réseau égocentrique. La communauté ayant le poids le plus important pour un élément de profil sera ainsi privilégiée dans la combinaison linéaire. Ceci nous permet de déduire le poids P^{social} d'un élément de profil dans la dimension sociale du profil de l'utilisateur suivant la formule 5. Dans cette formule, m représente le nombre total de communautés.

$$P^{social} = \sum_{j=1}^{m / P'''(C_{j-1}) < P'''(C_j)} (P'''_{AxVx}(C_j) * j) \qquad \textbf{(5)}$$

Pour calculer le poids P^{social} de chaque élément du profil, les communautés sont d'abord ordonnées de façon croissante en fonction de leur pondération (sémantico-structurelle) sur cet élément ($P'''_{(C_{j-1})} < P'''_{(C_j)}$). Ainsi, la communauté qui dispose du poids le plus important est privilégiée et son score est multiplié par son rang de classement, par exemple multiplication par 5 si l'on dispose de 5 communautés, le second meilleur score sera multiplié par 4,…, la communauté disposant du poids le moins important dans ce cas verra son score multiplié par 1.

Si l'on reprend l'exemple des tableaux 2, 3, 4, les calculs des poids P^{social} suivant la formule 5 sont présentés dans le tableau 5. Les résultats peuvent ensuite être normalisés en divisant par la somme des poids P^{social} calculés.

P^{social}	A	B	C	D
Dimension sociale	0.105*4+0.06*3+0.035*2 **= 0.67**	0.322*4+0.043*3 **=1.417**	0.566*4 **=2.264**	0.169*4+0.09*3+0.06*2 **=1.066**
Normalisation	**0.123**	**0.261**	**0.417**	**0.196**

Tableau 5. Exemple de dérivation de la dimension sociale

A travers cet exemple, on peut constater le poids important de l'élément de profil C dans la dimension sociale, relativement à l'élément B qui n'était pas spécifiquement caractéristique d'une communauté.

5. Evaluation

Le but de l'évaluation est de montrer la pertinence de l'algorithme basé sur les communautés décrit précédemment par rapport aux algorithmes équivalents basés plutôt sur les individus tels que ceux actuellement utilisés dans la littérature. La stratégie d'évaluation que nous adoptons ici consiste à : (i) considérer des utilisateurs très actifs dont la dimension utilisateur du profil est jugée très riche, (ii) dériver la dimension sociale de leur profil avec l'algorithme proposé basé sur les communautés et les algorithmes équivalents basés sur les individus, (iii) déduire l'algorithme de dérivation sociale qui prédit le mieux la dimension utilisateur (connue) du profil de l'utilisateur (usage du cosinus de similarité par exemple).

Pour mettre en œuvre ce processus, nous présentons dans un premier temps les algorithmes basés sur les individus (que nous avons définis à partir des travaux de la littérature) qui seront comparés à l'algorithme proposé. Ensuite, nous présentons les détails de l'évaluation réalisée dans Facebook.

5.1. *Algorithmes basés sur les individus*

Les algorithmes basés sur les individus qui sont utilisés dans la littérature peuvent être divisés en deux selon notre définition du réseau égocentrique : 1) ceux qui différencient chaque individu du réseau égocentrique de l'utilisateur (avec le poids de sa relation avec l'égo ou sa centralité dans le réseau égocentrique par exemple), 2) ceux qui ne font aucune différence entre les individus du réseau égocentrique.

Cas 1 : algorithme qui différencie chaque individu du réseau égocentrique de l'utilisateur (IBSP1)

Le principe de cet algorithme consiste à exploiter les individus du réseau égocentrique de l'utilisateur suivant des phases quasi similaires à celles décrites précédemment pour l'algorithme CoBSP. Concrètement, il s'agit d'une part, de réaliser une caractérisation sémantico-structurelle (cf. formule 4) en différenciant

chaque individu du réseau égocentrique par sa mesure de centralité (degré, proximité et intermédiarité) et la dimension utilisateur de son profil, et d'autre part, de combiner les profils des individus ainsi caractérisés par la fonction de combinaison linéaire définie dans la formule 5.

Cas 2 : algorithme qui ne fait aucune différence entre les individus du réseau égocentrique de l'utilisateur (IBSP2)

Il s'agit ici du cas trivial pour lequel la dimension sociale du profil de l'utilisateur est construite par simple agrégation (moyenne) des dimensions utilisateur des profils des membres de son réseau égocentrique. Aucune distinction n'est alors faite entre les membres du réseau égocentrique, ils ont tous la même importance pour l'égo (pas d'usage de mesure de structure).

5.2. *Extraction de données de réseaux égocentriques dans Facebook*

Pour l'évaluation des algorithmes présentés, nous avons exploité les données du réseau social Facebook pour deux raisons : la sémantique du lien (amitié) entre deux utilisateurs Facebook indique à priori que les deux utilisateurs se connaissent dans la vie réelle, et le réseau social Facebook est actuellement le plus utilisé en France et dans le monde.

Pour réaliser l'évaluation décrite précédemment pour un utilisateur donné dans Facebook, il était nécessaire d'accéder à la fois à ses activités (pour construire la dimension utilisateur de son profil), à son réseau égocentrique, et aux activités des utilisateurs de son réseau égocentrique (pour construire la dimension sociale de son profil avec chacun des trois algorithmes présentés dans les sections précédentes). Comme dans (Tchuente et al., 12), nous avons exploité l'API Facebook qui permet d'accéder à ces données sous réserve de l'accord explicite de chaque utilisateur accédant à une application tierce développée sur Facebook. Nous avons donc développé une application tierce[5] qui nous a permis d'accéder à des données d'utilisateurs volontaires sur Facebook. Le tableau 6 présente les principales données accessibles (correspondant au modèle de profil social proposé) à partir d'un utilisateur qui installe une application tierce dans Facebook.

De ce tableau, il ressort que pour accéder aux attributs statiques (nom et sexe par exemple), acquis (établissements fréquentés et historique des emplois par exemple) et évolutifs (publication de statuts, liens, ou connexion à des groupes par exemple) de l'utilisateur de l'application ainsi que pour ceux des membres de son réseau égocentrique, il faut demander des permissions supplémentaires à l'utilisateur. Nous avons donc demandé ces permissions supplémentaires à des utilisateurs volontaires. 64 utilisateurs (étudiants et enseignants) se sont portés volontaires pour installer et donner ces permissions supplémentaires à notre application. Cependant dans cette évaluation, nous avons uniquement analysé les réseaux égocentriques de 15 de ces

[5] apps.facebook.com/egoaccess/

utilisateurs qui ont été jugés « suffisamment » actifs pour construire une dimension utilisateur significative de leur profil (nous expliquons pourquoi dans la section qui suit). Les 15 réseaux égocentriques étudiés ici, ont en moyenne 296 utilisateurs chacun, pour un total de 4440 profils utilisateurs Facebook analysés dans cette évaluation.

	Données accédées par défaut		**Données accédées avec permissions supplémentaires de l'utilisateur**
Utilisateur	***Réseau égocentrique :***	*Accessible*	*N/A*
	Attributs statiques :	*Accessibles* (nom, sexe par exemple).	*Exemple :* intérêts explicitement renseignés par l'utilisateur, hobbies, émissions télés.
	Attributs Acquis :	*Non accessibles*	*Exemple :* historique des emplois, cursus académique, ville de résidence
	Attributs évolutifs:	*Non Accessibles*	*Exemple :* éléments publiés (statuts, liens, notes, photos, vidéos, groupes, pages, événements, etc.)
Amis de l'utilisateur	***Réseau égocentrique :***	*Non Accessible*	*Non Accessible*
	Attributs statiques :	*Accessibles* (nom, sexe par exemple).	*Exemple :* intérêts explicitement renseignés par l'utilisateur, hobbies, émissions télés.
	Attributs acquis :	*Non accessibles*	*Exemple :* historique des emplois, cursus académique, *ville de résidence*
	Attributs évolutifs :	*Non accessibles*	*Exemple :* éléments publiés (statuts, liens, notes, photos, vidéos, groupes, pages, événements, etc.*)*

Tableau 6. Classification et accessibilité des données du modèle de profil social dans Facebook par une application tierce

5.3. *Construction des profils à partir des activités dans Facebook*

Les données exploitées pour chacun des attributs du modèle de profil social proposé sont les suivants :

- *Attributs statiques :* nom et genre
- *Attributs acquis :* liste des établissements fréquentés et liste des emplois occupés
- *Attributs évolutifs* : liste des pages fans, liste des groupes rejoints, liste des événements auxquels l'utilisateur a participé.

Dans la suite de cette partie, nous désignerons par centres d'intérêts, les éléments des profils issus des attributs évolutifs.

Pour les attributs évolutifs, nous exploitons les pages, groupes et événements car ces applications sont déjà catégorisées par Facebook, ce qui permet de déduire plus facilement une connexion d'un utilisateur vers ces applications comme un intérêt de l'utilisateur pour les catégories (domaines d'intérêts) de ces applications. Si un utilisateur est par exemple fan de plusieurs pages liées à la catégorie politique, il parait logique d'en déduire qu'il s'intéresse à la politique ; et dans ce cas, ses centres d'intérêts en politique seront construits à partir des descriptions textuelles (titre et

description par exemple) des pages de la catégorie politique dont il est un fan. De plus, la structuration en catégorie des pages, groupes et événements nous permet d'avoir une taxonomie de référence pour tous les profils, ce qui correspond à la structure de données due modèle proposé.

A partir des catégories et descriptions des pages, groupes et événements, les profils des utilisateurs (dimension utilisateur et dimension sociale) sont construits selon le processus de la figure 5. La dimension utilisateur du profil d'un utilisateur est construite par analyse des pages, groupes et événements auxquelles cet utilisateur est connecté. Les 15 utilisateurs choisis parmi les 64 qui ont installé l'application, sont ceux qui sont connectés à au moins 200 groupes, pages et événements, et dont nous jugeons par conséquent qu'ils auront une dimension utilisateur assez riche en informations pour être comparée à la dimension sociale construite par chacun des algorithmes de dérivation de la dimension sociale. La dimension sociale est construite par analyse des pages, groupes et événements des utilisateurs ou communautés du réseau égocentrique de l'utilisateur. Le processus de construction des profils de la figure 5 est décomposé en 4 grandes étapes : extraction des catégories et unités sémantiques, construction des centres d'intérêts à partir des unités sémantiques, insertion des centres d'intérêts dans la feuille correspondante de la taxonomie de référence, remontée des centres d'intérêts dans la taxonomie (Tchuente et al., 12).

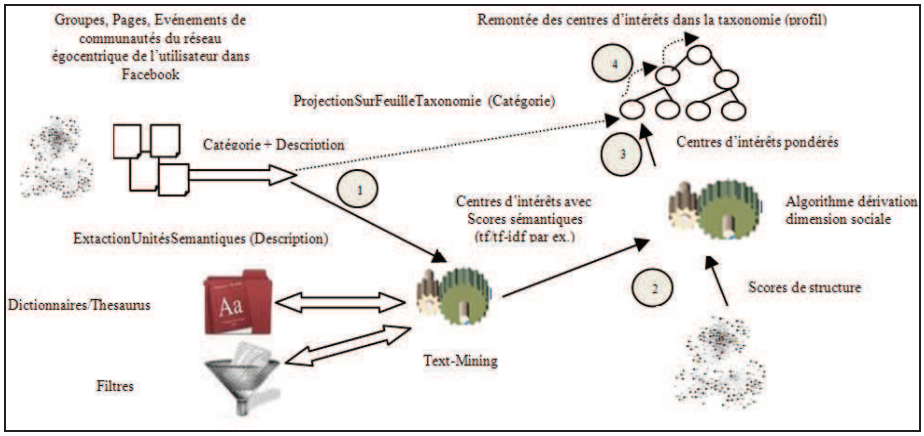

Figure 5. Méthodologie de construction des centres d'intérêts par fouille de textes dans Facebook

L'étape 1 (extraction des catégories et unités sémantiques) consiste à :

- Extraire la catégorie d'une page, groupe ou événement et la projeter sur la feuille correspondante de la taxonomie de référence.
- Extraire les mots distincts à partir du texte descriptif de la page, groupe ou événement avec un ensemble de séparateurs (espace, virgule, point virgule,

apostrophe, etc.). Chaque mot extrait est considéré comme une unité sémantique.

L'étape 2 (construction des centres d'intérêts à partir des unités sémantiques) consiste à :

- Enlever les unités sémantiques non significatives avec des listes de mots vides (articles de langue française par exemple) ou retenir uniquement des mots spécifiés par un expert du domaine à partir de filtres similaires à ceux exploités dans la plateforme Tétralogie[6] de l'IRIT (Tchuente et al., 12 bis). Fusionner éventuellement les unités sémantiques similaires (synonymes par exemple) en exploitant des dictionnaires ou des thesaurus (Tchuente et al., 12 bis). Une fois ces traitements réalisés, chaque unité sémantique (mot) retenue est considérée comme un centre d'intérêts de l'utilisateur.

- Calculer le poids de ce centre d'intérêts en fonction du score sémantique (tf ou tf-idf par exemple) et du score de structure (degré de centralité de communauté par exemple) tel que présenté dans les algorithmes de dérivation de la dimension sociale. Pour la dimension utilisateur du profil, seul le score sémantique est considéré.

- Utiliser chacun des algorithmes de dérivation de la dimension sociale à partir des profils des individus ou des communautés.

L'étape 3 (insertion des centres d'intérêts dans la feuille correspondante de la taxonomie de référence) consiste mettre à jour la feuille de la taxonomie correspondant à la catégorie de l'étape 1 avec les centres d'intérêts ainsi calculés.

L'étape 4 (remontée des poids des centres d'intérêts dans la taxonomie) consiste à mettre à jour la taxonomie au niveau de la feuille correspondante (à la catégorie projetée de l'étape 1) vers la racine selon le principe présenté dans la section 3.3 (figure 3).

Bien que les étapes présentées correspondent aux attributs évolutifs (taxonomie), le même principe est utilisé pour les attributs statiques (nom et sexe) et acquis (établissements et emplois) à la seule différence qu'aucune taxonomie n'est utilisée dans ces cas.

5.4. *Exemple de profil construit*

Les profils évolutifs construits étant représentés sous forme d'une taxonomie, ils peuvent être exploités à des niveaux de granularité plus ou moins élevée en fonction des besoins. La comparaison entre dimension utilisateur et dimension sociale peut également se faire à plusieurs niveaux de granularité.

---

[6] http://atlas.irit.fr

Figure 6. Dimension utilisateur catégorie « sport » du profil d'un utilisateur (à gauche) et dimension sociale (dérivée par l'algorithme proposé CoBSP) du profil « sport » du même utilisateur (à droite).

La figure 6 présente sous forme de nuages de mots la catégorie « sport » des profils construits pour un utilisateur très actif dans le domaine « sport ».

A droite (figure 6b) sont présentés les éléments de la dimension sociale du profil construit par l'algorithme basé sur les communautés *CoBSP*. Ce profil a été présenté à l'utilisateur en question (sans notifier qu'il a été construit à partir de son réseau social). Ce dernier a validé la quasi-totalité des termes qui y figurent. La dimension utilisateur construite à partir des activités de cet utilisateur est présentée sur la figure 6a. On peut constater que les centres d'intérêts les plus importants de la dimension utilisateur (figure 6a) se retrouvent également dans la dimension sociale (figure 6b), cas par exemple de BasketBall, France, NBA, Limoges, Judo. Ce qui montre bien que même si l'on ne disposait pas de la dimension utilisateur du profil « sport » de cet utilisateur, elle pourrait bien être dérivée à partir de son réseau social (algorithme basé sur les communautés). Cet exemple, nous permet de montrer que l'approche de dérivation par communautés sur le réseau social peut être pertinente. Toutefois, qu'en est-il des autres approches basées sur les individus ? Pour faire la comparaison des trois approches *(CoBSP, IBSP1, ISBP2)* nous avons calculé pour les 15 utilisateurs choisis, les cosinus de similarité entre la dimension utilisateur de leur profil et chacune des dimensions sociales construites par les trois approches que nous souhaitons évaluer. L'approche qui donnera le cosinus de similarité le plus important par rapport aux dimensions utilisateur des profils, sera considérée comme celle qui prédit le mieux le profil « réel » de l'utilisateur, donc la plus pertinente.

La figure 7 présente cette comparaison pour chacun des attributs statiques, acquis et évolutifs.

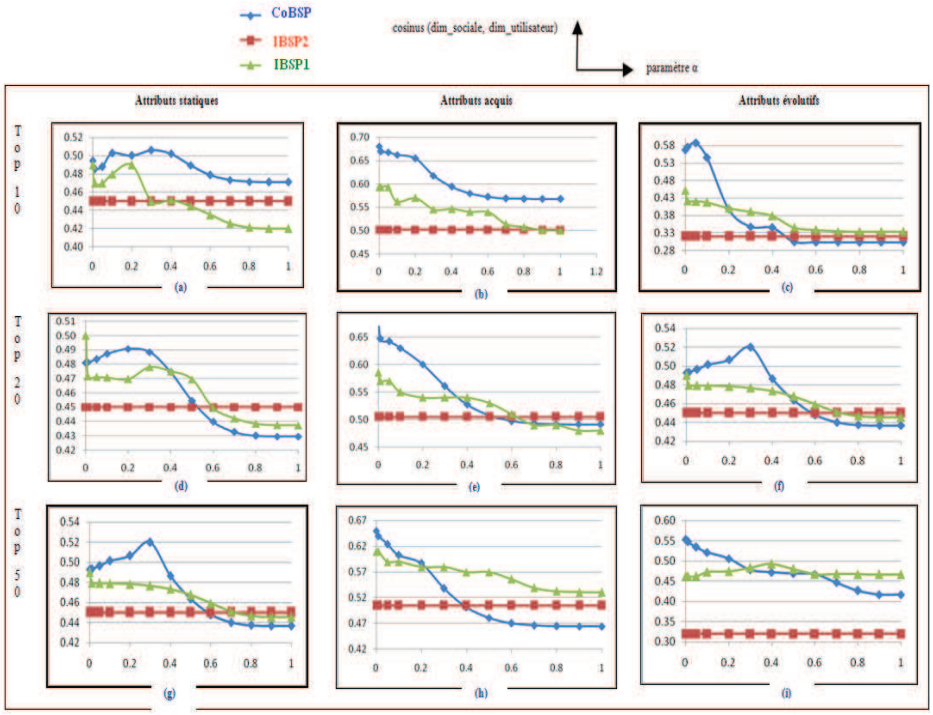

Figure 7. Comparaison cosinus de similarité entre les dimensions utilisateurs et les dimensions sociales construites par les algorithmes CoBSP (losanges), IBSP1 (triangles), ISBP2 (carrés)

Afin d'évaluer la pertinence de chaque algorithme sur les centres d'intérêts les plus importants (poids les plus élevés) qu'ils calculent, nous comparons les 10 premiers centres intérêts calculés par chaque algorithme (top 10), les 20 premiers centres d'intérêts calculés par chaque algorithme (top 20) et les 50 premiers centres d'intérêts calculés par chaque algorithme (top 50). Il est à noter que la mesure de structure utilisée dans cette expérimentation est le degré de centralité des communautés (pour l'algorithme CoBSP) et le degré de centralité des utilisateurs (pour l'algorithme IBSP1) dans le réseau égocentrique de chacun des 15 utilisateurs.

Trois principales conclusions ressortent de cette figure :

- Que ce soit pour les attributs statiques, acquis ou évolutifs, l'algorithme basé sur les communautés, courbe bleue (CoBSP) se rapproche le plus de la dimension utilisateur des profils des utilisateurs (cosinus le plus élevé parmi toutes les courbes).

- Pour de faibles valeurs de α (entre [0, 1] notamment), l'algorithme basé sur les communautés, courbe bleue (CoBSP) fournit de meilleurs résultats que les autres algorithmes. Pour le top 10, cet algorithme est largement meilleur

par rapport que pour le top 20 et le top 50 (bien qu'il reste tout de même meilleur). Ceci implique donc que l'algorithme CoBSP est beaucoup plus performant pour prédire les centres d'intérêts les plus importants de l'utilisateur.

- Les courbes représentant les algorithmes qui dépendent de α (CoBSP et IBSP1) décroissent lorsque les valeurs de ce paramètre augmentent. Ceci est logique dans la mesure où des valeurs élevées de ce paramètre, impliquent une moindre prise en compte des poids réels (sémantique) des centres d'intérêts. Il est normal que le poids sémantique d'un centre d'intérêt soit plus important que les mesures de structures des utilisateurs ou des communautés. Cependant, ces variations en fonction du paramètre α montrent que les mesures de structures des utilisateurs ou des communautés peuvent impacter la qualité des profils « sociaux » construits en améliorant les résultats pour les valeurs de α entre [0, 1] notamment.

Il ressort de cette expérimentation que l'algorithme proposé basé sur les communautés (*CoBSP*) est plus performant que les algorithmes basés sur les individus. Il est tout de même important de noter que cette expérimentation est réalisée sur 15 réseaux égocentriques, ce qui pourrait être considéré comme peu pour généraliser les résultats obtenus. Toutefois, l'expérimentation montre que l'approche basée sur les communautés peut se révéler très pertinente, et des expérimentations ultérieures sur un plus grand nombre de réseaux égocentriques seront réalisées pour confirmer ce résultat.

6. Conclusion et perspectives

Pour répondre efficacement à plusieurs mécanismes d'adaptation de l'information à l'utilisateur dans les systèmes d'information, nous avons proposé dans ce papier un modèle générique de profil utilisateur intégrant une dimension sociale via le réseau égocentrique de l'utilisateur. A partir de ce modèle, nous avons proposé un algorithme de dérivation de la dimension sociale du profil de l'utilisateur à partir des communautés de son réseau égocentrique. Cet algorithme s'appuie sur des scores portant sur la structure des communautés et sur des scores sémantiques de caractérisation des centres d'intérêts dans chaque communauté. Par rapport aux techniques existantes, cette approche s'appuie sur le principe d'affinité dans les communautés autour de l'utilisateur, plutôt que sur celui d'autorité (utilisateurs pris individuellement) dans le réseau social de l'utilisateur. Une expérimentation, réalisée avec l'analyse de 15 réseaux égocentriques très actifs dans Facebook, nous a permis de montrer la potentielle pertinence de l'utilisation cet algorithme, comparé aux approches exploitées dans la littérature, et l'importance que pourrait avoir les mesures de structures des utilisateurs ou des communautés dans la qualité des profils construits. Toutefois, pour confirmer les résultats obtenus, il serait important de réaliser les tests sur des échantillons de données plus importants dans des environnements où l'on dispose d'un accès plus ouvert aux données (réseaux de co-

auteurs DBLP par exemple). Plusieurs autres améliorations peuvent être ajoutées à l'algorithme proposé ou au processus d'évaluation : prise en compte des poids des liens entre utilisateurs lorsque ceux-ci existent dans le réseau social, prise en compte d'autres mesures de centralité dans les expérimentations (proximité, intermédiarité ou cohésion par exemple), analyse des résultats suivant la densité des réseaux égocentriques étudiés, évaluation des profils sociaux par confrontation à la perception humaine (demander aux utilisateurs de comparer eux-mêmes les profils construits par exemple), définition de techniques d'usage de la dimension sociale et de la dimension utilisateur dans des mécanismes d'adaptation de l'information à l'utilisateur (personnalisation ou recommandation de contenus par exemple).

7. Bibliographie

Abbar S, Bouzeghoub M., Lopes S., *Introducing Contexts into Personalized Web Applications*, in: International Conference on Information Integration and Web-based Applications & Services (IIWAS 2010), Paris, France, pp.155-162.

Bender M., Crecelius T., Kacimi M., Michel S., Neumann T.; Parreira J.X., Schenkel R., Weikum G., "*Exploiting social relations for query expansion and result ranking*" Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on , vol., no., pp.501-506, 7-12 April 2008.

Bouadjenek M. R., Hacid H., Bouzeghoub M., Daigremont J., *Une Nouvelle Approche d'Expansion Sociale de Requêtes dans le Web 2.0*. Huitième édition de la COnférence en Recherche d'Information et Applications, CORIA 2011: 41-48.

Bouzeghoub M., Kostadinov D.: *Personnalisation de l'information: aperçu de l'état de l'art et définition d'un modèle flexible de profils*. CORIA 2005: 201-218.

Cabanac G., *Accuracy of inter-researcher similarity measures based on topical and social clues,* Scientometrics 87: 3. 597-620 May 2011.

Cardon D., Ecole Grands Réseaux d'Interactions, Paris, Avril 2005.

Carmel D., Zwerdling N., Guy I., Ofek-Koifman S., Har'el N., Ronen I., Uziel Erel., Yogev S., Chernov S. 2009. *Personalized social search based on the user's social network*. *In* Proceedings of the 18th ACM conference on Information and knowledge management (CIKM '09).New York,USA, 1227-1236.

Cazabet, R.; Amblard, F.; Hanachi, C.; "*Detection of Overlapping Communities in Dynamical Social Networks*," Social Computing (SocialCom), 2010 IEEE Second International Conference, pp.309-314, 20-22 Aug. 2010.

Cazabet, R.; L. Maud; Amblard, F.; *Automatic Community Detection in online social networks, Useful ? Efficient ? Asking the users.* 4th International Workshop on Web Intelligence & Communities, Lyon, 2012.

Everett, M. G., & Borgatti, S. P. 1999. The *centrality of groups and classes*. Journal of Mathematical Sociology. 23(3): 181-201.

Fox E.A., Shaw J. A., "*Combination of Multiple Searches, the 2nd Text Retrieval Conference (TREC-2)",* NIST Special Publication 500-215, pp. 243-252, 1994.

Gauch S., Mirco S., Aravind C., Alessandro M. (2007). *User profiles for Personalized Information Acces*s, in: The Adaptive Web, Vol. 4321 (2007), pp. 54-89.

Goffman E. *The presentation of self in everyday life*. 1959. Garden City, NY, 2002.

Granovetter M.S., *The Strength of Weak Ties*, in: The American Journal of Sociology, Vol. 78. No. 6, May 1973, pp. 1360-1380.

Hubert G., Loiseau Y. et Mothe J., *Etude de différentes fonctions de fusion de systèmes de recherche d'information*, CIDE 10 : Le document numérique dans le monde de la science et de la recherche, Nancy, 02/07/2007-04/07/2007, EUROPIA, p. 199-207

Kautz H., Selman B., Shah M. 1997. *Referral Web: combining social networks and collaborative filtering*. *Commun. ACM* 40, 3 (March 1997).

Massa P., Avesani P. *Trust-aware recommender systems*. In *Proceedings of the 2007 ACM conference on Recommender systems* (RecSys '07). ACM, New York, NY, USA, 17-24.

Newman, M. E. J., *Modularity and community structure in networks*, Proc. Natl. Acad. Sci. USA 103, 8577-8582 (2006).

Ren X., Zeng Y., Qin Y., Zhong N., Huang Z., Wang Y., Wang C., *Social Relation Based Search Refinement: Let Your Friends Help You!*, International *Conferences* on Active Media Technology, AMT 2010: 475-485, 2010.

Salton G., Waldstein R. K. : "*Term relevance weights in on-line information retrieval*". Inf. Process. Manage. 14(1): 29-35, 1978.

Su X. and Taghi M. K. *A survey of collaborative filtering techniques. Adv*. in Artif. Intell. 2009, Article 4 (January 2009), 19 pages. DOI=10.1155/2009/421425.

Tchuente D., Canut M.F, Jessel N., Péninou A., Haddadi A. E., *Visualizing the evolution of users' profiles from online social networks*. International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2010), Odense, Danemark, 09/08/2010-11/08/2010, IEEE Computer Society, p. 370-374, août 2010.

Tchuente D., Jessel N., Péninou A., Canut M.F, Sèdes F. *A community based algorithm for deriving users' profiles from egocentrics networks.* IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), Istanbul, 26/08/2012-29/08/2012, IEEE Computer Society, 2012.

Tchuente D., Jessel N., Péninou A., Canut M.F, Sèdes F. *Visualizing the relevance of social ties in user profile modeling*. Dans : Web Intelligence and Agent Systems, An International Journal, IOS Press, Vol 10, N°2, Pages 261-274, mai 2012.

Viviani M., Bennani N., Egyed-Zsigmond E.. *A Survey on User Modeling in Multi-Application Environments*, The Third International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services CENTRIC'10, Nice, France. pp. 111-116. IEEE . ISBN 978-1-4244-7778-4. 2010.