
# THÈSE

**En vue de l'obtention du**

## DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

**Délivré par :**

Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)

---

**Présentée et soutenue par :**
**Dieudonné TCHUENTE**
**le** lundi 28 janvier 2013

**Titre :**

Modélisation et dérivation de profils utilisateurs à partir de réseaux sociaux : approche à partir de communautés de réseaux k-égocentriques

---

**École doctorale et discipline ou spécialité :**
ED MITT : Image, Information, Hypermedia

**Unité de recherche :**
Institut de Recherche en Informatique de Toulouse (IRIT)

**Directeur(s) de Thèse :**

Florence SEDES, Professeur, Université de Toulouse III Paul Sabatier
Nadine JESSEL, Maître de conférences (HDR), IUFM - Université de Toulouse II Le Mirail

**Jury :**

Rokia MISSAOUI, Professeur, Université du Québec en Outaouais, Rapporteur
Amel BOUZEGHOUB, Professeur, Télécom SudParis, Rapporteur
Christine LARGERON, Professeur, Université Jean Monnet Saint-Etienne, Examinatrice
Laure Pauline FOTSO, Professeur, Université de Dschang-Cameroun, Examinatrice
Claude CHRISMENT, Professeur, Université de Toulouse III Paul Sabatier, Président
Thomas STENGER, Maître de conférences, IAE - Université de Poitiers, Invité
Marie-Francoise CANUT, Maître de conférences, Université de Toulouse II Le Mirail, Invitée
André PENINOU, Maître de conférences, Université de Toulouse II Le Mirail, Invité

Dieudonné TCHUENTE

# Modélisation et dérivation de profils utilisateurs à partir de réseaux sociaux : approche à partir de communautés de réseaux k-égocentriques

*Directrice de thèse :* Florence SEDES, Professeur à l'Université de Toulouse III - IRIT- Toulouse, France

*Co-directrice :* Nadine JESSEL, Maître de conférences (HDR) à l'IUFM de Toulouse - IRIT- Toulouse, France

## Résumé

Dans la plupart des systèmes nécessitant la modélisation de l'utilisateur pour adapter l'information à ses besoins spécifiques, l'utilisateur est représenté avec un profil généralement composé de ses centres d'intérêts. Les centres d'intérêts de l'utilisateur sont construits et enrichis au fil du temps à partir de ses interactions avec le système. De par cette nature évolutive des centres d'intérêts de l'utilisateur, le profil de l'utilisateur ne peut en aucun moment être considéré comme entièrement connu par un système. Cette connaissance partielle du profil de l'utilisateur à tout instant *t* a pour effet de réduire considérablement les performances des mécanismes d'adaptation de l'information à l'utilisateur lorsque le profil de l'utilisateur ne contient pas (ou contient très peu) les informations nécessaires à leur fonctionnement. Cet inconvénient est particulièrement plus récurrent chez les nouveaux utilisateurs d'un système (*instant t=0, problème du démarrage à froid*) et chez les utilisateurs peu actifs. Pour répondre à cette problématique, plusieurs travaux ont exploré des sources de données autres que celles produites par l'utilisateur dans le système : utilisateurs au comportement similaire (utilisé dans le *filtrage collaboratif*) ou données produites par l'utilisateur dans d'autres systèmes (conception de *profil utilisateur multi-application* et *gestion des identités multiples* des utilisateurs).

Très récemment, avec l'avènement du Web social et l'explosion des réseaux sociaux en ligne, ces derniers sont de plus en plus étudiés comme source externe de données pouvant servir à l'enrichissement du profil de l'utilisateur. Ceci a donné naissance à de nouveaux mécanismes de filtrage social de l'information : systèmes de recherche d'information sociale, systèmes de recommandation sociaux, etc. Les travaux actuels portant sur les mécanismes de filtrage social de l'information démontrent que ce nouveau champ de recherche est très prometteur. Une étude sur les travaux existants nous permet tout de même de noter particulièrement deux faiblesses : d'une part, chacune des approches proposées dans ces travaux reste très spécifique à son domaine d'application (et au mécanisme associé), et d'autre part, ces approches exploitent de manière unilatérale les profils des individus autour de l'utilisateur dans le réseau social.

Pour pallier ces deux faiblesses, nos travaux de recherche proposent une démarche méthodique permettant de définir d'une part un *modèle social générique de profil de l'utilisateur* réutilisable dans plusieurs domaines d'application et par différents mécanismes de filtrage social de l'information, et à proposer d'autre part, une technique permettant de dériver de manière optimale des informations du profil de l'utilisateur à partir de son réseau social. Nous nous appuyons sur des travaux existants en sciences sociales pour proposer une *approche d'usage des communautés (plutôt que des individus) autour de l'utilisateur*. La portion significative de son réseau social est constituée des individus situés à une distance maximum k de l'utilisateur et des relations entre ces individus (*réseau k-égocentrique*). A partir de deux évaluations de l'approche proposée, l'une dans le réseau social numérique Facebook, et l'autre dans le réseau de co-auteurs DBLP, nous avons pu démontrer la pertinence de notre approche par rapport aux approches existantes ainsi que l'impact de mesures telles que la centralité de communautés (degré ou proximité par exemple) ou la densité des réseaux k-égocentriques sur la qualité des résultats obtenus. Notre approche ouvre de nombreuses perspectives aux travaux s'intéressant au filtrage social de l'information dans de multiples domaines d'application aussi bien sur le Web (personnalisation de moteurs de recherche, systèmes de recommandation dans le e-commerce, systèmes adaptatifs dans les environnements e-Learning, etc.) que dans les intranets d'entreprise (systèmes d'analyses comportementales dans les réseaux d'abonnés de clients télécoms, détection de comportements anormaux/frauduleux dans les réseaux de clients bancaires, etc.).

Dieudonné TCHUENTE

**Modeling and deriving users' profiles from social networks: a community based approach from k-egocentric networks**

*Supervisor :* Florence SEDES, Professor at University of Toulouse III - IRIT- Toulouse, France

*Co-supervisor :* Nadine JESSEL, Associate Professor (HDR) at IUFM of Toulouse - IRIT- Toulouse, France

# Abstract

In most systems that require user modeling to adapt information to each user's specific need, a user is usually represented by a user profile in the form of his interests. These interests are learnt and enriched over time from users interactions with the system. By the evolving nature of user's interests, the user's profile can never be considered fully known by a system. This partial knowledge of the user profile at any time *t* significantly reduces the performance of adaptive systems, when the user's profile contains no or only some information. This drawback is particularly most recurrent for new users in a system (*time t = 0, also called cold start problem*) and for less active users. To address this problem, several studies have explored data sources other than those produced by the user in the system: activities of users with similar behavior (e.g. *collaborative filtering* techniques) or data generated by the user in other systems (e.g., *multi-application user's profiles*, *multiple identities management* systems).

By the recent advent of Social Web and the explosion of online social networks sites, social networks are more and more studied as an external data source that can be used to enrich users' profiles. This has led to the emergence of new social information filtering techniques (e.g. social information retrieval, social recommender systems). Current studies on social information filtering show that this new research field is very promising. However, much remains to be done to complement and enhance these studies. We particularly address two drawbacks: (i) each existing social information filtering approach is specific in its field scope (and associated mechanisms), (ii) these approaches unilaterally use profiles of individuals around the user in the social network to improve traditional information filtering systems.

To overcome these drawbacks in this thesis, we aim at defining *a generic social model of users' profiles* that can be reusable in many application domains and for several social information filtering mechanisms, and proposing optimal techniques for enriching user's profile from the user's social network. We rely on existing studies in social sciences *to propose a communities (rather than individuals) based approach* for using individuals around the user in a specific part of his social network, to derive his *social profile* (profile that contains user's interest derived from his social network). The significant part of the user's social network used in our studies is composed of individuals located at a maximum distance k (in the entire social network) from the user, and relationships between these individuals (*k-egocentric network*). Two evaluations of the proposed approach based on communities in k-egocentric networks have been conducted in the online social network *Facebook* and the co-authors network *DBLP*. They allow us to demonstrate the relevance of the proposal with respect to existing individual based approaches, and the impact of structural measures such as the centrality of communities (degree or proximity) or user's k-egocentric network density, on the quality of results. Our approach opens up many opportunities for future studies in social information filtering and many application domains as well as on the Web (e.g. personalization of search engines, recommender systems in e-commerce, adaptive systems in e-Learning environment) or in Intranets business systems (e.g. behavioral analysis in networks of subscribers telecom customers, detection of abnormal behavior network bank customers, etc.).

Dieudonné TCHUENTE

**Modélisation et dérivation de profils utilisateurs à partir de réseaux sociaux : une approche à partir de communautés de réseaux k-égocentriques**

*Directrice de thèse :* Florence SEDES, Professeur à l'Université de Toulouse III - IRIT- Toulouse, France

*Co-directrice :* Nadine JESSEL, Maître de conférences (HDR) à l'IUFM de Toulouse - IRIT- Toulouse, France

# Mots-clés

- Modélisation utilisateur, réseaux sociaux, réseaux égocentriques, réseaux k-égocentriques, adaptation, profils utilisateurs, centres d'intérêts utilisateurs, filtrage social, analyse de communautés.

- Fouille de graphes, fouille de données, fouille de textes, filtrage d'information, filtrage collaboratif.

- Systèmes d'information, Web 2.0, Web social, sociologie.

- Facebook, DBLP, Mendeley.

Dieudonné TCHUENTE

**Modeling and deriving users' profiles from social networks: a community based approach from k-egocentric networks**

*Supervisor :* Florence SEDES, Professor at University of Toulouse III - IRIT- Toulouse, France

*Co-supervisor :* Nadine JESSEL, Associate Professor (HDR) at IUFM of Toulouse - IRIT- Toulouse, France

# Key-Words

- User modeling, social networks, egocentric networks, k-egocentric networks, adaptation, users' profiles, users' interests, social filtering, community analysis.

- Graph mining, data mining, text mining, information filtering, collaborative filtering.

- Information Systems, Web 2.0, social Web, sociology.

- Facebook, DBLP, Mendeley.

*Je dédie cette thèse*

*à l'âme de mon père,*

*à l'âme de ma grande mère,*

*à ma mère,*

*à Maman Fonga Jeanette,*

*à Maman Kom Véronique,*

*à Papa Kom Lucien,*

*à Danielle,*

*à Laure,*

*à Willy,*

*à Faustin,*

*et*

*à Francis.*

# Remerciements

*Je tiens tout d'abord à remercier Madame Nadine Jessel, co-directrice de cette thèse pour avoir initié ce travail de recherche, ainsi que pour son encadrement dans une atmosphère toujours conviviale et propice au bon déroulement de ce travail.*

*Je remercie Madame Florence Sèdes, directrice de cette thèse, qui n'a ménagé aucun effort pour s'assurer d'un suivi très pointilleux de l'avancée de mes travaux tout au long de ces années, ainsi que pour tous ses conseils qui m'ont permis d'avoir une vision plus large de mon travail, et pour son soutien à la réalisation de séjours à l'extérieur.*

*Je tiens à exprimer ma profonde gratitude à Madame Marie-Françoise Canut pour son encadrement, ses encouragements et sa grande disponibilité à mon égard depuis mon arrivée à l'IRIT en provenance du Cameroun, jusqu'à ce jour.*

*Je tiens également à exprimer ma profonde gratitude à Monsieur André Péninou, pour sa clairvoyance et pour le temps accordé aux vérifications et au bon avancement de mes travaux.*

*Je vous remercie collectivement, Mesdames Florence Sèdes, Nadine Jessel, Marie-Françoise Canut et Monsieur André Péninou, pour l'énergie consacrée tant à mon épanouissement au niveau professionnel qu'au niveau familial. L'ambiance chaleureuse de travail a été pour moi d'une très grande motivation pour la réalisation de cette thèse.*

*Je remercie Mesdames Rokia Missaoui, Professeur à l'Université du Québec en Outaouais, et Amel Bouzeghoub, Professeur à Telecom SudParis, Laure Pauline Fotso, Professeur à l'Université de Dschang-Cameroun, Chritine Largeron, Professeur à l'Université Jean Monnet de Saint-Etienne, d'avoir bien voulu consacrer du temps pour rapporter et examiner ce travail de thèse.*

*Je remercie le Professeur Claude Chrisment, et le Professeur Josiane Mothe successivement responsables de l'équipe SIG (Systèmes d'Information Généralisés) pendant la durée de cette thèse, pour m'avoir accueilli au sein de cette équipe de recherche, et pour tous les efforts qu'ils consacrent au bon déroulement des travaux des doctorants.*

*Je remercie le Professeur Michel Daydé, directeur de l'IRIT, et le professeur Daniel Hagimont de l'ENSEEIHT pour la confiance et l'opportunité qu'ils m'ont donnée d'effectuer mon stage de master à l'IRIT (en provenance du Cameroun), stage qui s'est poursuivi par la réalisation de cette thèse.*

*Je remercie Messieurs Thomas Stenger, Alexandre Coutant, et Olivier Rampnoux pour leur très enrichissante collaboration multidisciplinaire sur les projets de R&D avec le groupe La Poste, qui m'a permis d'entreprendre plusieurs choix et approches dans cette thèse.*

*Je remercie le Professeur Hideaki Takeda, directeur du centre R&D des ressources scientifiques du laboratoire NII (National Institute of Informatics) du Japon, pour m'avoir accueilli dans son équipe de recherche le temps d'un stage, et m'avoir prodigué de nombreux conseils qui m'ont permis de réaliser une partie des expérimentations de cette thèse.*

*Je remercie le Professeur Bernard Dousset pour tous ses encouragements durant cette thèse et pour tout le temps mis à ma disposition pour la compréhension de l'outil Tétralogie.*

*Je remercie Messieurs Guillaume Cabanac et Sébastien Laborie pour toutes les discussions enrichissantes qu'on a entretenues à mon arrivée à l'IRIT, ainsi qu'à certaines périodes du déroulement de cette thèse.*

*Je remercie toute l'équipe administrative de l'IRIT pour son accueil, sa réactivité et son efficacité. En ce sens, je remercie particulièrement Mesdames Chantal Morand, Jean Pierre Baritaud, et Agathe Baritaud.*

*Je tiens à renouveler mes amitiés vers les actuels et anciens doctorants de l'équipe SIG ou de l'IRIT, en particulier à Anass El Haddadi, Grégory Claude, Ana-Maria Manzat, Reda Jourani, Houssem Jerbi, Bachelin Ralalason, Dana Al Kukhun, Dana Codreanu, Madalina Mitran, Faten Atigui, Lamjed Ben Jabeur, Hamdi Chaker, Damien Dudognon, Firas Damak, Laure Soulier, Léa Laporte, Imen Megdiche…*

*Je remercie tous les étudiants de l'IUT de Blagnac qui se sont portés volontaires de participer à une des évaluations de cette thèse au travers de leur profil Facebook.*

*Je tiens à remercier mes camarades et amis de longue date pour leur confiance et leur présence quasi permanente dans ma vie : Noubissi Hervé, Merlain Fonguieng, Salomon Mahama, Thomas Ngueoko, Thierry Yatchoupou, Adrien Nouwomo Mangwa, Martial Guetcho, Mireille Nono, Jean Sans Terre Wankap, Ghislain Temgoua, William Mangoua, Raymond Essomba, …*

*Je remercie mes camarades et amis de l'université de Yaoundé I , de l'Université de Dschang et de Toulouse qui m'ont apporté leur soutien et encouragements lors de ma soutenance de thèse : Larissa Mayap, Rodrigue Chakode, Désiré Nuentsa, Lassina Camara, Franck Tagne, Hubert Zouatcham, Hugues Tameze, Fabrice Kouam, Henry Valery Teguiak, Raphael Ndzana, Armelle Ngoufack, Alvine Silatchom, Noeline Tsafack.*

*De tendres remerciements pour mes très chères Laure et Danielle qui ont dû supporter des éloignements parfois très fréquents pendant le déroulement de cette thèse. Vous avez toutefois*

*été toujours présentes  et sources de motivation supplémentaire pour moi. De la même manière, je remercie du fond du cœur ma maman Nguetchueng Agnès, et mes frères Kamseu William Yves, Awunti Nguetchueng Faustin, et Tagne Francis pour leur exemplarité.*

*Enfin, je ne saurais terminer sans remercier ceux qui ont toujours été au plus près de moi en veillant à ma réussite : mes parents, mes frères et sœurs au Cameroun et en France. Je remercie en particulier Maman Fonga Jeannette, Maman Kom Véronique, Papa Kom Lucien, Laure et Engelbert Mephu Nguifo.*

# Table des matières

# 1 Introduction générale



## 1.1 Contexte du travail

Depuis sa création, la croissance du volume d'informations numériques disponible sur le Web augmente de manière exponentielle. De nos jours plusieurs facteurs contribuent au renforcement de cette croissance : l'avènement du Web centré utilisateur (Web 2.0) suivant un modèle décentralisé de mise en ligne de contenus, les usages de plus en plus importants des interfaces mobiles (téléphones, Smartphones, tablettes, etc.), l'émergence du *cloud computing* avec le stockage de masses importantes de données (*big data*), etc. S'il devient plus facile de publier des contenus sur la toile, l'accès à ces contenus est cependant rendu plus difficile aux utilisateurs compte tenu du nombre de plus en plus important et de la diversité des informations susceptibles de les intéresser. Ceci pose en général des problèmes de surcharge cognitive à l'utilisateur qui aura de plus en plus du mal à retrouver l'information correspondant à ses attentes. Pour répondre à ces problèmes d'accès à l'information, des systèmes d'adaptation de l'information à l'utilisateur ont été proposés. Ces systèmes mettent en œuvre des mécanismes permettant de renvoyer ou de présenter à l'utilisateur des informations correspondant à ces besoins spécifiques : systèmes de recherche d'information personnalisée, système de recommandations d'informations, systèmes d'information adaptatifs,… . De très nombreux domaines d'applications sont concernés par ces travaux : les moteurs de recherche (Google, Yahoo, etc.), les librairies digitales (Di Giacomo et al., 01), le e-commerce (Money et al., 00), le e-learning (Conlan et al., 02), les services Web (Soukarrieh, 10), etc. Depuis les années 2000, l'intérêt pour ces systèmes est en forte croissance aussi bien sur le Web que dans les systèmes d'information d'entreprise comme en témoigne la forte progression des publications scientifiques (IEEE, ACM, Springer, Science Direct) relatives à la personnalisation et à la recommandation (Gao et al., 10) (figure 1.1).

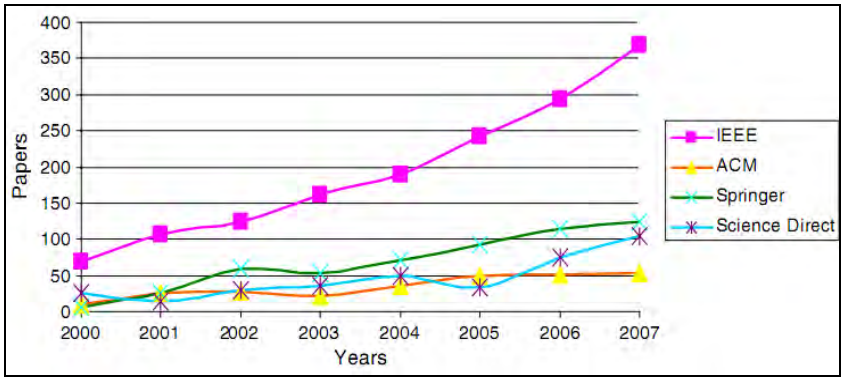



Figure 1. 1: progression des publications scientifiques liées à la personnalisation et à la recommandation de l'information

Chaque type de système d'adaptation de l'information utilise des mécanismes qui lui sont propres. Cependant, tous ces systèmes ont un point commun: l'utilisation d'un modèle de l'utilisateur. La modélisation de l'utilisateur vise à construire le profil de l'utilisateur qui sera utilisé dans le processus d'adaptation. La qualité des résultats des systèmes adaptatifs est en très grande partie dépendante de la qualité des profils utilisateurs construits, et dans une moindre mesure dépendante des techniques de *matching* entre les profils construits et ressources à adapter (documents, pages Web, produits, …). Le profil de l'utilisateur peut être renseigné par ce dernier (*profil explicite*) ou construit de manière automatique à partir de données issues des interactions entre ce dernier et un système (*profil implicite*). Cette dernière démarche est la plus courante et elle s'inscrit dans un cadre similaire au contexte plus général de processus d'extraction de connaissances à partir de données (Fayyad et al., 96). Au-delà des systèmes d'adaptation de l'information à l'utilisateur, la modélisation de l'utilisateur dans les systèmes d'information est également au centre des techniques d'analyses comportementales des utilisateurs pour la détection des comportements à risques ou pour faciliter la prise de décision : détection de fraudes et *scoring*[1] (banques, assurances, etc.), détection de réseaux terroristes (bases de données dédiées, Web, etc.), détection de pédophiles (Web), détection de leaders (Web, intranets d'entreprises, etc.), etc. La place centrale qu'occupe la modélisation de l'utilisateur dans les systèmes d'adaptation de l'information et les systèmes d'analyses comportementales impose d'y accorder un grand intérêt.

C'est dans ce contexte que se situent nos travaux. Ils s'inscrivent dans la continuité des travaux de notre équipe de recherche sur les systèmes d'adaptation de l'information : adaptation de documents semi-structurés (Zayani, 08), adaptation des interfaces de présentation de l'information (Encelle, 07), adaptation des services Web (Bouchra, 10), adaptation de documents multimédias (Laborie et al., 09). Tous ces systèmes utilisent des profils utilisateurs qui sont comparables. Orientés vers l'adaptation de l'information, ces travaux ont abordé la construction du profil utilisateur, soit de manière ad hoc, soit de manière insatisfaisante pour répondre aux besoins spécifiques de chaque utilisateur à tout instant. Pour pallier ces insuffisances, la construction des profils utilisateurs pour les systèmes d'adaptation de l'information fait l'objet de nos travaux.

## 1.2 Problématique

Le développement des profils utilisateurs suit un processus (similaire au processus d'extraction et d'analyse de données *KDD*[2]) par nature incrémental dans la mesure où le profil de l'utilisateur est construit et enrichi au fur et à mesure de ses interactions avec le système d'information. Ceci peut poser deux problèmes majeurs pour les mécanismes réutilisant les informations de ces profils :

- ***Comment gérer le cas où le profil de l'utilisateur ne comporte pas toutes les informations nécessaires au processus d'adaptation ?*** Ceci peut se décliner en deux problèmes :
    - *Comment gérer un nouvel utilisateur du SI qui n'a pas encore interagit ou qui interagi très peu avec le système ?* ce problème est connu dans la littérature comme celui du démarrage à froid (*cold*

[1] *En marketing, le scoring est une méthode qui consiste à affecter une note (un « score ») à chaque client ou prospect d'une base de données afin de cibler et prospecter avec une meilleure efficacité.*

[2] *Knowledge Discovery in Databases*

*start problem*). Le profil de l'utilisateur étant vide, aucune adaptation ne peut être réalisée pour l'utilisateur. Ce problème est très répandu dans les systèmes de recommandation basé sur le filtrage collaboratif[3] comme dans le e-commerce (Massa et Avesani, 04) (Massa et Avesani, 07) (Massa et Avesani, 09), et peut être généralisé pour tous les systèmes d'adaptation de l'information ou d'analyses comportementales (Zayani, 08).

- *Comment gérer les cas où le profil de l'utilisateur ne contient pas d'information utile pour un mécanisme à un instant donné ?* par définition, comme le profil de l'utilisateur est enrichi au fur et à mesure de ces interactions, il est par conséquent toujours incomplet. A supposer par exemple qu'un profil contient des informations sur les centres d'intérêts de l'utilisateur uniquement sur deux domaines nommés D1 et D2, que faire si un mécanisme a besoin des centres d'intérêts de cet utilisateur par rapport à un domaine D3 (non connu actuellement dans le profil de l'utilisateur) ?

- ***Où trouver l'information manquante au profil utilisateur ?*** problème corrélé à la résolution du précédent.

# 1.3 Contributions

Les contributions de ce mémoire visent à apporter une solution aux problèmes décrits précédemment. L'idée principale repose sur l'hypothèse que l'exploitation adéquate du ***réseau social***[4] de l'utilisateur permettra de dériver des centres d'intérêts non connus sur ce dernier. Des travaux issus des sciences sociales ou des expérimentations récentes en informatique démontrent que le réseau social d'un utilisateur est assez caractéristique de sa personne. Toutefois, reste encore à chercher le moyen le plus efficace permettant d'exploiter un réseau social pour l'enrichissement d'un profil utilisateur. Notre contribution dans le cadre de cette recherche se décline à deux niveaux :

- ***La proposition d'un modèle générique de profil utilisateur intégrant de manière significative son réseau social.*** Ce modèle bidimensionnel (***dimension utilisateur*** et ***dimension sociale***) d'un profil utilisateur s'appuie sur la définition d'une portion significative (***réseau k-égocentrique*** ou graphe de relations entre individus situés à distance k de l'utilisateur dans le réseau social entier) du réseau social d'un utilisateur. Ce modèle reste générique et chacune des dimensions qu'il contient peut être exploitée de manière flexible par tout mécanisme (personnalisation, recommandation, etc.).

- ***La proposition d'algorithmes permettant d'exploiter le réseau social et la sémantique des données inclus dans le modèle précédent pour dériver de manière efficace la dimension sociale du profil de l'utilisateur.*** A la différence des algorithmes existants qui s'appuient sur les individus du réseau social de l'utilisateur (approche que nous qualifions d'*autoritaire*), nous proposons un algorithme s'appuyant sur des communautés extraites du réseau k-égocentrique de l'utilisateur (approche que nous qualifions d'*affinitaire*). L'élaboration de cet algorithme s'appuie sur le constat simple déjà observé en sociologie (Goffman, 59) selon lequel chaque communauté (construite uniquement via

[3] *Technique visant à enrichir le profil d'un utilisateur à partir des utilisateurs qui lui sont similaires (voir chapitre 1).*

[4] *Le réseau social considéré ici est celui constitué de liens qu'entretiennent les utilisateurs dans la vie réelle.*

les liens sociaux) d'individus autour d'une personne dénote une certaine affinité entre ces individus, et cette affinité se rapporte très probablement à cette personne.

A la différence des travaux de la littérature sur l'adaptation de l'information à l'utilisateur, nous adoptons un procédé de validation de profils construits plutôt qu'un procédé de validation des mécanismes associés à ces profils. Ce procédé nous permet de comparer la pertinence de différentes approches de construction de la dimension sociale du profil afin d'en déduire la plus efficace. Ceci a pour but ultime de s'assurer de la qualité des profils ou des algorithmes à utiliser en amont des mécanismes, et de ce fait, d'anticiper la qualité des résultats de ces mécanismes.

Les expérimentations de validation sont réalisées dans deux contextes différents : les réseaux sociaux numériques (Facebook dans notre cas) et les réseaux d'auteurs et de co-auteurs d'articles scientifiques (DBLP[5] et Mendeley dans notre cas). Elles démontrent toutes les deux l'efficacité de notre approche de construction de profils s'appuyant sur les communautés extraites des réseaux k-égocentriques des utilisateurs, comparée aux principales méthodes utilisées dans la littérature.

## 1.4 Organisation du mémoire

Ce mémoire se décline suivant quatre chapitres.

Le premier chapitre présente le contexte général de nos travaux : la modélisation de l'utilisateur et son exploitation dans les systèmes d'adaptation de l'information à l'utilisateur (personnalisation et recommandation de l'information en particulier). Les travaux de la littérature sont présentés de manière structurée suivant les étapes d'un processus classique d'extraction de connaissance à partir de données. Pour chacune des étapes, les concepts clés sur lesquels il est nécessaire de se focaliser lors de la construction, la représentation et l'exploitation des profils utilisateurs sont identifiés et présentés. Ceci permet d'avoir une traçabilité plus claire du processus de développement des profils ainsi que l'identification des concepts potentiellement utiles pour notre contribution.

Le second chapitre est décomposé en deux parties. La première partie présente les travaux de la littérature liés aux problématiques de démarrage à froid et d'enrichissement des profils utilisateurs dans les systèmes d'adaptation de l'information à l'utilisateur qui s'appuient sur les réseaux sociaux. Compte tenu des limites de ces travaux, la seconde partie présente les éléments clés de l'analyse de réseaux sociaux susceptibles d'être exploités pour pallier ces limites.

Le troisième chapitre présente notre contribution suivant trois parties également. Dans un premier temps nous présentons le modèle générique de profils utilisateur intégrant les réseaux sociaux k-égocentrés. Chacun des concepts de ce modèle y est clairement présenté et justifié. Dans un second temps nous présentons différents algorithmes exploitant ce modèle pour dériver les centres d'intérêt de la dimension sociale du profil de l'utilisateur en se focalisant sur l'algorithme basé sur les communautés du réseau k-égocentrique. Les stratégies de validation des profils suivant les algorithmes proposés sont présentés en troisième partie de ce chapitre.

[5] *Digital Bibliography & Library Project*

Le quatrième chapitre détaille chacune des expérimentations réalisées sur *Facebook* et dans *DBLP*. Pour chacune des expérimentations, il est présenté l'accessibilité aux données, la construction des réseaux k-égocentriques, le processus de construction des dimensions des profils, la présentation et la discussion des résultats obtenus par comparaison de l'approche proposée avec des approches similaires dans l'état de l'art.

Enfin nous discutons en fin de ce mémoire des implications que peuvent avoir les propositions présentées dans l'amélioration des mécanismes associés dans les systèmes d'information, ainsi que des nombreuses pistes de recherche futures.

# 2 Chapitre 1 : Développement, représentation, et usage des profils utilisateurs



## 2.1 Introduction

Le contexte de notre travail est celui de la modélisation de l'utilisateur (*user modeling* en anglais) dans un système d'information. La modélisation de l'utilisateur fait partie du domaine des interactions homme-machine et consiste à étudier le processus d'extraction et d'analyse des interactions entre l'utilisateur et le SI pour construire des profils utilisateurs. La modélisation de l'utilisateur peut alors être perçue comme un cas particulier de processus d'extraction de connaissances à partir de données (*KDD* en anglais), lorsque celles-ci sont issues des interactions entre les utilisateurs et les SI. D'un point de vue du domaine de gestion des connaissances, les profils utilisateurs construits par ce processus peuvent être perçus comme des données, des informations ou des connaissances en fonction des usages qui en sont faits (Figure 2.1). Dans les systèmes de gestion de connaissances, une donnée (ou donnée brute) est un élément descriptif sur lequel aucun traitement (aucune analyse) n'a été effectué (exemple d'un champ de saisie d'un formulaire). Une information est regroupement ou une organisation de données ayant subi un traitement mécanique (traitement informatique ou manuel) sans interprétation humaine (exemple d'un ensemble de données d'un formulaire saisi par un

utilisateur). Une connaissance est une information qui a subi une interprétation humaine (par exemple, un algorithme identifiant des groupes de consommateurs ne générera que de l'information, c'est l'interprétation humaine de ces groupes obtenus, l'intérêt et les applications utiles qui en découlent qui en font une connaissance).

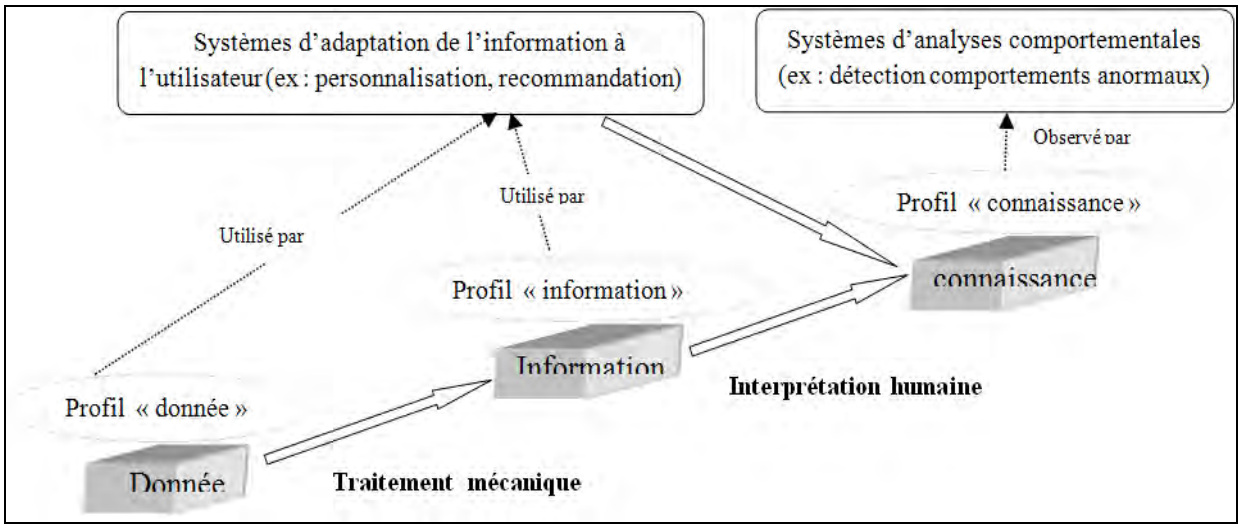



Figure 2. 1 : profils utilisateurs et gestion des connaissances

Ainsi, un profil utilisateur peut être perçu comme une donnée ou une information lorsqu'il est utilisé pour les systèmes désirant répondre à ses besoins spécifiques (adaptation de l'information à l'utilisateur) : personnalisation (en recherche d'information par exemple), ou recommandation (dans le e-commerce par exemple) par exemple. Par contre un profil utilisateur peut également être perçu comme une connaissance dans les cas des systèmes d'analyses comportementales : détection de fraudes (banques, assurances, etc.), détection de réseaux terroristes (bases de données dédiées, Web, etc.), détection de pédophiles (Web), détection de leaders (Web, intranets d'entreprises, etc.), etc.

Les paragraphes précédents donnent une définition des profils utilisateurs du point de vue de la gestion des connaissances qui permet de mieux présenter les mécanismes fonctionnels s'appuyant sur ces profils. Ceci permet de comprendre l'utilité (« *le pourquoi* ») ou la place centrale qu'occupe la modélisation de l'utilisateur dans les SI. Dans la suite de ce chapitre nous allons plutôt présenter les techniques relatives à la modélisation utilisateur dans les travaux de la littérature (« *le comment* »). Dans la mesure où la modélisation de l'utilisateur dans les SI est très proche des processus classiques d'extraction de connaissances à partir de données, nous structurons cet état de l'art suivant les différentes étapes de ce processus (Figure 2.2).

Ce processus comprend cinq grandes phases (Fayyad et al., 96) qui, dans le contexte de la modélisation utilisateur, peuvent être décrites de la manière suivante :

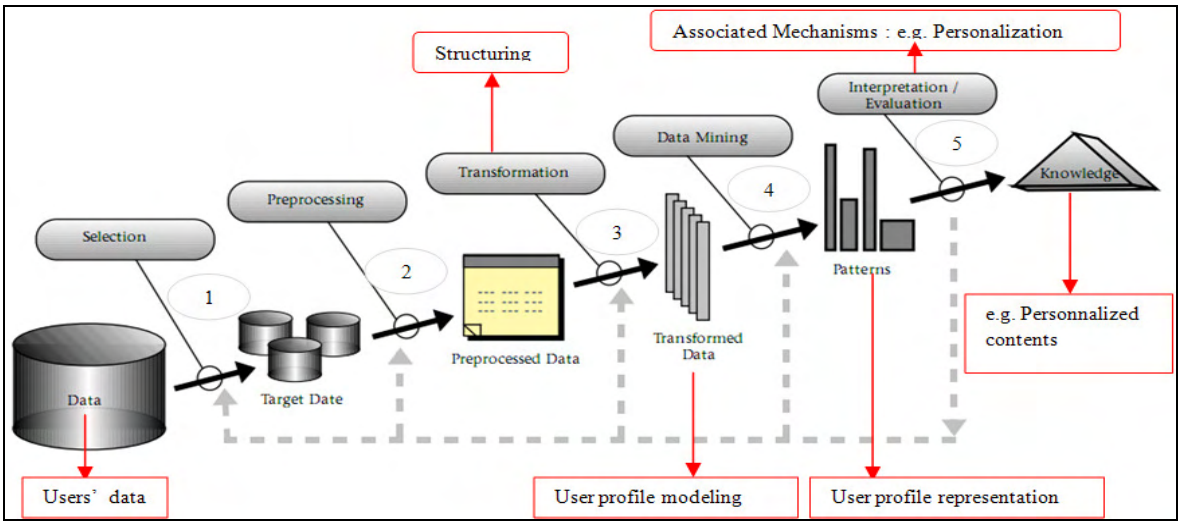



Figure 2. 2: Processus de développement de profils et d'usage des profils utilisateurs dérivé du processus classique d'extraction de connaissance à partir de données de Fayyad et al., 96 (les éléments en rouge sont les éléments spécifiques au contexte de la modélisation utilisateur).

1- **La sélection des données (*selection*)**. Toutes les données brutes que l'on possède ne sont pas nécessairement utiles pour générer de la connaissance. Cette étape filtre les données jugées pertinentes, les données cibles que l'on va réellement traiter. Dans le contexte de la modélisation utilisateur, les données brutes traitées sont des données issues des interactions entre les utilisateurs et le SI (*users data*).

2- **Le prétraitement des données (*preprocessing*)**. Une fois les données sélectionnées, des traitements mineurs leur sont appliqués pour éliminer le bruit ou gérer les valeurs manquantes par exemple. Cette étape de nettoyage de données fournit des données prétraitées pour améliorer la qualité des résultats obtenus par les étapes suivantes. Cette étape reste pratiquement la même que l'on soit dans des contextes classiques ou dans le cas particulier de la modélisation utilisateur.

3- **La transformation des données (*transformation*)**. Cette étape est entièrement dépendante de l'étape suivante (fouille de données). L'objectif est d'organiser les données prétraitées suivant une structure qui soit adaptée aux algorithmes de fouille de données (algorithme de classification par exemple). Dans le contexte de la modélisation utilisateur, il s'agit particulièrement de structurer (*structuring*) les différentes données qui seront utilisées pour la fouille de données suivant un modèle de profil utilisateur (*user profile model*).

4- **La fouille de données (*data mining*)**. Elle consiste à effectuer l'analyse des données et à appliquer des algorithmes d'extraction des connaissances (classification à base de règles, arbres de décisions, etc.) pour produire différents motifs (données ou informations non interprétées) (cf. figure 2.2). La qualité des *patterns* obtenus est directement liée à la qualité des traitements des trois étapes précédentes. Dans le contexte de la modélisation utilisateur, les *patterns* produits sont les éléments des profils utilisateurs qui peuvent être représentés de plusieurs manières (*user profile representation*).

5- **L'interprétation et l'évaluation (*interpretation / evaluation*)**. Cette étape dépend de la nature des éléments du profil (figure 2.1). Les éléments du profil peuvent être perçus comme des données ou information d'une part (s'ils sont utilisés par les systèmes d'adaptation de l'information à l'utilisateur

pour produire de la connaissance interprétée par l'humain), ou comme des connaissances d'autre part (s'ils sont interprétés directement par les humains dans les systèmes d'analyse comportementale). Les évaluations sont réalisées par l'analyse des performances de ces systèmes d'adaptation de l'information à l'utilisateur ou d'analyse comportementale.

La quasi-totalité des travaux de la littérature relatifs à la modélisation utilisateur dans les SI peuvent être classifiés suivant ces cinq étapes. Ces étapes n'ont été décrites que de manière très succincte jusqu'ici. La suite de ce chapitre vise à étudier plus en détail les concepts et techniques utilisés dans les travaux de la littérature sur la modélisation utilisateur suivant deux grandes parties : les travaux relatifs au développement proprement dit des profils utilisateurs (étapes 1 à 4) et les travaux relatifs à l'usage de ces profils dans différents systèmes (étape 5).

Dans cette partie, nous allons présenter des travaux de la littérature relatifs aux quatre premières étapes du processus de développement des profils utilisateurs (de la phase de sélection des données à la phase de représentation des profils). Pour chacune des étapes, nous allons insister sur les principaux concepts et techniques qui sont mis en jeu au travers des travaux existants. Cette partie de l'état de l'art a pour but d'aboutir d'une part à une bonne modélisation et traçabilité du processus de développement des profils (qui est un gage de la qualité des profils construits), et d'autre part à identifier les points saillants sur lesquels on pourrait s'appuyer pour aborder notre problématique de l'enrichissement des profils utilisateurs.

# 2.2 Sélection de données pour les profils utilisateurs

Dans le processus de développement des profils utilisateurs, la phase de sélection de données fait intervenir deux concepts (Figure 2.3) : les producteurs de données (individus qui produisent les données à utiliser pour construire le profil de l'utilisateur) et les sources de données (provenance et accessibilité des données). Ces deux concepts très en lien avec la problématique de cette thèse seront présentés brièvement dans les sous-sections qui suivent, et plus en détail (au niveau des techniques utilisées) dans la deuxième partie de ce chapitre.

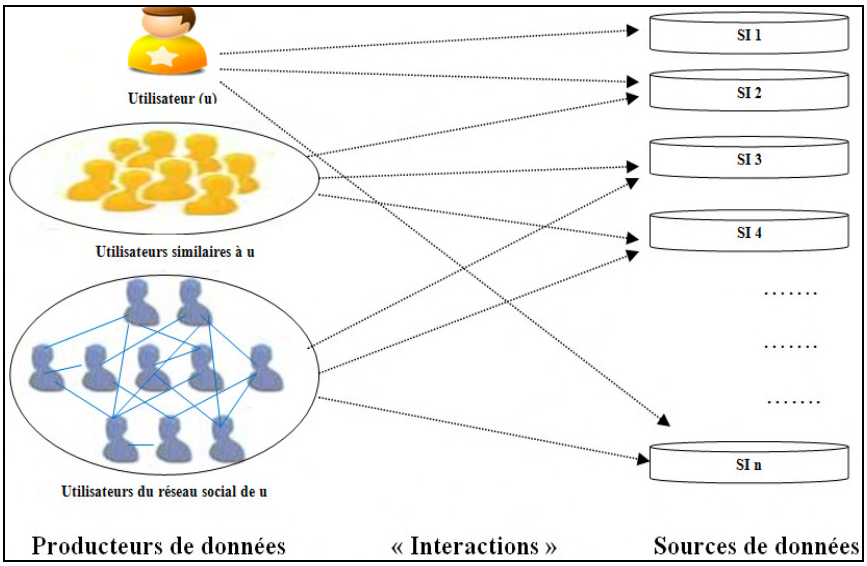



Figure 2. 3: producteurs et sources de données : concepts majeurs lors de la sélection de données

### 2.2.1 Producteurs de données

Dans la littérature, les données sélectionnées pour construire les profils utilisateurs peuvent être produites soit (Figure 2.3) :

- ***par l'utilisateur lui-même :*** c'est naturellement le cas le plus courant dans la littérature. Toutefois, lorsque les systèmes disposent de très peu ou pas de données générées par l'utilisateur (nouvel utilisateur ou utilisateur peu actif), des données produites par d'autres utilisateurs sont particulièrement utilisées (utilisateurs similaires et utilisateurs du réseau social notamment).

- ***par les individus qui lui sont similaires :*** ces données sont généralement utilisées pour prédire (instant $t + \Delta t$) des éléments non connus du profil de l'utilisateur à partir des individus ayant un profil similaire à celui de l'utilisateur à l'instant actuel t. La similarité est généralement calculée par croisement (cosinus de similarité par exemple) (Gao et al., 10) du profil de l'utilisateur à l'instant t avec ceux de tous les autres utilisateurs du système à ce même instant. Ce principe est à la base des différentes techniques de filtrage collaboratif. Ceci est, par exemple, très utilisé dans les sites de e-commerce (cas d'Amazon) utilisant des règles d'association : les utilisateurs qui ont acheté le produit de gamme X, ont également acheté les produits de gamme Y ; Donc intuitivement, pour un utilisateur ayant acheté un produit de gamme X, les produits de gamme Y peuvent être utilisés pour enrichir son profil. Toutefois, ces techniques peuvent devenir très coûteuses en temps de calcul pour de très grands systèmes ayant un nombre très élevé d'utilisateurs, car il serait alors nécessaire de construire, de maintenir et d'exploiter d'énormes matrices creuses (Massa et Avesani, 04) (Massa et Avesani, 07) (Massa et Avesani, 09). De plus, cette technique ne peut être exploitée lorsque le profil de l'utilisateur est vide ou très pauvre à l'instant t (nouvel utilisateur dans le système par exemple), car il serait évidemment impossible de retrouver des individus similaires à ce dernier. Le réseau social de l'utilisateur peut être utilisé comme alternative pour pallier ces inconvénients.

- ***par les individus de son réseau social :*** un réseau social est un graphe de relations entre individus. Les relations entre les individus peuvent être des relations explicites lorsqu'il s'agit des relations créées par les individus eux-mêmes (relations d'amitié dans la vie réelle par exemple) ou des relations implicites lorsqu'elles sont interprétées à partir des activités des individus (relations entre utilisateurs annotant les mêmes ressources par exemple). Dans la mesure où l'entourage social d'un individu peut être très révélateur de sa personne (Goffman, 59)(Granovetter, 73), de plus en plus de travaux s'appuient sur les réseaux sociaux pour dériver des informations pertinentes sur ce dernier (Carmel et al., 09)(Cabanac, 11)(Zeng et al., 09)(Guy et al., 08). Toutefois, de nombreuses pistes restent encore très peu explorées par rapport à l'analyse efficiente des réseaux sociaux pour dériver de l'information utile pour l'utilisateur. Par exemple, comment intégrer de manière efficiente le réseau social dans un modèle utilisateur ? Comment sélectionner de manière efficiente les nœuds du graphe social qui seraient les plus caractéristiques de l'utilisateur ? Comment exploiter de manière efficiente les propriétés liées à la structure du graphe et les propriétés liées aux profils des nœuds ? … Nous tenterons d'apporter des réponses à ces questions (chapitres 3, 4, et 5).

Dans leurs activités, ces producteurs de données interagissent avec un ou plusieurs systèmes d'information qui vont contenir les sources de données à exploiter pour construire les profils utilisateurs.

## 2.2.2 Sources de données

Par sources de données, nous entendons les systèmes (systèmes d'exploitation, logs de bases de données, logs de serveurs Web, etc.) et les applications (email, social bookmarking, etc.) par lesquels les données utilisateurs peuvent être collectées. Deux principales problématiques se dégagent généralement par rapport aux sources de données : leur interopérabilité (du fait de la multiplicité) et leur fiabilité (ou qualité) (Figure 2.4).

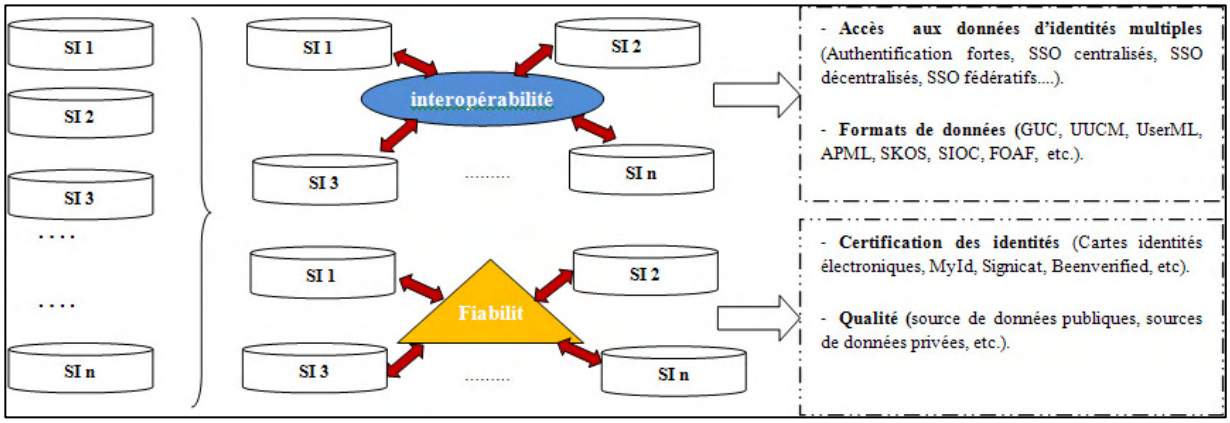



Figure 2. 4 : problématiques liées aux sources de données

### *2.2.2.1 Interopérabilité des sources de données*

De nos jours les utilisateurs disposent de plus en plus d'interfaces qui génèrent des traces issues de leurs activités dans les systèmes d'information. C'est le cas particulier du Web 2.0 avec la multiplication des identités numériques. Ainsi, les utilisateurs disposent de plus en plus de données partagées au sein de diverses applications, et le challenge le plus important consiste à recouper ces données ou à faire inter-opérer ces applications. Dans les travaux et technologies actuels, deux approches permettent d'aborder cette problématique : (i) les systèmes d'identification et d'authentification uniques au travers de diverses applications (*Single Sign On* ou SSO), (ii) les formats standards d'échange de données entre applications.

***(i)* Technologies SSO :** elles permettent d'identifier et d'authentifier aisément un même utilisateur au sein du système d'information. Elles permettent également le partage des données utilisateurs entre ces systèmes. Ces technologies peuvent être regroupées en quatre catégories :

- ***les systèmes à forte authentification* :** ils s'appuient sur des certificats numériques (X509 en général) embarqués dans des supports physiques (clés USB, cartes à puces, etc.). Ils peuvent être délivrés par des gouvernements (exemple des cartes nationales d'identité électroniques en Belgique, Estonie ou Finlande) ou par des organismes publics ou privés (Idénum, Certigreffe, Certinomis, etc.). Ces technologies nécessitent cependant d'énormes infrastructures qui rendent difficile leur déploiement. Ils sont surtout utilisés pour la dématérialisation des services publics ou privés sur le Web. Même si certaines solutions dédiées au Web grand public s'appuient déjà sur ces technologies (certification des comptes E-bay en Belgique par exemple), la forte sensibilité des données utilisateurs dans ce contexte est un frein à son adoption par les internautes.
- ***Les SSO centralisés* :** ce sont des architectures SSO beaucoup plus souples au niveau de leur déploiement, et qui sont donc mieux adaptés pour des applications Web grand public (e-commerce, réseaux sociaux, *social bookmarking*, etc.). Dans ces architectures, une seule entité joue le rôle de fournisseur d'identité (identification, authentification, éléments du profil de

l'utilisateur, etc.). C'est le cas de la plateforme projet *PassPort* de Microsoft ou les APIs de réseaux sociaux numériques de plus en plus en vogue (API facebook, Google OpenSocial, etc.). Même si ces technologies sont très souples à déployer et à utiliser pour le Web grand public, le fait de centraliser toutes les données utilisateur chez un seul fournisseur de données n'inspire pas la confiance des internautes. Ceci peut par exemple expliquer l'échec du projet PassPort de Microsoft qui a évolué vers une architecture de SSO décentralisé (*CardSpace*).

- ***Les SSO décentralisés* :** ils permettent de pallier l'inconvénient des SSO centralisés dans lesquels l'utilisateur confie toutes ses données chez un seul fournisseur. Dans une architecture de SSO décentralisée, l'utilisateur choisit lui-même son ou ses fournisseurs de données d'identité. OpenID[7] est la principale solution qui a été conçue à cet effet. Windows CardSpace peut également être classé dans cette catégorie. Des travaux de recherche au sein des laboratoires Sun et du MIT étudient la possibilité de simplifier encore plus le processus des SSO centralisés via l'usage de vocabulaire du Web social *FOAF (Friend Of A Friend)* par exemple, et du protocole SSL (Story et al., 09). Les SSO décentralisés semblent bien adaptés pour le Web, cependant même si une confiance est établie entre l'utilisateur et son fournisseur de données d'identités (puisque l'utilisateur le choisit lui-même), ces architectures ne gèrent pas la confiance entre un site fournisseur de données d'identité et un site récepteur qui exploite ces données. La confiance (entre site fournisseur et site récepteur) est pourtant très importante dans un contexte d'interopérabilité entre systèmes institutionnels (universités par exemple) ou des collaborations entre intranets d'entreprises.

- ***Les SSO fédératifs* :** dans le but de pallier le problème de gestion de la confiance entre un site récepteur et un site fournisseur de données dans une architecture SSO décentralisée, les SSO fédératifs rajoutent une couche de gestion de confiance entre ces deux derniers. Ces SSO supposent au préalable un accord de partage des données d'identité entre tous les systèmes impliqués (cercles de confiance). Pour mettre en œuvre ces SSO, les étapes de communications ainsi que les échanges de messages et surtout des assertions (jetons de sécurité) entre fournisseurs et récepteurs sont généralement gérés via le protocole et le format d'échange SAML (Security Assertion Markup Language). Au lieu de SAML, la spécification WS-Fédération peut également être utilisée. WS-Federation se base sur WS-Trust (pour sécuriser les architectures orientées services) et sur WS-Security pour assurer la sécurité des messages. Il est important de noter que des efforts de collaboration sont menés entre SAML (notamment par Sun et l'implémentation Liberty Alliance) et WS-Federation (par Microsoft) et ont abouti à l'élaboration de nouvelles spécifications : le protocole Web de SSO *MetaDataExchange* et le profil Web de SSO *Interoperability* (Angal et al., 05). Liberty Alliance et Shiboleth sont les principaux SSO fédératifs utilisés par les institutions ou les entreprises pour le partage des données utilisateurs. Ces solutions sont bien adaptées à leurs contextes, mais elles restent bien entendu plus lourdes à déployer (car usage de SAML et de services Web SOAP) dans le contexte du Web grand public par rapport aux SSO décentralisés (qui s'appuient uniquement sur le protocole http).

***ii) Formats standards d'échange de données utilisateurs :*** au-delà des techniques SSO, les données utilisateurs peuvent être échangées entre systèmes par l'usage des formats de données. On distingue deux approches de formats de données : les approches de standardisation et les approches de médiation (Viviani et al., 10).

[7] http://openid.net/

- ***Approches par standardisation :*** elles s'appuient sur l'usage de spécifications standardisées de type XML : *User Modeling Markup Language* – UserML[8], *Unified User Context Model* – UUCM[9], *Attention Profile Markup Language* – APML[10], etc.) ou de type RDF (*General User Model Ontology* – GUMO[11], *Friend Of A Friend* – FOAF[12], *Semantically Interlinked Online Communities* – SIOC[13], *Simple Knowledge Organization System* - SKOS[14], etc.). Ces standards visent la généricité, toutefois ils peuvent s'avérer insuffisants pour prendre en compte les spécificités de certains systèmes.

- ***Approches par médiation :*** elles visent à réconcilier les standards existants dans le but de prendre en compte les spécificités d'un système donné. Le modèle utilisateur *Generic User Model Component* (GUC) (Sluijs et al., 05) suit particulièrement cette approche.

Au même titre que l'interopérabilité des sources de données, la fiabilité des sources de données est également un enjeu majeur lors de la sélection des données à utiliser pour construire les profils utilisateurs.

#### *2.2.2.2 Fiabilité des sources de données*

La fiabilité d'une source de données peut être relative à : (i) la fiabilité des utilisateurs à partir desquels ces données sont produites (certification des identités), (ii) la fiabilité de l'application qui produit ces données (qualité de l'application).

***i) Certification des identités :*** la question de la certification des identités se pose surtout dans le contexte du Web (réseaux sociaux numériques par exemple) où les utilisateurs peuvent aisément déclarer de fausses identités ou usurper des identités. Ce problème est donc assez récent et surtout abordé par de nouvelles solutions technologiques qui se proposent de certifier une identité sur le Web en vérifiant qu'elle correspond bien à une identité physique dans le monde réel. Cette vérification peut s'appuyer sur des sources de données jugées plus fiables pour identifier un utilisateur dans la vie réelle telles que : les données bancaires et adresses postales physiques des utilisateurs (exemple de MyId[15]) les données issues des fichiers publics disponibles dans des registres publics des états (BeenVerified[16], Trufina[17], NorthID[18], etc.), ou les données déjà vérifiées des cartes nationales d'identité numérique (exemple de Signicat[19]). Nous n'avons pas retrouvé de travaux de recherche qui se penchent sur ce problème.

***ii) Qualité des applications qui produisent les données :*** en fonction de l'application qui produit les données, la qualité des profils utilisateurs construits peut être plus ou moins bonne. Des auteurs comme (Guy et al., 08) démontrent clairement, via une expérimentation, l'impact de la fiabilité d'une source de données sur la qualité des profils utilisateurs construits. Dans leur expérimentation, des poids distincts sont affectés à plusieurs sources de données ou applications (email, tchat, blogs, social bookmarks, etc.) utilisées

[8] Voir (Heckman, 03)

[9] http://www.acronymfinder.com/Unified-User-Context-Model-%28data-architecture%29-%28UUCM%29.html

[10] http://apml.areyoupayingattention.com/

[11] Voir (Heckman, 03)

[12] http://www.foaf-project.org/

[13] http://sioc-project.org/

[14] http://www.w3.org/2004/02/skos/

[15] http://myid.is/

[16] http://www.beenverified.com/

[17] http://www.crunchbase.com/company/trufina-2

[18] http://www.northid.com/

[19] http://www.signicat.com/

pour construire des profils utilisateurs. L'expérimentation montre que la variation du système de pondération des sources de données impacte considérablement la qualité des profils construits.

Nous pouvons donc noter que dans la phase de sélection des données pour la construction des profils utilisateurs il est important d'étudier les contours des concepts de producteurs et sources de données présentés dans cette section. Les données sélectionnées à cette étape doivent ensuite être prétraitées.

## 2.3 Prétraitement de données

Une fois les sources de données définies à l'étape de sélection des données, l'étape du prétraitement traite les contenus de ces sources de données afin qu'ils soient conformes aux modèles ou aux entrées des algorithmes des étapes ultérieures du processus. En fait les sources de données issues de la sélection des données peuvent contenir de nombreuses inconsistances telles que :

- ***des données incomplètes :*** valeurs manquantes pour certains champs ou attributs par exemple,
- ***des bruits :*** erreurs ou exceptions produites lors des saisies ou de la collecte automatique des données (duplication de données par exemple),
- ***des incohérences :*** nommages et codages différents dans les données par exemple,
- ***etc.***

Pour prendre en compte ces inconsistances, plusieurs types de traitements peuvent être réalisés (Lemlouna et Layaida, 02) :

- ***nettoyage de données :*** dans le cas de données incomplètes, il peut s'agir par exemple d'ignorer les données manquantes, utiliser la valeur moyenne d'un attribut en remplacement, utiliser la valeur la plus probable (formule bayésienne ou arbre de décision) en remplacement, etc.
- ***discrétisation des données :*** convertir des attributs continus vers des attributs nominaux ou ordinaux,
- ***réduction des données :*** obtenir une représentation réduite du jeu de données, plus petite en volume, mais qui produit (ou presque) les mêmes résultats analytiques,
- ***transformation de données :*** ne conserver qu'un résumé d'un texte au lieu du texte entier, traduction d'un texte d'une langue à une autre, etc.
- ***transcodage de données :*** modifier l'encodage d'un média, modification d'une vidéo encodée en AVI en MPEG par exemple,
- ***transmodage des données :*** modifier le format d'un média, modification d'un fichier sonore en un fichier textuel par exemple,
- ***etc.***

Les concepts et techniques cités précédemment restent les mêmes que ceux utilisés de manière générale.

# 2.4 Structuration des données pour les profils utilisateurs

Dans cette étape, on peut regrouper les travaux de la littérature qui visent à structurer les différents types de données à prendre en compte lors de la construction des profils utilisateurs. Plusieurs modèles de données sont proposés dans ce sens, en fonction des objectifs de chaque auteur (Bouzhegoub et al., 05) (Zayani, 08) (Amato et al., 99) , FIDIS[20] (Future of Identity in the Information Society) etc. Une synthèse de ces travaux permet de distinguer cinq grandes catégories de données : les données explicites, les données implicites, les données de sécurité, les données de contexte et les données de sémantique.

## 2.4.1 Données explicites (feedback explicites)

Les données explicites sont les données fournies de manière explicite par l'utilisateur (via des formulaires par exemple). Ces données peuvent comprendre : les données d'identité (identifiant, nom, prénom, etc.), les données démographiques (âge, genre, adresse, etc.), les caractéristiques physiques (taille, couleur des yeux, etc.), le cursus académique, les emplois occupés, etc. (FIDIS). On peut également rajouter dans cette catégorie, des données ou des jugements explicites (*feedback explicites*) exprimés par les utilisateurs au cours de leurs activités. Par exemple, fournir une note sur une échelle de valeurs prédéfinies (notes que les internautes indiquent sur des produits qu'ils achètent sur Internet par exemple), faire une action de recommandation (article qu'un utilisateur recommande à un autre utilisateur par exemple), exprimer une opinion polarisée sur un objet (exemple du bouton « j'aime » sur Facebook), etc. Les données explicites sont des éléments de données du profil utilisateur qui peuvent directement être utilisées par les mécanismes de personnalisation tels que *MyYahoo !*[21], *Syskill & Weber* (Pazzani et al., 96), *WAWA* (Shavlik et Elassi-Rad, 99) (Shavlik et al., 98). Cependant l'acquisition des données explicites est une tâche qui peut se révéler très lourde pour les utilisateurs qui doivent se porter volontaires de fournir les données demandées (Kobsa, 01) et qui peuvent entraîner une surcharge cognitive suite aux nombreuses demandes de jugements explicites par les systèmes. La plupart du temps, cela génère un abandon ou un désintéressement des utilisateurs, et il résulte alors une détérioration de l'efficacité des mécanismes s'appuyant sur ces données.

## 2.4.2 Données implicites (feedback implicites)

Les données implicites sont les données collectées en observant le comportement des utilisateurs ou en scrutant leurs activités (et qui sont généralement utilisées pour déterminer leurs centres d'intérêts). L'activité peut correspondre à :

- ***l'utilisation d'un moteur de recherche :*** requêtes et documents sélectionnés,

[20] http://www.fidis.net/resources/deliverables/profiling/

[21] http://dir.yahoo.com/

- ***la navigation sur le Web :*** pages Web consultées, liens consultés,
- ***les publications sur un réseau social numérique :*** statuts, commentaires, opinions,
- ***les diverses applications :*** applications de bureau (exemple de MS Office), outils de messagerie électronique, éditeurs de textes, fichiers de logs,
- ***la consultation de bases de données ou de bases documentaires,***
- ***etc.***

Le principal avantage de la collecte de données implicites est qu'elle ne nécessite aucune implication directe de l'utilisateur, ni de temps passé à émettre des jugements, ni un effort d'attention particulier lors de ses interactions. En effet, toute interaction de l'utilisateur avec le système est considérée comme un jugement d'intérêts (Oard et Kim, 01). Basé sur la durée des interactions, le procédé d'acquisition de données implicites peut être classifié comme suit (Razmerita, 03) :

- ***procédé peu profond :*** il est basé sur l'observation du comportement d'interaction relativement à court-terme avec un système ; il ne tient pas compte des interactions avec le système durant les sessions précédentes,
- ***procédé profond :*** il observe le comportement de l'utilisateur durant son interaction à long terme avec un système ; il tient compte de l'historique entier des interactions de l'utilisateur avec le système.

La difficulté de l'exploitation de données implicites est la définition d'un processus d'interprétation du comportement observé dans un contexte d'application spécifique. Dans les systèmes où l'utilisateur manipule des documents par exemple, une catégorisation des comportements observables de l'utilisateur est généralement réalisée en fonction des unités élémentaires manipulées par l'utilisateur (Tableau 2.1) (Nichols, 98)(Oard et Kim, 01) (Claypool et al., 01).

| ***Catégorie comportements*** / **Unités élémentaires** | **Segment** | **Objet** | **Classe** |
|---|---|---|---|
| ***Examiner*** | Regarder, Ecouter, Défiler, Trouver, Soumettre une requête | Sélectionner | Naviguer |
| ***Retenir*** | Imprimer | Marquer, Sauvegarder, Supprimer, Envoyer un mail | |
| ***Référencer*** | Copier - Coller | Répondre, Ajouter un lien, Citer | |
| ***Annoter*** | Marquer | Juger, Publier | Organiser |
| ***Créer*** | Taper, Editer | Autre | |

Tableau 2. 1 : Catégorisation du comportement de l'utilisateur selon (Oard et Kim, 01) (Kelly et Teevan, 03).

Sur ce tableau, les catégories de comportement (*examiner, retenir, référencer, annoter*) se rapportent au but fondamental du comportement observé. L'unité élémentaire manipulée (*segment, objet, classe*) se rapporte à la plus petite unité informationnelle manipulée par l'utilisateur. (Kelly et Teevan, 03) a ajouté une cinquième catégorie de comportement (créer) aux quatre catégories d'Oard et Kim (Oard et Kim, 01). Cette catégorie décrit des comportements de l'utilisateur lors de la création de nouvelles unités informationnelles comme par

exemple l'écriture d'un papier. On peut rajouter à cette catégorisation des comportements obtenus par des dispositifs externes tels que le suivi des mouvements des yeux sur un écran ou l'analyse de fréquences cardiaques lors des interactions de l'utilisateur (ClayPool et al., 01)(Haolin et al., 09).

L'interprétation de ces comportements nécessite l'utilisation d'indicateurs traduisant l'intérêt de l'utilisateur pour les données manipulées. En effet, pour obtenir les centres d'intérêts de l'utilisateur à partir de son comportement, il est nécessaire de définir des indicateurs traduisant ces intérêts. La définition et le choix de ces indicateurs constituent la principale problématique pour la détermination des centres d'intérêts de l'utilisateur à partir des données implicites. La majorité des travaux sur le feedback implicite se sont focalisés sur la définition d'indicateurs permettant d'obtenir de bonnes prédictions sur la pertinence du comportement observé de l'utilisateur. Le tableau 2.2 présente une synthèse des principaux comportements observables ainsi que les indicateurs associés. Pour chacun de ces comportements, il est indiqué les conclusions de leurs performances obtenues par les expérimentations effectuées dans la littérature.

**(+) :** Bonnes performances de l'indicateur.

**(-) :** Mauvaises performances de l'indicateur.

**(#) :** Performances moyennes de l'indicateur.

| **Comportements** | **Indicateurs** | **Performances** | **Utilisation** |
|---|---|---|---|
| **Lecture** | Durée de lecture | (+) | Durée complète<br>Durée d'activation |
| | Mouvement souris | (+) (Kim et Chan, 05) | |
| | | (-) (Jung, 01)(Le et Waseda, 00) | |
| | Nombre de clics de la souris | (+) (Jung, 01) | |
| | | (#) (Kim et Chan, 05) | |
| | | (-)(Le et Waseda, 00) | |
| | Durée de défilement du scrollbar | (+)(Le et Waseda, 00) | Combiner avec la durée du défilement (clavier & souris) |
| | Défilement avec souris | (+)(Kim et Chan, 05)<br>(-)(Le et Waseda, 00) | Combiner avec d'autres indicateurs |
| | Nombre de clics Scrollbar | (+)(Jung, 01)(Goecks et al., 05)(Claypool et al., 01) | |
| | Défilement avec les touches clavier | (#)(Le et Waseda, 00) (Jung, 01) (Kim et Chan, 05) | Combiner avec l'indicateur de la souris |
| | Sélection du texte | (#)(Jung, 01) (Kim et Chan, 05) | Définir une bonne mesure de distance |
| **Sauvegarde**<br>**Impression**<br>**Annotation** | Action observée | (-) (Kim et Chan, 03) | Combiner avec d'autres indicateurs |

Tableau 2. 2 : Typologies des comportements & indicateurs d'intérêts associés (Zemirli, 08).

Les indicateurs implicites n'étant pas toujours évidents à exploiter pour juger de la pertinence des profils construits, plusieurs auteurs ce sont intéressés à l'évaluation de la qualité des profils construits à partir de ces indicateurs (Quiroga et Mostafa, 99)(White et al., 01)(Teevan et al., 05). Prises dans leur ensemble, les résultats de ces évaluations montrent que les profils construits par les indicateurs implicites donnent des résultats satisfaisants (lors de leur intégration dans les mécanismes qui s'y réfèrent, personnalisation par exemple) autant que les profils définis de manière explicite par les utilisateurs.

### 2.4.3 Données de contexte

Au-delà des données implicites issues du comportement de l'utilisateur, les profils utilisateurs doivent prendre en compte les données de contexte qui influencent très souvent ce comportement. La notion de contexte utilisateur est définie de plusieurs manières dans la littérature. (Dey et al., 00) proposent une définition très générique du contexte : « *any information that can be used to characterize the situation of an entity. Any person, place, or object that can be considered relevant to the interaction between a user and an application, including the user and the application themselves".* Pour (Abbar et al., 08) (Abbar et al., 09), « *a context is a set of features which describe the environment within which the user interacts with the information system".* Pour être plus précis dans cette définition, (Abbar et al., 09) définissent un modèle multidimensionnel de contexte utilisateur intégrant plusieurs dimensions plus ou moins présentes dans les travaux de la littérature :

- ***Le contexte temporel :*** qui est lié au paramètre *temps*. Ce contexte permet par exemple de définir des profils à court à terme ou à long terme de l'utilisateur (Tchuente et al., 10)(Tchuente et al., 12).

- ***Le contexte spatial :*** qui est lié au paramètre *lieu*. Le profil d'un utilisateur *au travail* n'est pas par exemple le même que son profil *à la maison* (Holland et Werner, 04).

- ***Le contexte matériel :*** qui est lié au dispositif matériel utilisé par l'utilisateur (taille de l'écran, vitesse du processeur, etc.) et qui pourra servir à définir des contraintes spécifiques de présentation de documents multimédias par exemple (Dromzee et al., 12) (Bouchra, 10). Plusieurs standards permettent de représenter ce type de données : CC/PP « *Composite Capability/ Preference Profiles* » (Klyne et al., 04), UAProf « *User Agent Profile* » (WAP Forum, 01), CSCP « *Comprehensive Structured Context Profiles* » (Buchloz et al., 04), etc.

- ***Le contexte environnement :*** qui est lié aux caractéristiques ambiantes de l'environnement d'interaction de l'utilisateur (température par exemple).

- ***Le contexte émotionnel :*** qui est lié à l'état psychique de l'utilisateur à un moment donné. Une analyse des sentiments ou des émotions sur les sites de réseaux sociaux numériques peuvent permettre de détecter ces éléments de contexte (Saad Missen et al., 10).

Les données de contexte de l'utilisateur sont essentielles afin de pouvoir définir des mécanismes d'adaptation sensibles aux différents contextes d'un utilisateur. Dans la construction du profil, une étape de « *contextualisation* » permet d'intégrer les données de contexte aux algorithmes permettant de construire le profil de l'utilisateur (Abbar et al., 09).

### 2.4.4 Données sémantiques

Dans le cas de la construction des profils utilisateurs par traitement de données textuelles, les données sémantiques permettent d'enrichir la sémantique des profils construits. Ceci est nécessaire pour lever d'éventuelles ambigüités liées à la polysémie de termes. Par exemple, « Java » peut désigner un langage de programmation en informatique ou île indonésienne en géographie. Comme données de sémantiques couramment utilisées dans la construction des profils, on peut distinguer :

- ***Les dictionnaires et thesaurus:*** il s'agit par exemple des dictionnaires de synonymes permettant de regrouper des ensembles de termes de la terminologie manipulée (Tchuente et al., 10). Les termes peuvent également être classifiés selon des thesaurus (Dousset, 06).

- ***Les ontologies :*** elles permettent d'avoir des structures sémantiques plus avancées avec la prise en compte de différents types de relations entre les termes (ou concepts) de la terminologie manipulée. Les ontologies peuvent être construites à partir de la terminologie manipulée pour construire les profils (Haddadi et al., 09). Des ontologies existantes (ontologies de référence) peuvent également être utilisées (WordNet, ODP, etc.) (Daoud et al., 09). Avec la multiplicité des langues utilisées sur la toile, des ontologies multilingues sont également utilisées (De Luca et al., 10).

Les données sémantiques sont donc fondamentales pour enrichir le sens des analyses et par conséquent d'affiner la qualité des profils construits.

### 2.4.5 Données de sécurité

Il s'agit des paramètres de sécurité que l'utilisateur définit pour limiter ou autoriser le traitement de ses données afin de construire son profil. Ces données de sécurité prennent également en compte les législations des états en ce qui concerne la manipulation des données personnelles des utilisateurs (Kobsa, 01). De par la nature très sensible des données utilisateurs, il est donc très important de prendre en compte les paramètres de sécurité définis par les utilisateurs et les restrictions liées à la législation dans le processus de développement de profils utilisateurs. Dans le cas d'Internet et de l'explosion du Web social en particulier, des tiers disposent de plus en plus d'outils leur permettant de collecter des masses importantes de données utilisateurs (Bonneau et al., 09), et les législations des états ne sont pas forcément à jour pour contrôler ces accès. Aujourd'hui, la protection de la vie privée est devenue un enjeu majeur sur Internet avec toutes les questions liées à l'identité numérique (identités multiples, identités certifiées, usurpation d'identités, etc.) (cf. FIDIS[22]).

## 2.5 Fouille de données pour les profils utilisateurs

Une fois les données structurées, elles sont utilisées par les algorithmes de fouille de données pour construire le profil de l'utilisateur. Plusieurs techniques de fouille de données peuvent être utilisées. (Gao et al., 10)

[22] http://www.fidis.net/resources/deliverables/profiling/

classifient les techniques utilisées suivant les modèles de profils utilisateurs à construire (modélisation du comportement, modélisation des centres d'intérêts, modélisation des intentions) (figure 2.5).

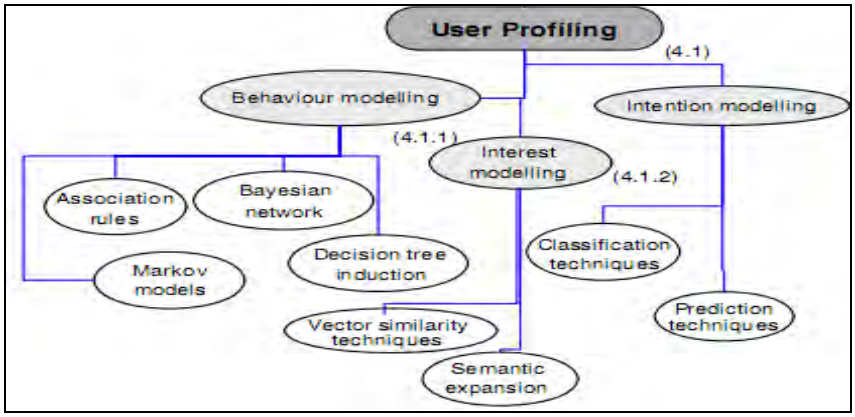



Figure 2. 5 : Familles d'algorithmes de fouille de données en fonction des profils à construire

## 2.5.1 Modélisation comportementale (*behaviour modeling*)

Ce type de modélisation est généralement utilisé dans le contexte du Web. Elle consiste à analyser les comportements des internautes via les historiques de navigation ou les transactions qu'ils effectuent sur des serveurs Web (Frias-Martinez et al., 06), dans le but de déterminer des parcours de navigation récurrents, de valider des stratégies marketing ou de vérifier la pertinence des campagnes marketing. Il y a des méthodes spécifiques à ce type d'analyses (Nanopoulous et al., 01) : (i) les règles d'association qui sont utilisées pour prédire les futures requêtes de l'utilisateur ; (ii) les chaînes de Markov utilisées pour prédire les futures URL qui seront visités par l'utilisateur (Sarukkai, 00) ; (iii) les arbres de décisions qui sont les plus utilisés et qui servent généralement à prédire les futures interactions de l'utilisateur (Kim et al., 02) (Cho et al., 02).

## 2.5.2 Modélisation des centres d'intérêts (*interest modeling*)

Les centres d'intérêts sont définis par une fonction *c(i)* qui donne le degré d'intérêt ou de désintérêt d'un utilisateur pour un item i en analysant son comportement antérieur (Jung et al., 05). Plusieurs méthodes ont été définies pour construire les centres d'intérêts des utilisateurs. Trois approches sont fréquemment utilisées (Schubert et Koch, 02) : l'approche directe, l'approche semi-directe et l'approche indirecte.

L'approche directe consiste à demander directement aux utilisateurs ce qu'ils aiment, en listant par exemple toutes les catégories de centres d'intérêts et en leur demandant de faire des sélections.

L'approche semi-directe consiste à demander aux utilisateurs d'attribuer des notes à des items (produits par exemple) qu'ils ont manipulés (achetés par exemple). Ces deux premières approches s'appuient donc sur les données explicites présentées dans la section précédentes.

L'approche indirecte consiste à construire les centres d'intérêts par analyse des données issues des interactions antécédentes de l'utilisateur. Cette approche s'appuie sur les données implicites présentées dans la

section précédente. La pondération des centres d'intérêts pour cette approche peut être représentée de trois manières : pondération vectorielle, pondération probabiliste, pondération par règle d'association.

- ***La pondération vectorielle :*** chaque centre d'intérêt est pondéré par son degré de pertinence pour l'utilisateur. Elle s'appuie généralement sur des mesures caractérisant l'importance de termes dans les documents manipulés (système de recherche d'information par exemple) par l'utilisateur. Pour trouver les termes d'un document qui représentent le mieux son contenu sémantique, (Robertson et Jones, 76) ont défini la fonction de pondération d'un terme dans un document connu sous la forme TF.IDF, qui est reprise dans différents travaux de pondération des centres d'intérêts du profil utilisateur (Joachim, 97)(Bilssus et al., 99).

    - *TF (Term Frequency)* : cette mesure est proportionnelle à la fréquence du terme dans le document. L'idée sous-jacente est que plus un terme est fréquent dans un document, plus il est important dans la description de ce document.
      Le TF est souvent utilisé suivant l'une des déclinaisons suivante :
        1. $TF$ *:* utilisation brute.
        2. $\log(1+TF)$

    - *IDF (Inverse Document Frequency) :* mesure l'importance d'un terme dans toute la collection. L'idée sous-jacente est que les termes qui apparaissent dans peu de documents de la collection sont les plus représentatifs du contenu de ces documents que ceux qui apparaissent dans tous les documents de la collection.
      Cette mesure est exprimée selon l'une des déclinaisons suivantes :
        1. $IDF = \log(\frac{N}{d_f})$
        2. $IDF = \log(\frac{N - d_f}{d_f})$

      Où $d_f$ est la proportion de documents contenant le terme et N le nombre total de documents dans la collection.

      La fonction de pondération TF.IDF consiste à multiplier les deux mesures TF et IDF. Une formule largement utilisée est la suivante :

      $$TF.IDF = \log(1+TF) * \log(\frac{N}{d_f}) \qquad (1)$$

      La fonction TF.IDF a parfois du mal à retrouver les termes pertinents d'un document lorsque celui ci comporte trop de termes. Dans ces cas, des solutions dérivées telles que *TF.IDF reduction* et *LSA (Latent Semantic Analysis)* peuvent être utilisées (Isokazi et al., 02) ou encore les machines à vecteurs de support (SVM) (Billsus et al., 99).

- ***La pondération probabiliste :*** chaque centre d'intérêt est pondéré via sa probabilité de pertinence pour l'utilisateur. Cette probabilité est utilisée pour prédire le comportement futur de l'utilisateur, dans un réseau bayésien par exemple (Friedman et al., 97) (Joachims, 97).

- ***La pondération par règles d'association :*** chaque centre d'intérêt est pondéré par la mesure de la règle d'association entre ce centre d'intérêt et les centres d'intérêts antécédents de l'utilisateur (Nanopoulous et al., 01).

### 2.5.3 Modélisation des intentions (*intention modeling*)

Une intention ici désigne le but (objectif) que l'utilisateur veut atteindre via le système d'information, ou la raison pour laquelle l'utilisateur utilise le système d'information (Frias-Martinez et al., 06). La modélisation des intentions des utilisateurs consiste donc à construire un modèle qui permettra d'identifier le but de chaque utilisateur du système d'information. Par exemple, des clients d'un site de e-commerce peuvent être divisés en deux groupes : ceux qui ont réellement pour but d'acheter et ceux qui n'ont pas pour but d'acheter. La modélisation des intentions s'appuie largement sur les techniques de classification avec des catégories prédéfinies (Frias-Martinez et al., 06). Pour modéliser les intentions utilisateurs, (Ruvini, 03) présente une approche qui infère l'objectif de l'utilisateur à partir des machines à vecteurs de supports (SVM). Les réseaux bayésiens, les arbres de décisions ou les réseaux de neurones produisent également de bons résultats (Horvitz et al., 98) (Chen et al., 02) (Frias-Martinez et al., 06). Dans certains travaux, la modélisation des intentions est traitée comme la modélisation comportementale ou la modélisation des centres d'intérêts. Toutefois, la modélisation des intentions peut être perçue comme une modélisation ultérieure qui réutilise les profils construits dans la modélisation comportementale et la modélisation des centres d'intérêts.

Ces algorithmes de fouille de données produisent des profils utilisateurs qui peuvent être représentés de plusieurs manières.

## 2.6 Représentation des profils utilisateurs

Les résultats produits par la modélisation comportementale et la modélisation des intentions représentent généralement de la connaissance qui peut directement être analysée et exploitée par les décideurs. Par exemple, l'observation de parcours de navigation qui ne convertissent pas les visiteurs d'un site en acheteurs seront réétudiés par des experts du domaine. Les résultats de ces modélisations comportementales et des intentions représentent en elles-mêmes la connaissance qui est le but ultime du processus dans ce cas. Par contre, la modélisation des centres d'intérêts permet de construire des profils utilisateurs qui ne sont que des données ou des informations qui seront réutilisées par les mécanismes appropriés d'adaptation de l'information à l'utilisateur (personnalisation, recommandation par exemple) afin de produire la connaissance. Ces profils sont généralement représentés de trois manières dans la littérature : représentation ensembliste, représentation par réseaux sémantiques, et représentation conceptuelle.

## 2.6.1 Représentation ensembliste

L'approche ensembliste consiste à représenter le profil de l'utilisateur par des paquets de termes pondérés. On parle également de représentation vectorielle par analogie au modèle vectoriel de Salton (Salton, 71) sur lequel elle se base. Ces paquets de termes, traduisant les centres d'intérêts de l'utilisateur peuvent être regroupés différemment selon l'approche suivie pour exploiter le profil de l'utilisateur.

On distingue dans la littérature quatre grandes approches de représentation de profils utilisateurs basés sur la représentation ensembliste :

- Par liste de mots clés où chaque mot correspond à un centre d'intérêt spécifique (Freitag et al., 95).
- Par vecteur de termes pondérés pour chaque centre d'intérêts (Tebri et al., 05).
- Par ensemble de vecteurs de termes pondérés (ou non) indépendants, pour prendre en compte des centres d'intérêts multiples (Somlo et Howe, 03) où chaque vecteur correspond à un domaine d'intérêt (Pazzani et al., 96).
- Par définition d'une relation d'ordre entre les centres d'intérêts du profil, on parle dans ce cas de préférences (Kießling, 02).

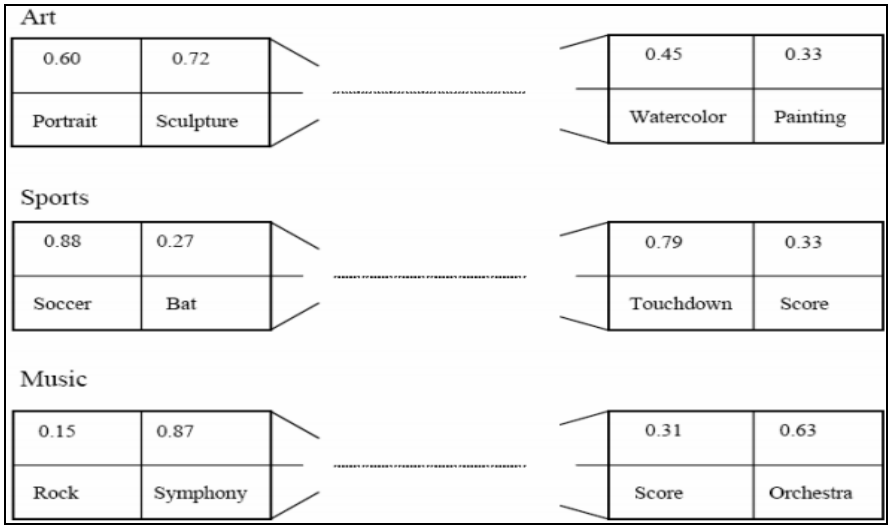



Figure 2. 6: Un exemple de profil représenté par des mots clés

La représentation ensembliste fait partie des premières représentations de profils utilisateurs qui ont été utilisées. La pondération des termes est généralement basée sur le schéma tf.idf communément utilisé en recherche d'information (Salton et Yang, 73). Le poids associé à chaque terme permet de représenter son degré d'importance dans le profil de l'utilisateur. La figure 2.6 donne un exemple de profil utilisateur représenté par des mots clés pondérés. Ce profil contient trois centres d'intérêts : *Art, Sport, et Musique.* Chaque centre d'intérêt est représenté par un ensemble de termes pondérés. Music = <(Rock, 0.15), (Symphony, 0.87) … > est un extrait de l'ensemble de termes pondérés représentant le centre d'intérêt Music.

Plusieurs systèmes d'accès personnalisé utilisent ce type de représentation. On peut citer *Anatagonomy* (Sakagami et Kamba, 97), un système personnalisé de consultation de nouvelles et de journaux en ligne, *Fab* (Balabanović et Shoham, 97) un système de recommandation de pages Web, *Letizia* (Lieberman, 95), un

système d'aide à la navigation, et *Syskill & Webert* (Pazzani et al., 96), un système de recommandation. Tous ces systèmes proposent des profils utilisateurs représentés par une liste de mots clés.

Dans le même cadre, *PEA* (Montebello et al., 97), un système d'aide personnalisé à la navigation établit des profils utilisateurs basés sur la représentation vectorielle, en utilisant des termes extraits des pages annotées par l'utilisateur lors de sa navigation (Cabanac et al., 07). Cependant, à la différence des autres systèmes, plutôt que de créer un profil unique pour chaque utilisateur, dans *PEA*, l'utilisateur est représenté par un ensemble de vecteurs de termes pondérés, pour chaque annotation. Le principe de base de ce système est que l'utilisateur peut avoir plusieurs centres d'intérêts lors de sa recherche. La combinaison des termes représentant ces centres dans un même vecteur permet d'obtenir un profil couvrant tous ses centres d'intérêts.

*WebMate* (Chen et Sykara, 98) établit également des profils utilisateurs contenant un vecteur de termes par centre d'intérêt, tandis qu'*Alipe* (Widyantoro et al., 99) augmente cette approche en représentant chaque centre d'intérêt avec trois vecteurs : un descripteur à long terme et deux descripteurs à court-terme : un négatif et un positif (représentant les centres d'intérêts non intéressant et intéressant , respectivement).

La représentation par liste de mots clés et/ou par classes de vecteurs de termes apporte l'avantage de la simplicité de mise en œuvre. Néanmoins, même si ces systèmes prennent en considération des centres d'intérêts multiples en utilisant plusieurs vecteurs, cette représentation manque de structuration. Elle ne facilite ni l'interprétation, ni la prise en compte de différents niveaux de granularité caractérisant l'utilisateur (Bottreaud et al., 04). En effet, la plupart des utilisateurs ont des intérêts multiples, leur généralisation dans un vecteur simple n'est pas représentative de la réalité.

L'efficacité des profils dans cette approche dépend fortement du degré de généralisation pour représenter les centres d'intérêts. Le problème est lié à l'application d'une analyse statistique de mots clés indépendamment de toute information contextuelle. Dans le cas d'une représentation en hiérarchies ou classes de concepts, le rapport généralisation/spécification existant naturellement dans ce genre de structure permet d'avoir une représentation plus réaliste du profil utilisateur. Ceci est par exemple utilisé dans la représentation par réseaux sémantiques.

### 2.6.2 Représentation par réseaux sémantiques

Afin d'adresser le problème de polysémie des termes inhérents à la représentation ensembliste, une première solution consiste à représenter le profil par un réseau de nœuds pondérés dans lequel chaque nœud représente un concept traduisant un centre d'intérêt utilisateur. Ce type de représentation offre le double avantage de la structuration et de la représentation associative (relations entre les termes) permettant de considérer l'ensemble des aspects représentatifs du profil.

Les centres d'intérêts sont souvent représentés par des relations de paires de nœuds dans lesquelles chaque nœud contient un terme issu de données implicites utilisées pour construire le profil. Les arcs reliant les nœuds sont créés sur la base de cooccurrences entre ces termes.

Cependant, la représentation séparée de chaque mot par des nœuds dans le réseau sémantique n'est pas assez précise pour décliner les différentes significations des centres d'intérêts de l'utilisateur. Une alternative possible est d'exploiter les sources externes telles que les ontologies pour établir les liens entre les nœuds.

Dans ce cadre, le système *SiteIF* (Stefani et Strapparava, 98) propose d'utiliser les concepts inhérents à WordNet pour regrouper des termes semblables dans des concepts appelés des « *ensembles de synonymes* » ou des *synsets*. Le profil utilisateur est alors représenté comme un réseau sémantique dans lequel les nœuds sont les *synsets* et les arcs sont des cooccurrences des membres de *synset* avec le document intéressant l'utilisateur. Les nœuds et les poids des arcs représentent le niveau d'intérêt de l'utilisateur.

Une approche similaire a été étudiée par le système *InfoWeb* (Gentili et al., 03). Initialement, chaque réseau sémantique contient une collection de nœuds unitaires dans laquelle chaque nœud représente un concept. Les nœuds du concept, appelés *planètes,* contiennent un vecteur unique de termes pondérés. Lorsque de nouvelles informations sur l'utilisateur sont collectées, le profil est enrichi en intégrant les termes pondérés dans les concepts correspondants. Ces termes sont stockés dans les nœuds auxiliaires *(satellites)* qui sont liés aux nœuds concepts (*planètes*) associés. La figure 2.7 montre un exemple extrait d'un modèle d'utilisateur basé sur cette représentation. Sur cette figure, les cercles représentent les *planètes* (*Painting, Sculpture, Restauration, Environnement, Pollution*), et les carrés aux bords arrondis représentent les nœuds auxiliaires *satelites* liés aux planètes (pour la *Sculpture* par exemple, les *satelites* associés dans ce profil sont *Museum*, *Sculptor*, *Painter*, et *Chisel*). Les *planètes* peuvent également être reliées entre elles (cas de la relation entre *Pollution* et *Environnement* par exemple). Ces deux types de relations (entre *planètes* et entre *planètes* et *satelites*) permettent d'obtenir des profils sémantiques beaucoup plus structurés que de simples vecteurs de termes.

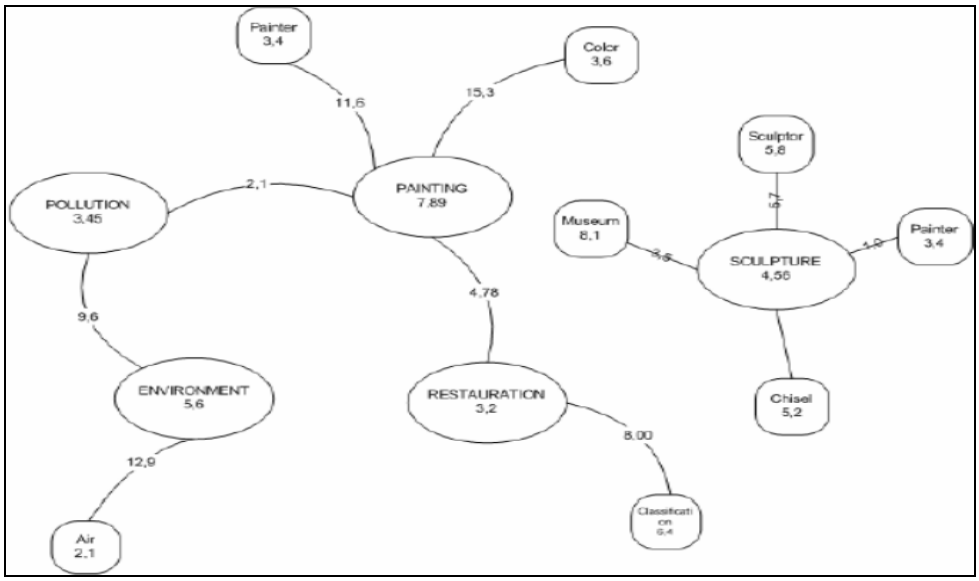



Figure 2. 7: Exemple de profil sémantique de l'utilisateur

Cette représentation a été prolongée dans *WIFS* (Micarelli, Sciarrone, 04), une interface de filtrage d'information pour personnaliser les résultats du moteur de recherche d'*AltaVista*[23]. Dans ce système, le profil de l'utilisateur est représenté par trois composantes : *un entête*, intégrant les données personnelles de

[23] http://fr.altavista.com/

l'utilisateur, un ensemble de *stéréotypes*, et une *liste de centres d'intérêts*. Un stéréotype comporte un ensemble de centres d'intérêts, représentés par une classe d'information. Chaque classe contient trois champs : *domaine (Domain)*, *matière (Topic)*, *poids (Weight)*, et deux autres informations supplémentaires : *les liens sémantiques* (*Semantic Links*) et la *justification des liens (Justification Links)*.

Le domaine identifie un centre d'intérêt de l'utilisateur, la matière contient le terme spécifique employé par l'utilisateur pour décrire son centre d'intérêt, et le poids indique le degré d'intérêt de l'utilisateur pour ce centre d'intérêt. Sur la figure 2.8 par exemple, les *slots* désignent les classes, le slot-1 représente le centre d'intérêt (matière) « *Learning* » qui a un poids « *9* » dans le domaine « *Artificial Intelligence* ». L'information *Semantic Links* donne la liste des mots clés co-occurents dans le document lié à la classe et ayant un degré de similarité avec le centre d'intérêt. Cette information permet d'avoir une représentation du profil suivant des réseaux sémantiques (la *planète* ici est le centre d'intérêt, et les *satelites* sont les mots clés co-occurents). L'information *Justification Links* est un supplément d'information permettant de connaître comment a été acquis le centre d'intérêt (par *interview* par exemple pour le domaine « *Artificial Intelligence* » de la figure 2.8).

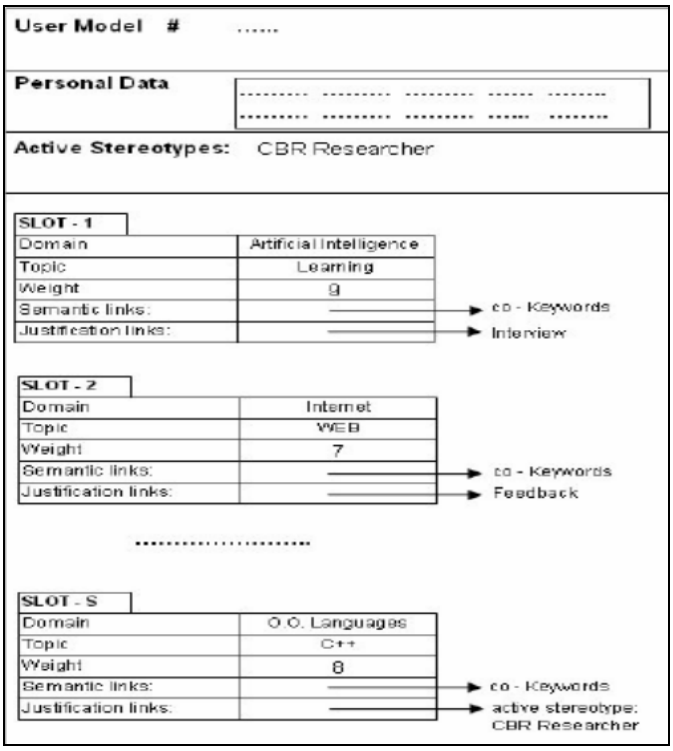



Figure 2. 8 : Extrait de profil sémantique WIFS

Comme on peut le voir avec les représentations *SiteIF* et *WIFS*, on peut imaginer plusieurs types de relations entre centres d'intérêts dans un réseau sémantique en fonction des besoins contextes. Pour être plus générique et représenter de multiples relations entre centres d'intérêts, les ontologies (représentation conceptuelle) sont une alternative plus efficace.

### 2.6.3 Représentation conceptuelle

La représentation du profil met en évidence, dans cette approche, une ou plusieurs relations sémantiques entre les informations du profil. Suivant une direction proposée dans un contexte plus général (Huhns et Stevens, 99), cette représentation offre une alternative intéressante à l'approche des réseaux sémantiques.

En effet, les travaux actuels tendent à représenter le profil sous forme d'une ontologie de concepts personnels en se basant sur les connaissances contenues dans les ontologies plutôt que de construire les profils utilisateur seulement à partir des données implicites collectées de ses interactions (Chaffee et Gauch, 00). La représentation est essentiellement basée sur l'utilisation d'ontologies (Gauch et al., 03)(Brini et al., 05)(Golemati et al., 07) ou de réseaux probabilistes (Lin et al., 05)(Wen et al., 04). Dans ce type d'approche, les liens entre les concepts sont explicitement induits de l'ontologie et le profil résultant inclura des relations informationnelles plus diverses et spécifiques.

La représentation conceptuelle est semblable à la représentation par réseaux sémantiques, dans le sens où elle représente les centres d'intérêts de l'utilisateur par un réseau de nœuds conceptuels. Cependant, dans l'approche conceptuelle, les nœuds correspondent à des domaines abstraits représentant les centres d'intérêt de l'utilisateur, contrairement à l'approche sémantique où les centres d'intérêts sont représentés par des mots spécifiques, ou ensemble de mots relatifs. La représentation conceptuelle peut également être assimilée à une approche ensembliste (vectorielle) du fait que les domaines sont souvent représentés comme les vecteurs de termes pondérés. Néanmoins, les termes de ces vecteurs sont regroupés pour former un domaine spécifique et non de simples mots clés.

De l'association des centres d'intérêts de l'utilisateur aux concepts des domaines de l'ontologie, on obtient un profil représenté sous forme d'une hiérarchie de concepts. Les données implicites issues des interactions de l'utilisateur sont classifiées dans ces concepts et l'intérêt de l'utilisateur pour de tels concepts est enregistré. Plusieurs mécanismes peuvent être appliqués pour exprimer le degré d'intérêt de l'utilisateur pour chaque domaine. Néanmoins la technique la plus utilisée est l'affectation d'un poids d'une valeur numérique pour chaque domaine (Gauch et al., 03).

On trouve dans la littérature plusieurs types de structures hiérarchiques et ressources sémantiques. Les plus simples sont construits sur la base d'une taxonomie de concepts ou d'un thesaurus de référence. A titre d'exemple, les systèmes (Guarino et al., 99)( Knight et Luk, 94) utilisent l'ontologie *Sensus*, une taxonomie d'approximativement 70 000 nœuds, et un sous ensemble de l'annuaire Yahoo ! (Labrou et Finin, 99) en tant que hiérarchie de référence.

On trouve également *ODP (Open Directory Project)* qui est une hiérarchie de concepts open source au format RDF largement adoptée par de nombreux systèmes utilisant l'approche conceptuelle telle que *OBIWAN (Ontology Based Informong Web Agent Navigation)* (Prestchner, 99), *Personae* (Tanudjaja et Mui, 02), *Outride* (Pitkow et al., 02).

L'utilisation d'ODP comme source conceptuelle diffère d'un système à un autre. Ainsi dans OBIWAN les profils sont représentés en utilisant 1869 concepts des trois principaux niveaux de la hiérarchie de concepts d'ODP (Chaffee et Gauch, 00). Avec l'élargissement du contenu d'ODP, le nombre de concepts a augmenté d'environ 2991 concepts extraits également des trois niveaux. En outre, le système *Personae* ne se limite pas

aux premiers niveaux d'ODP, mais exploite les différents concepts extraits des différents niveaux de la hiérarchie ODP. Il établit ainsi des profils utilisateurs plus spécifiques. Le système Personae gère ainsi des profils de moyenne taille. Quant au système *Outride* (Pitkow et al., 02), dont les profils utilisateurs sont également plus petits que ceux utilisés dans OBIWAN, il exploite seulement 1. 000 concepts d'ODP.

L'organisation sous forme hiérarchique des concepts du profil est adaptée différemment pour chaque utilisateur afin de mieux exprimer les caractéristiques de généralisation/spécification inhérentes à ce type de structure (Bloedorn et al., 96). Les niveaux de la hiérarchie peuvent être statiques et fixes (Trajkova, Gauch, 04) ou changer dynamiquement selon le degré d'intérêt de l'utilisateur accordé à chaque concept (Chen et al., 01).

Dans un contexte aussi dynamique et volumineux que le Web, la représentation des profils basée sur les ontologies engendre certains problèmes liés à l'hétérogénéité et la diversité des centres d'intérêts de l'utilisateur. En dépit du fait que ces profils peuvent contenir un nombre considérablement élevé de concepts, ces concepts n'englobent que partiellement le nombre potentiellement infini des centres d'intérêts spécifiques de chaque utilisateur. Par exemple, *Yahoo !* peut représenter le concept *base-ball* à l'intérieur de celui du *sport* mais ne pas représenter un intérêt plus spécifique pour une équipe ou un joueur célèbre (ou pas) de *base-ball.* En outre, les ontologies imposent leur organisation de concepts aux profils utilisateurs qui ne sont pas nécessairement en correspondance avec les perceptions de l'utilisateur. D'ailleurs les utilisateurs peuvent avoir différentes perceptions pour un même concept, ce qui peut engendrer des représentations peu précises de l'utilisateur (Godoy, 06).

Une fois les profils utilisateurs construits et représentés suivant les types de représentation présentés dans les trois dernières sections, ils sont soit directement interprétés par l'humain dans les systèmes d'analyse comportementale (le profil représente alors la connaissance extraite à partir des données, dernière étape du processus), soit exploités par les systèmes d'adaptation de l'information à l'utilisateur (Figure 2.1). Dans la partie qui suit, nous nous intéressons à ce dernier cas d'usage des profils utilisateur.

# 2.7 Usage des profils dans les systèmes d'adaptation de l'information à l'utilisateur

## 2.7.1 Généralités

Les profils utilisateurs construits et représentés peuvent être perçus en fonction de leur usage comme de la connaissance (systèmes d'analyses comportementales) ou comme des données ou informations qui seront utilisées pour renvoyer à l'utilisateur les contenus correspondant à ses besoins spécifiques (systèmes adaptatifs). A l'origine, les systèmes adaptatifs sont apparus pour essayer de résoudre deux problèmes : la surcharge cognitive et la désorientation (Conklin et al., 87). Le problème de surcharge cognitive est lié à la difficulté que peut avoir l'utilisateur à sélectionner l'information correspondant à ses besoins spécifiques face à une masse importante de données. Par exemple un utilisateur informaticien qui soumet la requête « java » à un moteur de recherche et qui s'attend à des résultats liés au langage de programmation java, aura du mal à

choisir les résultats qui l'intéresse dans une masse de résultats où s'entremêlent ceux liés au tourisme et ceux liés à la région « java » en Indonésie. Le problème de désorientation est lié au fait que l'utilisateur ne sait plus quel chemin suivre lorsqu'il navigue via interface utilisateur (par exemple lors de la navigation dans un site Web). Trois questions principales lors de la mise en œuvre d'un système adaptatif peuvent être posées (Brusilovsky, 96)( Brusilovsky, 01): dans quel domaine l'adaptation s'applique-t-elle ? Qu'est ce qui peut être adapté ? Sur quelles données et méthodes les systèmes adaptatifs se basent t-ils ?

Pour répondre à la première question « Dans quel domaine l'adaptation s'applique-t-elle ? », Brusilovsky a distingué six espaces d'application (Brusilovsky, 96) : les Systèmes Hypermédias Educatifs (Conlan et al., 02)(Chen et Duh, 08), les Systèmes d'Information en ligne (Lai et al., 03)(Das et al., 07), les Systèmes d'aide en ligne (Mock et Vemuri, 97), les Systèmes Hypermédias de Recherche d'Information (Zayani, 08)(Zemirli, 08), les Systèmes d'Information Institutionnels et les Vues Personnalisées.

Pour répondre à la deuxième question « Qu'est-ce qui peut être adapté ? », deux types de dimensions sont proposées (Brusilovsky, 96) : d'une part, les dimensions d'adaptation du contenu et de la présentation pour répondre à la problématique de la surcharge cognitive et, d'autre part, la dimension d'adaptation de la navigation pour répondre à la problématique de la désorientation.

Pour répondre à la troisième question « Sur quelles données et techniques le SA se base t'il ? », des modèles de référence ont été définis, parmi lesquels nous citons le modèle récent de Gao qui contient la modélisation utilisateur (*user profiling*), la modélisation des contenus (*content modeling*), et les techniques de filtrage d'information (*filtering*) (figure 2.9).

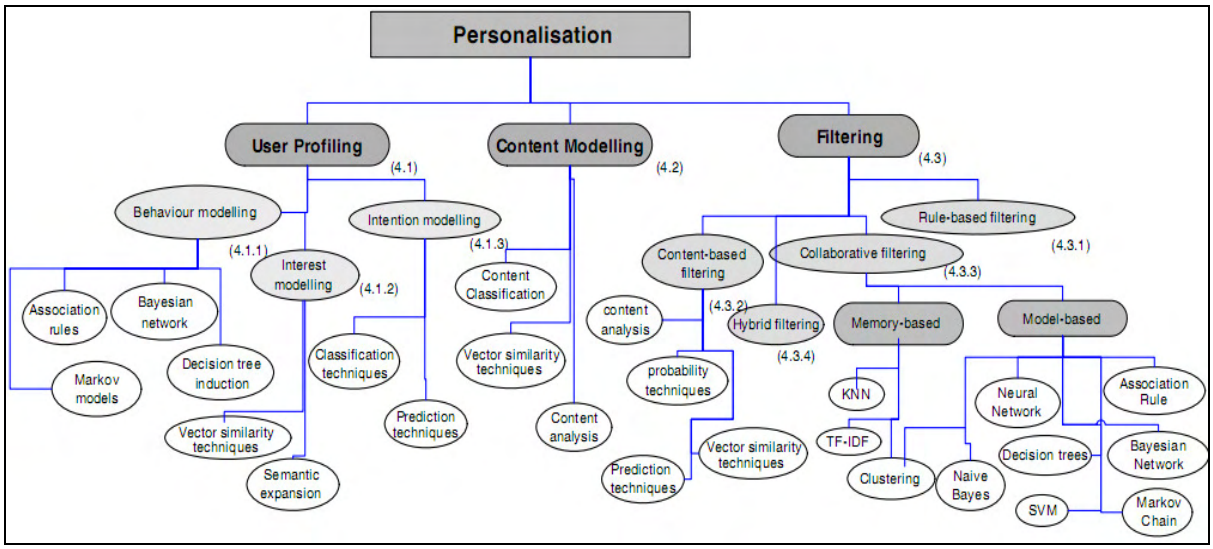



Figure 2. 9 : Données et techniques pour les systèmes adaptatifs (Gao et al., 10)

La modélisation utilisateur a été étudiée dans les sections précédentes. La modélisation des contenus consiste à définir les termes caractéristiques des documents qui seront adaptés (pages Web, documents dans une plateforme de e-learning, produits dans un site de e-commerce, etc.).

Les documents sont généralement représentés de la même manière que les profils utilisateurs (vecteurs de termes, réseaux sémantiques, ontologies, etc.). Les techniques récentes utilisées pour modéliser les documents s'appuient principalement sur LSA/LSA (*Latent Semantic Analysis/Indexing*) et PLSA/LSA (*Probabilistic*

*Latent Semantic Analysis/Indexing*). LSA est utilisée pour obtenir des relations termes-documents et concepts-documents à partir d'une matrice termes-documents (Hofmann, 04). PLSA est une vue statistique de LSA qui estime plutôt les probabilités dans la matrice concepts-documents à partir d'une collection de documents (Hofmann, 99)(Jin et al., 04). D'autres méthodes de classification permettent également d'affecter des termes ou des concepts à des documents (SVM, arbres de décisions, tf.idf, kNN, etc.) (Gauch, 03). Toutes ces techniques permettent donc de construire en quelque sorte les « profils des documents ». Les techniques de filtrage mettent en relation les profils et les documents (et éventuellement des requêtes utilisateurs) afin de renvoyer les contenus correspondant aux besoins spécifiques de l'utilisateur. On distingue plusieurs techniques : filtrage basé sur les contenus, filtrage basés sur les règles, filtrage collaboratif, et filtrage hybride. Nous les décrivons dans les sections qui suivent.

## 2.7.2 Filtrage par contenus (*content based filtering*)

Le filtrage basé sur les contenus facilite le filtrage (ou l'adaptation) de l'information correspondant aux besoins spécifiques de l'utilisateur en comparant le profil des documents (ou ressources) avec le profil de l'utilisateur (Chedrawi et Abidi, 06). On peut distingue deux cas de filtrage par contenus : lorsque le filtrage est réalisé sans requête préalable de l'utilisateur (recommandation par contenus) et lorsque le filtrage est réalisé suite à une requête de l'utilisateur (recherche d'information personnalisée).

### *2.7.2.1 Recommandation par contenus*

Il est généralement utilisé dans les sites de e-commerce qui vendent de très importantes variétés de produits alors que chaque client potentiel est très souvent intéressé par un nombre très limité de produits. C'est également le cas de systèmes de recommandation de « news » (*tweets* par exemple) qui doivent sélectionner des « news » intéressantes pour un utilisateur parmi un nombre très important de « news » publiées à un moment donné. Un système de recommandation par contenus va comparer le profil des ressources (produit ou « news » par exemple) avec le profil de l'utilisateur pour sélectionner les ressources correspondantes à ses besoins spécifiques (Figure 2.10). Le filtre utilisé pour sélectionner les ressources pertinentes pour l'utilisateur est une fonction de similarité entre le profil de l'utilisateur et chaque ressource. Cette fonction peut être déterminée de plusieurs manières, la plus courante est le cosinus de similarité qui mesure le cosinus de l'angle entre le vecteur représentant le profil de l'utilisateur et le vecteur représentant une ressource (Adamavicius et Thuzelin, 05), formule (2).

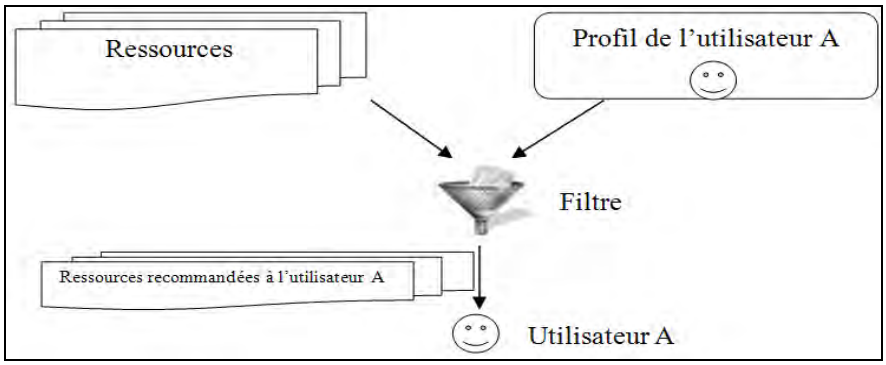



Figure 2. 10 : Recommandation par contenus

Si k est le nombre total de termes, et que les vecteurs $\vec{W}_u$ et $\vec{W}_r$ représentent le profil de l'utilisateur et le profil d'une ressource, *i* est un terme à la fois dans $\vec{W}_u$ et $\vec{W}_r$ ($\vec{W}_{i,u}$ est le poids du terme i dans le profil de l'utilisateur u, et $\vec{W}_{i,r}$ est le poids du terme i dans le profil de la ressource *r*).

$$Sim\ (\vec{W}_u, \vec{W}_r) = Cos\ (\vec{W}_u, \vec{W}_r) = \frac{\vec{W}_u * \vec{W}_r}{\left\|\vec{W}_u\right\|^2 * \left\|\vec{W}_r\right\|^2} = \frac{\sum_{i=1}^{K} W_{i,u} * W_{i,r}}{\sqrt{\sum_{i=1}^{K} W_{i,u}^2} * \sqrt{\sum_{i=1}^{K} W_{i,r}^2}} \qquad \textbf{(2)}$$

Les ressources les plus proches du profil de l'utilisateur (suivant la fonction cosinus par exemple) sont alors recommandées en priorité à l'utilisateur.

### *2.7.2.2 Recherche d'information personnalisée*

La recherche d'information (RI) désigne l'ensemble des méthodes, procédures et techniques pour l'acquisition, l'organisation, le stockage, la recherche et la sélection d'information (données, textes, images, vidéos). A la différence de la recommandation par contenus, l'utilisateur exprime son besoin d'information par une requête dans un système de RI. Dans un système de RI classique, la requête de l'utilisateur (liste de mots pondérés) est appariée avec la liste des documents indexés (liste de mots pondérés) afin de renvoyer des documents pertinents correspondant à la requête de l'utilisateur (moteurs de recherche Google, AOL, Yahoo !, etc.). Dans un environnement hypertexte comme le Web, les documents (pages Web) ne sont pas indexés uniquement sur leur contenu (liste de mots clés), mais aussi via des métriques additionnelles basées sur la structure des liens entre les pages. L'algorithme PageRank de Google (Brin et Page, 98) est l'un des plus connu dans ce contexte.

Les requêtes exprimées par les utilisateurs en RI classique sont généralement courtes et peuvent comporter des ambigüités (Ruthven et Lalmas, 03). Par exemple, deux utilisateurs qui saisissent la requête « Java » dans un moteur de recherche peuvent s'attendre à des résultats complètement différents (Java en programmation pour l'un, Java en tourisme pour l'autre). Ceci se traduit par des problèmes de surcharge cognitive pour les utilisateurs. La RI personnalisée vise à améliorer les problèmes de surcharge cognitive de systèmes de RI classiques en intégrant le profil de l'utilisateur dans le mécanisme. Selon (Zimirli, 08) cette intégration peut se faire de 3 principales manières dans le processus de recherche d'information (Figure 2.11).

1. ***Sélection personnalisée de l'information :*** elle consiste à intégrer les paramètres du profil utilisateur dans la fonction de similarité entre la requête de l'utilisateur et chaque document.

2. ***Modification (reformulation ou expansion) de la requête :*** elle consiste à introduire dans la structure de la requête les termes issus du profil de l'utilisateur. C'est la méthode la plus répandue (Speretta et Gauch, 05).

3. ***Réordonnancement des résultats :*** elle consiste à utiliser les termes du profil de l'utilisateur pour réordonnancer les résultats issues d'un système de RI classique (Zayani, 08)(Kostadinov, 03).

4. Quelque que soit la technique d'intégration du profil utilisée, chacun de ces systèmes produit de meilleurs résultats qu'un système de RI classique.

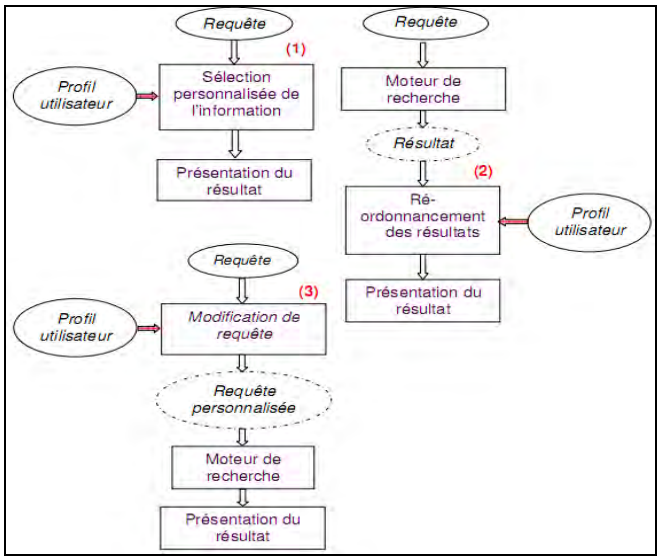



Figure 2. 11 : Phases d'intégration du profil utilisateur dans un système de RI personnalisée

Les techniques de filtrage par contenus sont très efficaces dans les domaines où les contenus de documents sont majoritairement textuels. Elles ne sont pas appropriées pour les contenus multimédias (images, vidéos, sons) (Mobasher, 07).

## 2.7.3 Filtrage collaboratif (*collaborative filtering*)

Le filtrage collaboratif est une technique complémentaire au filtrage par contenus (Das et al., 07), qui utilise les profils des individus ayant un profil similaire à celui de l'utilisateur à un instant donné t, pour recommander de l'information à l'utilisateur à un instant ultérieur t+Δt (Figure 2.12).

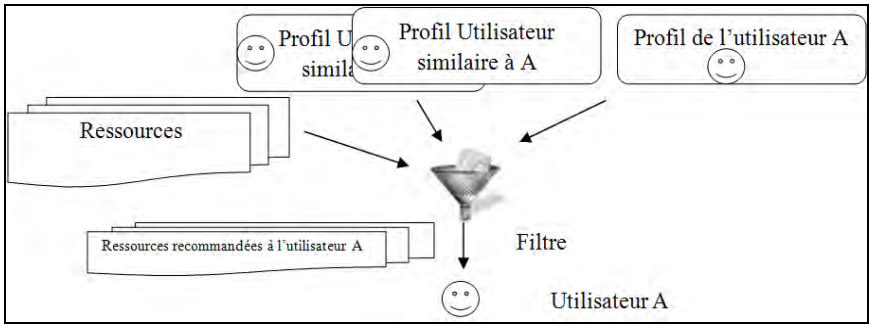



Figure 2. 12 : Filtrage collaboratif (centré utilisateur)

Le filtrage collaboratif est très utilisé dans les environnements de type e-commerce dans lesquels les utilisateurs attribuent des scores à des produits qu'ils consultent ou achètent. L'approche consiste alors à faire des recommandations en recherchant des corrélations entre les produits « aimés » et « pas aimés » parmi les utilisateurs du système. Dans un système de recommandation de livres par exemple, le système recherchera les individus similaires à l'utilisateur ; et seuls les livres bien notés par ces individus seront recommandés à l'utilisateur.

**Le filtrage collaboratif (centré utilisateur)**[24] est celui qui est le plus utilisé dans la littérature (Konstan et al., 97)(Li et al., 05). Cependant pour le mettre en œuvre dans ce type de filtrage, le système doit stocker et maintenir une matrice (Utilisateurs*Items) contenant les scores attribués par chaque utilisateur (U) sur tous les items[25] (I) du système. La complexité des calculs sur cette matrice croit linéairement avec le nombre d'items et le nombre d'utilisateurs du système. Cette matrice est aussi généralement très éparse car le ratio entre le nombre total d'items du système et le nombre d'items notés par un utilisateur est très faible. La complexité de calcul de cette matrice et le fait qu'elle soit généralement très éparse pose de nombreux problèmes lors du passage à l'échelle et lors des stockages en mémoire (Mobasher, 07)(Das et al., 07). Pour pallier ces faiblesses, plusieurs stratégies d'optimisation ont été proposées (Das et al., 07). Elles incluent la réduction des dimensions de la matrice, l'indexation des similarités entre utilisateurs (matrices Utilisateurs-Utilisateurs), et le clustering (hors-ligne) des utilisateurs. Comme autre solution, on peut avoir recours à une extension du filtrage collaboratif centré sur l'utilisateur appelée filtrage collaboratif centré sur les items[26] (Kitts et al., 00).

**Le filtrage collaboratif centré sur les items** utilise une matrice Items*Items plutôt qu'une matrice Utilisateurs*Items ou Utilisateurs*Utilisateurs pour le calcul des items à recommander à l'utilisateur (Figure 2.13). Les items similaires aux items dont l'utilisateur a déjà attribué des scores important lui sont recommandés. Cette technique peut être beaucoup plus rapide car la matrice Items*Items peut être calculée à l'avance et varie très peu. De plus, des travaux montrent que les résultats de cette technique sont comparables à ceux obtenus par le filtrage collaboratif centré utilisateur (Li et al., 05).

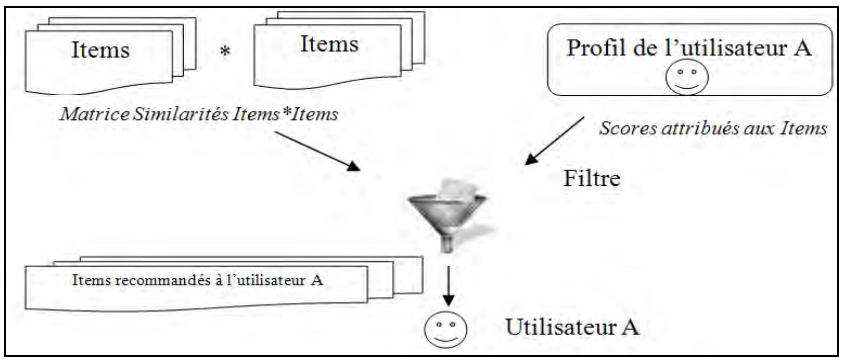



Figure 2. 13 : Filtrage collaboratif (centré sur les Items)

Plusieurs travaux catégorisent aussi le filtrage collaboratif suivant les techniques utilisées pour estimer les scores des utilisateurs sur les items (Das et al., 07)(Hofman, 04)(Adamavicius et Tuzhilin., 05) : le filtrage collaboratif basé sur la mémoire (*memory based collaborative filtering)* et le filtrage collaboratif basé sur les modèles (*model-based collaborative filtering*). La principale différence est que le filtrage collaboratif basé sur les modèles prédit les scores non pas à partir des heuristiques, mais à partir des modèles appris par des algorithmes de fouille de données ou de statistiques (Adamavicius et al., 05). Le filtrage collaboratif basé sur la mémoire est suffisant pour de nombreux cas concrets de problèmes, cependant ils ont quelques inconvénients (Hofman., 04) : (a) la précision des résultats obtenus n'est pas optimale, (b) les profils utilisateurs ne sont pas

[24] User-Based collaborative filtering en anglais

[25] Un article à vendre dans un site de e-commerce par exemple

[26] Items-Based collaborative filtering en anglais

utilisés et rien ne peut être réutilisé car aucun modèle n'est construit, (c) le passage à l'échelle est difficile car la technique nécessite trop de mémoire, (d) il est difficile de catégoriser ces techniques pour des tâches précises. Les techniques de filtrage collaboratifs basés sur les modèles peuvent pallier ces inconvénients : (a) les prédictions sont plus précises, (b) les données sont analysées suivant des modèles plus légers qui permettent d'identifier automatiquement des communautés d'utilisateurs, (c) les calculs peuvent être faits en temps réels, (d) on a plus de flexibilité à définir les modèles pour des tâches bien précises.

Quelque soit la technique de filtrage collaboratif, un problème majeur demeure : c'est celui du démarrage à froid (*cold start problem*) (Massa et Avesani, 04) (Massa et Avesani, 07) (Massa et Avesani, 09). Ce problème se pose pour les utilisateurs ayant un profil vide car n'ayant pas encore attribué des scores sur des items (nouveaux utilisateurs par exemple). Leur profil étant vide, il n'est pas possible de trouver des utilisateurs qui leur sont similaires, et donc impossible d'inférer des recommandations. Le même problème se pose pour les items ayant très peu ou pas de scores attribués par les utilisateurs (nouveaux items par exemple). Nous reviendrons sur cette problématique dans le chapitre suivant.

### 2.7.4 Filtrage hybride *(hybrid filtering)*

Le filtrage par contenus et le filtrage collaboratif étudiés précédemment possèdent chacun leurs avantages et leurs inconvénients. D'une part, le filtrage collaboratif pallie des inconvénients du filtrage par contenus tels que le manque de données subjectives (données explicites de l'utilisateur) ou de scores attribués par les utilisateurs eux-mêmes. D'autre part le filtrage par contenus pallie des inconvénients du filtrage collaboratif tel que le problème des données très éparses ou la complexité de calculs (Montaner, 03). Le filtrage hybride vise à intégrer ces deux types de filtrages en gardant les avantages de chacun (Balabanovic et al., 97)(Claypool et al., 99)(Melville et al., 02). Cette intégration peut se faire de différentes manières (Adomavicius et al., 05) : (i) implémenter chacune des méthodes séparément et intégrer leurs résultats ; (ii) rajouter certaines caractéristiques du filtrage collaboratif dans le filtrage par contenus ; (iii) rajouter certaines caractéristiques du filtrage par contenus dans le filtrage collaboratif ; (iv) construire un modèle unifié incorporant les caractéristiques de chacune des méthodes.

La dernière catégorie de filtrage s'éloigne de part son principe aux autres catégories de filtrage (collaboratif, par contenus et hybride): il s'agit du filtrage à base de règles.

### 2.7.5 Filtrage à base de règles *(rules based filtering)*

Le filtrage à base de règles est une méthode simple qui consiste à utiliser les données explicites et statiques (sexe, âge, etc.) renseignés par l'utilisateur (formulaires ou questionnaires par exemple). Des règles prédéfinies (de type *if.. then..)* sont appliquées pour sélectionner l'information pertinente pour chaque utilisateur (Liang et al., 02) (Figure 2.14). Cette approche s'appuie sur des groupes (ou classes) prédéfinies d'utilisateurs afin de définir les contenus à présenter ou les services à fournir. C'est le cas par exemple de segments de clients par catégories socioprofessionnelle en marketing. Ce sont les règles définies par les experts du domaine (e-learning, marketing, etc.) qui représentent le centre de cette approche. La difficulté réside donc dans

l'établissement et la validation de ces règles. Cette approche pose également des problèmes de maintenance des règles qui peuvent évoluer (Kim, 01). De plus, les données considérés dans les profils utilisateurs sont subjectives et peuvent donc être biaisées (Mobasher, 07).

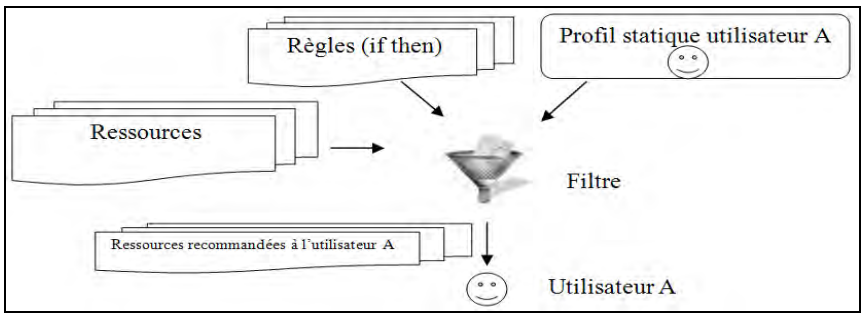



Figure 2. 14 : Filtrage à base de règles

Le tableau 2.3 récapitule chacune des techniques de filtrage présentées dans cette section.

# 2.8 Conclusion

Au travers des travaux de la littérature, nous avons présenté dans ce chapitre les principaux concepts et techniques qui entrent en jeu dans le développement des profils utilisateurs et dans leur usage dans les systèmes adaptatifs. Le cycle de développement et d'usage des profils utilisateurs étant similaire un processus classique d'extraction de connaissances à partir de données, nous avons structuré cette présentation suivant les différentes phases de ce processus.

| Approches | Contexte Applicatif | Utilisation du profil | Avantages | Inconvénients | Références |
|---|---|---|---|---|---|
| **Filtrage à base de règles** | Services personnalisés basiques | Matching entre profils explicites statiques (sexe, âge, etc.) et une base de règles prédéfinies | Facilité d'usage des règles simples *if then*… | Difficulté à définir et maintenir les règles | MyLibrary (Di Giacomo et al., 01). |
| **Filtrage par contenus** | Domaine à fort usage de textes | Matching entre les éléments du profil de l'utilisateur et les éléments caractéristiques des ressources. | Très adapté aux domaines à fort usage de textes | Pas adapté aux données multimédias. Profils utilisateurs peu précis. | (Zayani, 08) (Zemirli, 08) (Daoud et al., 09). |
| **Filtrage collaboratif** | Domaine pas fortement textuel avec beaucoup de feedback explicite utilisateur | Matching entre les profils des individus similaires à l'utilisateur et les ressources. | Très utilisé dans le e-commerce ; profils utilisateurs plus précis ; va au-delà des domaines textuels | Problème de démarrage à froid (*cold start problem*) | Amazon (Linden et al., 03 ; Das et al., 07). |
| **Filtrage hybride** | Services avancés personnalisés | Combinaison du filtrage collaboratif et du filtrage par contenus. | Combine les avantages du filtrage collaboratif et du filtrage par contenus | Peu être très compliqué à mettre en œuvre | Explanet (Masters et al., 08). |

Tableau 2. 3 : Résumé des techniques des systèmes adaptatifs

Dans la phase de collecte de données, deux concepts majeurs sont à prendre en compte : les producteurs de données et les sources de données. Le concept de producteur de données soulève plusieurs problématiques dont la plus importante est le choix efficace des individus à partir desquels le profil de l'utilisateur peut être dérivé. Le concept de source de données soulève des problématiques liées à la multiplicité et la fiabilité des sources de données.

Dans la phase de prétraitement de données, les techniques classiques existantes peuvent être utilisées : réduction de données, transformation de données, transcodage de données, transmodage de données, etc.

Dans la phase de préparation de données, les données qui seront utilisées par les algorithmes de fouille de données peuvent être idéalement structurées suivant quatre concepts : les données explicites (fournies explicitement par l'utilisateur et qui peuvent être réutilisées telles quels par les mécanismes d'usage du profil) ; les données implicites (recueillies de manière implicites à partir des interactions de l'utilisateur) qui vont permettre de définir des indicateurs permettant de déduire les intérêts de l'utilisateur ; les données de contexte qui vont permettre un meilleur usage en fonction des contextes des profils construits ; les données sémantiques qui vont permettre de lever les ambigüités sémantiques sur les données implicites utilisées pour construire les profils.

Dans la phase de fouille de données, plusieurs familles d'algorithmes peuvent être utilisées en fonction du type de modèle de profil à construire (modélisation du comportement, modélisation des centres d'intérêts, modélisation des intentions). La modélisation des centres d'intérêts est généralement préalable à la modélisation des intentions ou la modélisation comportementale. Les centres d'intérêts construits peuvent être représentés sous forme ensembliste (vecteurs de termes pondérés représentant les centres d'intérêts de l'utilisateur), sous forme connexionniste (réseaux sémantiques) ou sous forme conceptuelle (ontologies).

Dans la phase d'interprétation et d'évaluation qui correspond dans notre cas à l'usage de profils dans les systèmes adaptatifs, nous avons distingué quatre principales approches dans la littérature. Les deux approches les plus importantes et les plus utilisées sont le filtrage par contenus et le filtrage collaboratif. Ces deux approchent souffrent toutefois d'un problème majeur : celui du démarrage à froid. Il n'est pas possible d'adapter l'information pour les utilisateurs ayant des profils très pauvres ou vides (nouveaux utilisateurs par exemple). Dans la littérature, les approches qui visent à enrichir ce type de profils s'appuient sur l'un ou l'autre des concepts de la phase de collecte de données (sources de données et producteurs de données).

Des sources de données externes à une application peuvent être utilisées pour rechercher d'autres données utiles pour le développement du profil de l'utilisateur. On parle alors de développement de profils utilisateurs multi-applications (Viviani et al., 10) qui sont cependant très difficiles à mettre en œuvre dans un cadre comme le Web avec les problématiques de fiabilité et de gestion des identités numériques multiples (cf. section 2.2.2).

Les producteurs de données tels que les individus du réseau social de l'utilisateur sont par contre facilement accessibles avec l'avènement du Web 2.0. De manière intuitive, les profils de ces individus sont de plus en plus utilisés pour l'amélioration des systèmes adaptatifs. Ceci a donné lieu à une nouvelle approche de filtrage de l'information : *le filtrage social,* qui peut être perçue comme une extension du filtrage collaboratif. Au lieu de s'appuyer uniquement sur les individus similaires à l'utilisateur, le filtrage social s'intéresse particulièrement aux individus du réseau social de l'utilisateur qui non seulement lui seraient similaires, mais qui seraient aussi

les plus susceptibles de l'influencer (et donc d'influencer son profil). La question principale qui se pose dans les systèmes de filtrage social est celle de savoir comment exploiter efficacement le réseau social de l'utilisateur pour enrichir les mécanismes d'adaptation de l'information à l'utilisateur. Plusieurs travaux s'intéressent à cette problématique qui constitue également la motivation principale de cette thèse. Nous les présentons dans le chapitre suivant.

# 3 Chapitre 2 : Filtrage social de l'information et éléments de l'analyse des réseaux sociaux



## 3.1 Introduction

Dans le chapitre précédent nous avons soulevé le problème de démarrage à froid dans les techniques de filtrage d'information des systèmes adaptatifs. Le problème de démarrage à froid peut être étendu à celui de l'enrichissement des profils utilisateurs qui consiste à rechercher tout centre d'intérêt non connu dans le profil de l'utilisateur à un instant t. Les approches de solution existantes visant à résoudre ces problèmes s'appuient principalement sur le réseau social de l'utilisateur. Dans la suite de ce document, nous appellerons filtrage social, les mécanismes de filtrage d'information qui s'appuient sur le réseau social de l'utilisateur. L'expression filtrage social est souvent utilisée dans la littérature pour désigner le filtrage collaboratif. Cependant, dans ce mémoire, nous considérons le sens premier de cette expression qui implique forcément l'usage d'un réseau social à la différence de l'utilisation des utilisateurs similaires dans le filtrage collaboratif. Nous retenons la définition classique d'un réseau social qui désigne un ensemble d'individus (ou organisations) reliés entre eux par des liens créés lors d'interactions sociales.

Les premières approches s'orientant vers le filtrage social de l'information ont consisté à étendre les systèmes collaboratifs en y intégrant les réseaux sociaux (Kautz et al., 97) (Massa et Avesani, 04) (Massa et Avesani, 07) (Massa et Avesani, 09). De nos jours les techniques de filtrage social de l'information se distinguent complètement des techniques de filtrage collaboratif et ont donné naissance à de nouveaux champs de recherche, notamment les systèmes de recommandations sociaux (Amatrian et al., 10) et les systèmes de recherche d'information sociale (Boughanem et al., 10). Toutefois la nature des réseaux sociaux ainsi que les techniques d'exploitation de ces réseaux diffèrent considérablement en fonction des contextes de travail et de la nature de liens sociaux considérés. Ces champs de recherche sont assez récents et n'ont réellement pu voir le jour qu'avec l'avènement du Web 2.0 et la disponibilité de données sociales sur le Web (multiplication de création de liens et d'interactions sociales entre utilisateurs via les environnements tels que les réseaux sociaux numériques). Les travaux de recherche associés exploitent encore très peu la richesse des observations et des résultats de l'analyse des réseaux sociaux obtenus en sciences sociales ou dans les éléments de la théorie des graphes depuis les années 1930 (Breslin et Decker, 07).

Dans ce chapitre nous étudions ces deux aspects de l'analyse de réseaux sociaux (filtrage social et éléments de l'analyse des réseaux) relatifs à notre problématique suivant deux grandes parties. Dans la première partie, nous présentons les principales techniques actuelles de filtrage social de l'information (utiles pour pallier le problème de démarrage à froid et d'enrichissement des profils utilisateurs). Dans la seconde partie nous étudions les éléments plus théoriques de l'analyse des réseaux sociaux (sociologie, sociométrie, théorie des graphes) qui peuvent aider à améliorer les travaux actuels sur le filtrage social de l'information.

# 3.2 Filtrage social de l'information

Nous regroupons les travaux de la littérature relatifs au filtrage social de l'information en trois parties : extension du filtrage collaboratif au filtrage social, systèmes de recommandation sociaux, systèmes de recherche d'information sociaux.

## 3.2.1 Du filtrage collaboratif au filtrage social

Le filtrage collaboratif possède des inconvénients dont les plus importants sont les suivants (Massa et Avesani, 07) :

- ***Le démarrage à froid :*** que nous avons présenté dans le chapitre précédent.

- ***Les résultats non satisfaisants ou non optimaux :*** la plupart des utilisateurs attribuent des scores à très peu d'items parmi les millions (voire milliards) que comptent très souvent les systèmes (matrices Utilisateurs*Items très éparses). Il devient alors difficile de trouver des individus similaires à un utilisateur ayant un profil comportant très peu d'informations. La sélection des individus similaires à partir desquelles les recommandations seront déduites devient alors peu précise et ceci peut conduire à des recommandations de mauvaise qualité. Certains profils d'individus du système potentiellement pertinents pour l'utilisateur ne sont pas exploités, car leurs profils comportent également très peu d'informations.

- ***La complexité élevée des calculs des algorithmes de recherche des individus similaires à l'utilisateur :*** le profil de l'utilisateur doit en général être comparé à celui de chaque utilisateur du système. Les temps de calculs deviennent énormes pour des systèmes ayant des millions d'utilisateurs par exemple (Amazon, Web semantic, etc.) (Berners-Lee et al., 01). Ceci impose en général de calculer et de stocker en « off » de manière périodique les similarités entre profils, impliquent de ce fait que les recommandations réalisées ne sont pas toujours à jour par rapport aux profils utilisateurs.

- ***Les attaques par des utilisateurs malicieux :*** le filtrage collaboratif est généralement utilisé dans les sites de e-commerce (Amazon par exemple) : dans ces contextes, un utilisateur peut être tenté d'influencer les recommandations faites à d'autres utilisateurs (O'Mahony et al., 03) (O'Mahony et al., 05). Par exemple, un utilisateur peut « forcer » Amazon à recommander le livre qu'il a écrit à d'autres utilisateurs. Une manière très simple serait que cet utilisateur (sous une fausse identité) copie le profil (scores attribués aux items) d'un autre utilisateur qui achète des livres. Ainsi, le système va penser que cet utilisateur malicieux est très similaire à l'utilisateur correspondant au profil copié, et par conséquent fournir des recommandations en se basant sur le profil intégral de l'utilisateur malicieux. Dans les systèmes à serveurs centralisés comme les systèmes actuels, créer de fausses identités (profils) nécessite beaucoup de temps. De ce fait, ce type d'attaque ne constitue pas une réelle menace. Toutefois, avec un usage de plus en plus important de serveurs et de profils décentralisés tels que RVW (Eaton, 04) ou FOAF (Goldbeck, 03), ce type d'attaque pourrait être très facile à déployer.

Pour faire face à ces inconvénients (Kautz et al., 97) furent parmi les précurseurs à motiver l'usage du réseau social de l'utilisateur afin de réduire la taille et la pertinence des individus à partir desquels les recommandations seront faites à l'utilisateur. Plusieurs types de réseaux sociaux sont généralement considérés en fonction des contextes d'études : réseaux de collègues ou personnes familières (Kautz et al., 97)(Guy et al., 10), réseaux d'amis (Guy et al., 09)(Said et al., 10), réseaux de confiance (Massa et Avesani, 04) (Massa et Avesani, 07) (Massa et Avesani, 09)(Goldbeck, 06)(O'Donovan et al., 07) etc.

(Massa et Avesani, 04)(Massa et Avesani, 07) (Massa et Avesani, 09) proposent par exemple de rajouter au processus classique de filtrage collaboratif des réseaux de confiance entre utilisateurs (Figure 3.1).

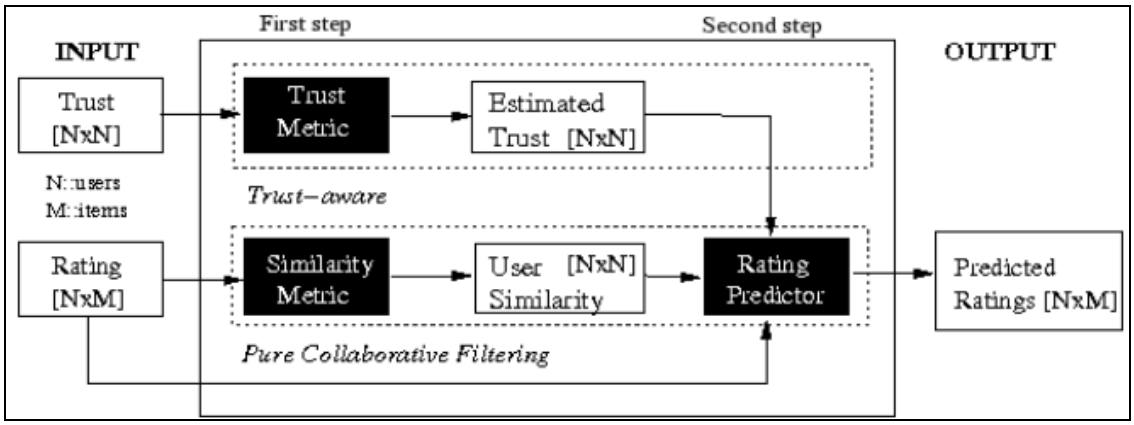



Figure 3. 1 : Exemple d'extension du filtrage collaboratif avec les réseaux de confiance

Le filtrage collaboratif classique utilise uniquement la matrice *Rating* NxM dans laquelle sont stockés les scores attribués par les utilisateurs aux items (profils utilisateurs dans ce cas). L'application d'une fonction de

similarité (module *Similarity Metric*) sur cette matrice permet d'obtenir la matrice *User Similarity* NxN de similarité entre les utilisateurs. Le coefficient de similarité de Pearson est très souvent utilisé comme fonction de similarité (Herlocker et al., 99). Pour deux utilisateurs u1 et u2, ce coefficient est calculé selon la formule (3) dans laquelle *i* représente un item, *m* représente le nombre total d'items, $r_{u1,i}$ représente le score attribué par u1 sur l'item i et $\bar{r}_{u1}$ représente la moyenne de scores attribués par u1.

$$S_{u1,u2} = \frac{\sum_{i=1}^{m} (r_{u1,i} - \bar{r}_{u1})(r_{u2,i} - \bar{r}_{u2})}{\sqrt{\sum_{i=1}^{m} (r_{u1,i} - \bar{r}_{u1})^2 \sum_{i=1}^{m} (r_{u2,i} - \bar{r}_{u2})^2}} \quad \textbf{(3)}$$

La matrice de confiance entre utilisateurs *Trust* NxN contient les scores de degré de confiance exprimés de manière explicite entre utilisateurs. Des sites de e-commerce comme Epinions.com, Ebay ou Amazon offrent cette fonctionnalité à leurs utilisateurs. Le graphe orienté de la figure 3.2 est un exemple de réseau de confiance dans lequel l'utilisateur C a indiqué ne pas faire confiance à l'utilisateur B (valeur 0) et l'utilisateur B a indiqué avoir pleinement confiance à l'utilisateur A (valeur 1). Un tel graphe de confiance n'est pas symétrique : par exemple, B a pleinement confiance à A (valeur 1), mais le contraire n'est pas entièrement vrai (valeur de 0.4 seulement). Dans la mesure où chaque utilisateur n'exprime sa confiance qu'à un nombre très souvent limité d'utilisateurs autour de lui, beaucoup de liens de confiance doivent être estimés à partir de la matrice initiale *Trust* NxN afin d'avoir une matrice de confiance peu éparse (*Estimated Trust* NxN). Pour ce faire, le module *Trust Metric* utilise une métrique de confiance pour estimer des liens inexistants dans la matrice initiale NxN. Dans la figure 3.2, il s'agira par exemple d'estimer la confiance de A vers D. Plusieurs métriques de confiance existent dans la littérature pour ce type de réseau. (Massa et Avesani, 04)(Goldbeck, 06)(Massa et Avesani, 07) utilisent par exemple une métrique de confiance qui est fonction de la distance *n* entre deux utilisateurs et d'une distance maximum de propagation de confiance *d*, $T_M = (d - n + 1)/d$. Si d=4 par exemple, ceci implique que la confiance estimée entre tout couple d'utilisateurs dont la distance est >= 5 vaudra 0). Pour le cas de la figure 2, si d=4, la confiance estimée de A vers D vaudra (4-2+1)/4 = 0.75 dans la matrice *Estimated Trust* NxN.

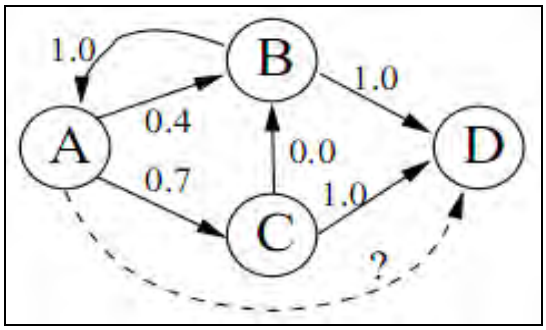



Figure 3. 2 : Exemple simplifié de réseau de confiance

A partir de la matrice de similarité *User Similarity NxN* (seule matrice utilisée dans le filtrage collaboratif classique) et de la matrice de confiance *Estimated Trust NxN*, le module de prédiction des scores utilisateurs (*Rating Predictor* qui enrichit les profils) utilise des fonctions de prédictions classiques (Herlocker et al., 99) telle que celle définie dans la formule (4).

$P_{u,i}$ est le score prédit de l'utilisateur u sur l'item i, $k$ est le nombre d'utilisateurs similaires à u et à qui il fait confiance (calculé à partir de seuils prédéfinis au travers de la matrice de similarité et de confiance estimée), $S_{u,v}$ est la similarité entre u et v définie dans la formule (4).

$$P_{u,i} = \bar{r}_u + \frac{\sum_{v=1}^{k} S_{u,v} (r_{u,i} - \bar{r}_u)}{\sum_{v=1}^{k} S_{u,v}} \quad \textbf{(4)}$$

Les résultats de l'expérimentation menée par (Massa et Avesani, 07) en utilisant ce processus démontrent une meilleure qualité des recommandations et une complexité de calculs réduite par rapport au processus de filtrage collaboratif classique. L'usage de la matrice de confiance en plus de la matrice de similarité permet de ne conserver que les individus les plus pertinents pour l'utilisateur lors de l'enrichissement de son profil. Si l'utilisateur a un profil vide, la matrice de confiance sera utilisée toute seule pour l'enrichissement de son profil (le problème de démarrage a froid peut ainsi être en partie résolue). Le réseau social de confiance permet également d'éviter l'usage des profils malicieux dans le mécanisme (peu d'utilisateurs exprimeront leur confiance envers ces profils). Toutefois, il n'est pas toujours évident pour un système d'acquérir les données de confiance entre les utilisateurs.

D'autres études empiriques ont également montré l'intérêt d'intégrer d'autres types de réseaux sociaux au filtrage collaboratif. Même si la taille de l'échantillon d'utilisateurs est réduite dans l'expérimentation de (Sinha et al., 01), leurs résultats montrent que les recommandations de livres et de films réalisées par Amazon, RatingZone[27] & Sleeper, MovieCritic, Reel.com[28]) sont de moins bonne qualité que les recommandations faites par les amis (dans la vie réelle) des utilisateurs. Toutefois, les films et livres recommandés par les sites en ligne fournissent plus de nouveautés et de résultats inattendus par rapport à ceux des amis. Dans une étude de recommandation de clubs à partir d'un réseau social numérique allemand, (Groh et al., 07) montrent que les scores attribués aux clubs par les utilisateurs sont statistiquement dépendants des scores des utilisateurs de leur réseau social, et que l'usage du réseau social de l'utilisateur dans le processus de recommandation donne de meilleurs résultats que le filtrage collaboratif classique. Dans le même ordre d'idée, (Said et al., 10) montrent que le réseau social des utilisateurs impacte leurs choix dans l'attribution des scores sur des films en ligne et que ce réseau peut être utilisé pour améliorer la qualité des recommandations. (Konstas et al., 09) montrent que la recommandation d'albums de musique dans le site *lastfm* par le filtrage collaboratif est considérablement améliorée par l'usage des réseaux d'amis ainsi que les annotations faites par ces derniers sur les musiques dans ce même site. Au travers d'une expérimentation sur la recommandation de films dans laquelle les raisons de chaque recommandation sont clairement expliquées aux utilisateurs, (Bonhard et al., 06) montrent l'intérêt d'intégrer les réseaux sociaux de personnes familières (connus dans la vie réelle) et de personnes similaires (mêmes caractéristiques démographiques). Une expérimentation assez similaire est proposée par (Guy et al., 09) dans un contexte de recommandation de ressources à partir d'interactions entre utilisateurs dans des outils collaboratifs d'un intranet d'entreprise.

[27] http://www.ratingzone.com/

[28] http://www.reel.com/

Tous ces travaux montrent l'intérêt de l'usage des réseaux dans les systèmes de recommandation. Toutefois, plusieurs travaux se sont uniquement intéressés au filtrage social. Nous les présentons dans les deux sections suivantes.

## 3.2.2 Systèmes de recommandation sociaux (*Social Recommander Systems*)

### *3.2.2.1 Usage des réseaux de similarité et de familiarité*

Les systèmes de filtrage collaboratif intégrant les réseaux de confiance entre utilisateurs furent les premières approches de systèmes de recommandation sociaux. Dans un sens beaucoup plus large, les systèmes de recommandation sociaux s'appuient sur différents types de réseaux sociaux extraits des médias sociaux ou d'outils collaboratifs tels que les réseaux sociaux numériques, les sites de social *bookmarking*, les wikis, les blogs, etc. (Guy et al., 08). Les réseaux sociaux exploités peuvent être déduits des interactions directes entre utilisateurs (score de confiance, amitié, tags entre utilisateurs, organigramme d'entreprise, etc.) ou des interactions entre utilisateurs et ressources (co-auteur d'un même article, commentaires ou tags sur une même page de wiki, etc.) tels que proposé par (Guy et al., 08)(Guy et al., 10) (figure 3.3). Au-delà des scores (données explicites) qui sont utilisés dans le filtrage collaboratif, les systèmes de recommandations sociaux s'appuient également sur l'analyse des contenus (données implicites) des interactions entre utilisateurs sur les médias sociaux (commentaires, tags, etc.) (Siersdorfer et al., 09). Des sites comme StumbleUpon[29] proposent par exemple des recommandations de pages web à partir des similarités des scores attribués par les utilisateurs, les scores attribués par les amis, et les centres d'intérêts de l'utilisateur et de ses amis sélectionnés dans une liste de près de 500 domaines d'intérêts.

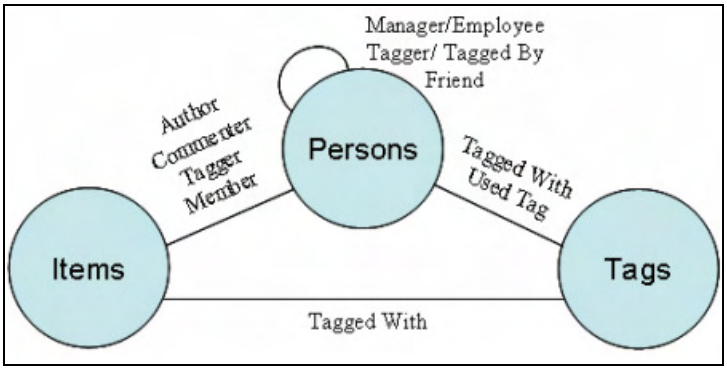



Figure 3. 3 : Exemple d'interactions directes et indirectes pour construction d'un réseau social

A travers des interactions présentes sur la figure 3.3, (Guy et al., 09) distinguent par exemple trois types de réseaux sociaux :

- ***Les réseaux sociaux de familiarité :*** qui sont extraits à partir des relations directes (amis, collègues, etc.) d'un utilisateur (personnes que l'utilisateur connaît réellement).

[29] www.stumbleupon.com

- ***Les réseaux sociaux de similarité :*** qui sont extraits à partir des activités sociales des utilisateurs qui se recouvrent (Co-écriture d'au moins un article, tags sur au moins 5 ressources communes, etc.). Par rapport au filtrage collaboratif, la similarité ici n'est pas calculée à partir des scores, mais à partir des interactions utilisateurs sur des mêmes ressources dans des médias sociaux (logs, wikis, *social bookmarking*, réseaux sociaux numériques, etc.).

- ***Les réseaux sociaux entiers :*** qui combinent les réseaux de similarité et les réseaux de familiarité.

Afin de déterminer quel réseau social serait le plus approprié dans un système de recommandation sociale, (Guy et al., 09) réalisent une expérimentation dans laquelle pour un utilisateur *u*, et un item *i*, le poids prédit en fonction du réseau social de *u* sur l'item *i* est déterminé par la formule (5).

$$RS(u,i) = e^{-\alpha d(i)} . \sum_{v \in N^T(u)} S^T[u,v] \sum_{r \in R(v,i)} W(r) \quad \textbf{(5)}$$

Dans cette formule :

- $T$ représente le type de réseau considéré, T∈ {réseau social de familiarité de *u*, réseau social de similarité de *u*, réseau social entier de *u*} ;

- $N^T(u)$ représente l'ensemble des individus considérés dans le réseau social de u (les 30 individus directement connectés à *u* et dont le poids de relation sont les plus importants).

- $S^T[u,v]$ est le poids de la relation entre l'utilisateur *u* et l'utilisateur *v* de son réseau social, en fonction du réseau social déterminé par T (Guy et al., 08).

- $R(v,i)$ est l'ensemble de toutes les relations (co-auteur, commentateur, etc.) existantes entre l'utilisateur *v* et l'item *i*.

- $W(r)\ r \in R(v,i)$ est le poids de la relation entre l'utilisateur *v* et l'item *i*. Ce poids dépend de la relation R entre *v* et *i*. Pour l'écriture d'un article (R= « écriture », i= « article »), $W(r)$ vaut par exemple 0.6, alors qu'un commentaire sur un article (R= « commentaire », i= « article»), $W(r)$ vaut plutôt 0.4 (Guy et al., 08).

Dans cette formule, $RS(u, i)$ dépend d'une part du poids des liens entre *u* et chacun des individus *v* de son réseau social ( $\sum_{v \in N^T(u)} S^T[u,v]$ ), et d'autre part des activités (écriture d'article, appartenance à des groupes, commentaire sur des blogs, etc.) de chacun des ces individus ( $\sum_{r \in R(v,i)} W(r)$ ). Le facteur $e^{-\alpha d(i)}$ est une constante qui permet de donner une priorité aux items créés récemment ($d(i)$ est le nombre de jours depuis la création de l'item *i*, *a* est un paramètre valant 0.025 dans cette expérimentation (Guy et al., 08)).

Par exemple si on considère l'exemple suivant :

- *i* est un article écrit sur un blog, *i* est créé depuis un jour (d(i)=1).

- Le réseau social (de familiarité) de u comprend deux individus *v1* et *v2*, v1 et v2 sont liés à l'article *i* : v1 l'a créé, v2 l'a commenté.

- Le poids du lien entre u et v1 vaut 0.4, et le poids du lien entre u et v2 vaut 0.5.

- Le poids prédit entre l'utilisateur *u* et l'article *i* en fonction du réseau social de familiarité de *u* est alors déterminé par :*RS(u, i)*= exp(-0.025*1)*[(04*0.6)+(0.5*0.3)]. Pour rappel, W(r) vaut 0.6 pour la création d'un article, et 0.3 pour le commentaire d'un article.

A travers une expérimentation sur 290 utilisateurs qui ont indiqué de manière explicite des feedback sur les recommandations qui leur ont été faites, (Guy et al., 09) évaluent quel serait le meilleur réseau social (parmi les trois listés précédemment) à exploiter pour recommander des informations les plus pertinentes possibles à l'utilisateur.

Les résultats de cette expérimentation montrent que les réseaux sociaux de familiarité permettent d'obtenir de meilleures recommandations que les réseaux sociaux de similarité et le réseau social entier. Dans cette expérimentation, seules les relations entre personnes, et entre personnes et items (figure 3.3) sont utilisées dans (Guy et al., 09). Les contenus (description textuelle par exemple) des items (articles, commentaires, groupes, etc.) ne sont pas pris en compte.

Pour aller plus loin, (Guy et al., 10) intègrent les contenus des tags (figure 3.3) en plus des relations entre personnes et items, pour prédire les scores (poids) entre utilisateurs et items en fonction de leur réseau social. Trois types de tags sont considérés : (i) les tags (directs) utilisés par l'utilisateur lui-même sur les items (exemple d'un tag sur un article) ; (ii) les tags (directs) sur l'utilisateur effectués par d'autres individus (exemple d'un tag pour décrire une autre personne) ; (iii) les tags (indirects) effectués par d'autres individus sur les items liés à l'utilisateur (exemple les tags effectués par d'autres personnes sur un article créé par l'utilisateur). L'expérimentation vise à analyser les contenus de tous ces tags afin de prédire avec plus de précision le profil de l'utilisateur. Le poids prédit de l'utilisateur *u* sur chaque item *i* est calculé par la formule (6) :

$$RS(u,i) = e^{-\alpha d(i)} * [\beta \sum_{v \in N(u)} w(u,v) * w(v,i) + (1-\beta) \sum_{t \in T(u)} w(u,t) * w(t,i)] \qquad \textbf{(6)}$$

Dans cette formule :

- *N(u)* représente les individus du réseau social de familiarité de *u* (30 individus ayant les liens les plus importants avec l'utilisateur).

- *w(u, v)* est le poids de la relation entre l'utilisateur *u* et le membre *v* de son réseau social.

- *w(v, i)* est le poids de la relation entre l'individu *v* et l'item *i* (cf. *W(r)* de la formule 5).

- *T(u)* est l'ensemble des termes (ou centres d'intérêts) extraits des tags de l'utilisateur *u*, donc les éléments du profil de l'utilisateur de u.

- *w(u, t)* est le poids du terme *t* dans le profil de l'utilisateur *u*.

- $w(t, i)$ est le poids du terme $t$ dans l'item $i$. $w(t,i)$ est déterminé relativement au nombre d'utilisateurs ayant appliqué le tag $t$ sur l'item $i$ (Amitay et al., 09).

Dans cette formule, le poids prédit de l'utilisateur $u$ sur un item $i$ est déterminé d'une part en fonction des individus du réseau social de u ( $\sum_{v \in N(u)} w(u,v) * w(v,i)$ ) et d'autre part en fonction des termes contenus dans les tags de l'utilisateur ( $\sum_{t \in T(u)} w(u,t) * w(t,i)$ ).

Cette double dépendance est relativisée par le paramètre $\beta$ qui est déterminé expérimentalement. Le facteur $e^{-\alpha d(i)}$ est une constante ayant la même signification que dans la formule (5) précédente.

Les résultats de l'évaluation réalisée avec cette technique montrent que la combinaison des utilisateurs du réseau social et de leurs tags permet de dériver un profil utilisateur plus pertinent comparé au cas où seul le réseau social de l'utilisateur est utilisé (Guy et al., 09). Une expérimentation assez similaire sur la recommandation de films dans le site *MovieLens* montre également l'intérêt de l'usage des profils utilisateurs construits à partir de leurs tags pour l'amélioration du filtrage collaboratif (Sen et al., 09).

### *3.2.2.2 Usage des réseaux de co-auteurs d'articles scientifiques*

Dans un tout autre contexte (Cabanac, 11) propose un système de recommandation social d'articles scientifiques aux chercheurs. Les systèmes de recommandation ordinaires dans ce contexte s'appuient généralement sur les graphes de co-auteurs (ou citations) ainsi que sur les centres d'intérêts des auteurs (McNee et al., 06)(Porcel et al., 09). Pour améliorer la qualité de ces recommandations, il rajoute à ces systèmes un autre réseau social qui permettra d'exploiter les potentielles interactions entre auteurs dans la vie réelle (participation à des conférences ou *workshops* communs) (figure 3.4). Ces deux graphes sont exploités pour définir une liste d'auteurs (*social list*) déterminée uniquement à partir d'un filtrage social. Cette liste est ensuite combinée avec une liste d'auteurs dont les profils sont similaires à l'auteur à qui le système souhaite faire des recommandations (*topical list*) afin de produire une liste finale d'auteurs à recommander (*ST list*) (figure 3.5).


Figure 3. 4 : Graphes de co-auteurs (à gauche) et de participation à des événements communs (à droite)

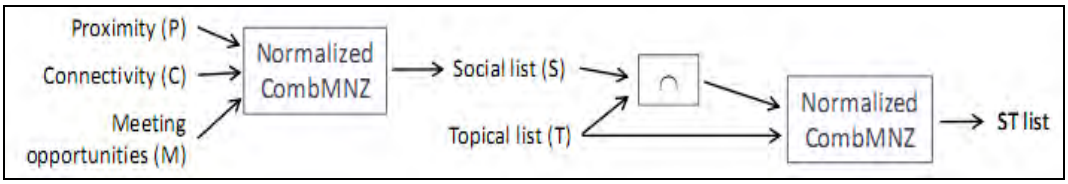



Figure 3. 5 : Combinaison de la recommandation sociale et collaborative

Pour déterminer la liste sociale, trois mesures sont exploitées à partir des graphes de la figure 3.4 : la proximité dans le graphe de co-auteurs, la connectivité dans le graphe de co-auteurs, la probabilité de rencontre dans le graphe de participation à des événements communs (tableau 3.1).

| | ***Graphe utilisé*** | ***Méthode de calcul*** | ***Aspect social à capturer*** |
|---|---|---|---|
| ***Proximité*** | Co-auteurs | Inverse de la longueur du plus court chemin entre deux chercheurs. | Proximité entre deux chercheurs |
| ***Connectivité*** | Co-auteurs | Nombre de chemins les plus courts entre deux chercheurs. | Possibilités pour un chercheur d'accéder à d'autres chercheurs par un nombre important de chemins (et donc d'intermédiaires). |
| ***Probabilité de rencontre*** | Participation événements communs | Relatif au nombre de participation à des événements communs entre deux chercheurs. | Possibilités qu'il y ait déjà eu des interactions dans la vie réelle entre deux chercheurs. |

Tableau 3. 1 : Mesures et intérêts de calcul de la liste sociale d'auteurs

Ces trois mesures sont combinées suivant les techniques de combinaisons (*CombMNZ)* de scores décrites dans (Fox et Shaw, 94) afin de définir la liste sociale d'auteurs.

Pour déterminer la liste des auteurs similaires (topical list), les profils des auteurs sont calculés par extraction des termes dans les titres de leurs publications (pondération via la mesure *tf.idf*) et comparés via la fonction de similarité *cosinus*.

Les auteurs à recommander sont finalement déterminés en combinant les deux listes suivant les techniques de combinaisons (*CombMNZ)* de scores définies dans (Fox et Shaw, 94). Nous reviendrons sur l'explication ces techniques de combinaison dans le chapitre présentant nos contributions (chapitre 4).

Les résultats de l'expérimentation réalisée montrent la pertinence des résultats de cette méthode par rapport à ceux des techniques existantes. Des auteurs comme (Ben Jabeur et al., 10) proposent des modèles de représentation de données sociales plus fins en représentant d'une part, plusieurs types de ressources (documents, auteurs, utilisateurs, tags) et d'autre part, les différentes relations qu'elles peuvent avoir (auteurs, co-auteurs, citations, références, bookmarks, *tagging*, annotations, amitiés). Bien que plus élaboré, ce type de modèles reste toutefois difficile à mettre en œuvre car toutes ces données et relations ne sont pas toujours accessibles.

## 3.2.3 Recherche d'information sociale (*Social Information Retrieval*)

### *3.2.3.1 Usage des réseaux de similarité et de familiarité*

Les techniques de recherche d'information sociale consistent à répondre aux besoins d'information des utilisateurs (exprimés généralement par des requêtes) en utilisant son profil et celui de son réseau social. De la même manière que les systèmes de recherche d'information personnalisée, l'aspect social des profils peut être

pris en compte soit dans le système d'appariement requête/ressources lui-même (Ren et al., 10) (Amitay et al., 09), soit pour l'expansion de la requête (Bender et al., 08)(Bouadjenek et al., 11), soit pour le ré-ordonnancement des résultats de la requête (Carmel et al., 09). Par exemple, dans le même contexte des travaux de (Guy et al., 08)(Guy et al., 09) (Guy et al., 10) présentés dans la section précédente, (Carmel et al., 09) propose un système de recherche d'information sociale s'appuyant sur des réseaux de similarité, des réseaux de familiarité, et des réseaux entier (similarité et familiarité) à partir de données sociales issues des outils collaboratifs (wikis, *social bookmarking*, blogs, etc.) et des relations entre collègues (organigramme d'entreprise, collaboration à des projets communs, appartenance aux mêmes groupes de travail, etc.) dans un Intranet d'entreprise (figure 3.3). De même que dans la formule (6), si *N(u)* représente les individus directement connectés à l'utilisateur dans le réseau social considéré (similarité, familiarité, ou les deux), *T(u)* représente les termes du profil de l'utilisateur extrait à partir de ces tags, et si *q* est la requête de l'utilisateur, et *i* un item correspondant à la requête de l'utilisateur, chaque résultat *i* de la requête est réordonnancé suivant le score défini dans la formule (7) (Carmel et al., 09). Le couple P(u)=<N(u),T(u)> représente le profil de l'utilisateur (qui intègre son réseau social).

$$S_p[q,i|P(u)] = \alpha S_{np}(q,i) + (1-\alpha)[\beta \sum_{v \in N(u)} w(u,v) * w(v,i) + (1-\beta) \sum_{t \in T(u)} w(u,t) * w(t,i)] \quad \textbf{(7)}$$

Dans cette formule, le score de chaque résultat potentiel *i* de la requête *q* est déterminé en fonction de trois scores :

- ***Un score non personnalisé $S_{np}(q,i)$*** qui exploite uniquement la similarité sémantique entre la requête *q* et l'item *i* (croisement entre les termes caractéristiques de la requête et ceux de l'item).

- ***Un score personnalisé en fonction du réseau social N(u) de l'utilisateur u*** ( $\sum_{v \in N(u)} w(u,v) * w(v,i)$ ). Ce score dépend du poids de la relation entre l'utilisateur *u* et chaque membre *v* de son réseau social *(w(u, v))* et du poids entre chaque individu v et l'item i *(w(v, i),* cf W(r) formule 5*)*.

- ***Un score personnalisé en fonction du profil de l'utilisateur u*** ( $\sum_{t \in T(u)} w(u,t) * w(t,i)$ ). Ce score dépend du poids du chaque terme *t* dans le profil de l'utilisateur u (*w(u,t)*) et du poids de la relation entre le terme t et l'item *i (w(t, i)*, cf. (Amitay et al., 09)*)*.

L'importance de chacun de ces scores dans le score final est relativisé les paramètres $\alpha$ et $\beta$ qui sont déterminés expérimentalement. Une double évaluation automatisée d'une part, et avec des feedback utilisateur sur les résultats renvoyés d'autre part, montre deux choses : (i) la personnalisation en utilisant le réseau social de l'utilisateur est de loin meilleure que le système non personnalisé ; (ii) quelque soit le type de réseau social considéré (similarité, familiarité, ou les deux) la personnalisation en utilisant le réseau social est de loin meilleure à la personnalisation utilisant uniquement le termes *T(u)* du profil de l'utilisateur.

### *3.2.3.2 Cas de l'usage des réseaux de co-auteurs d'articles scientifiques*

Dans un tout autre contexte, (Zeng et al., 09)(Ren et al., 10) proposent un système de recherche d'information sociale d'articles scientifiques via les réseaux de co-auteurs de la librairie DBLP. Chaque auteur est caractérisé par deux ensembles de centres d'intérêts : ses centres d'intérêts individuels calculés à partir de

ses publications (formules 8 et 9) et ses centres d'intérêts calculés à partir de son réseau social (formule 10). Ces deux ensembles de centres d'intérêts sont utilisés pour raffiner les résultats des requêtes de l'utilisateur (figure 3.6).

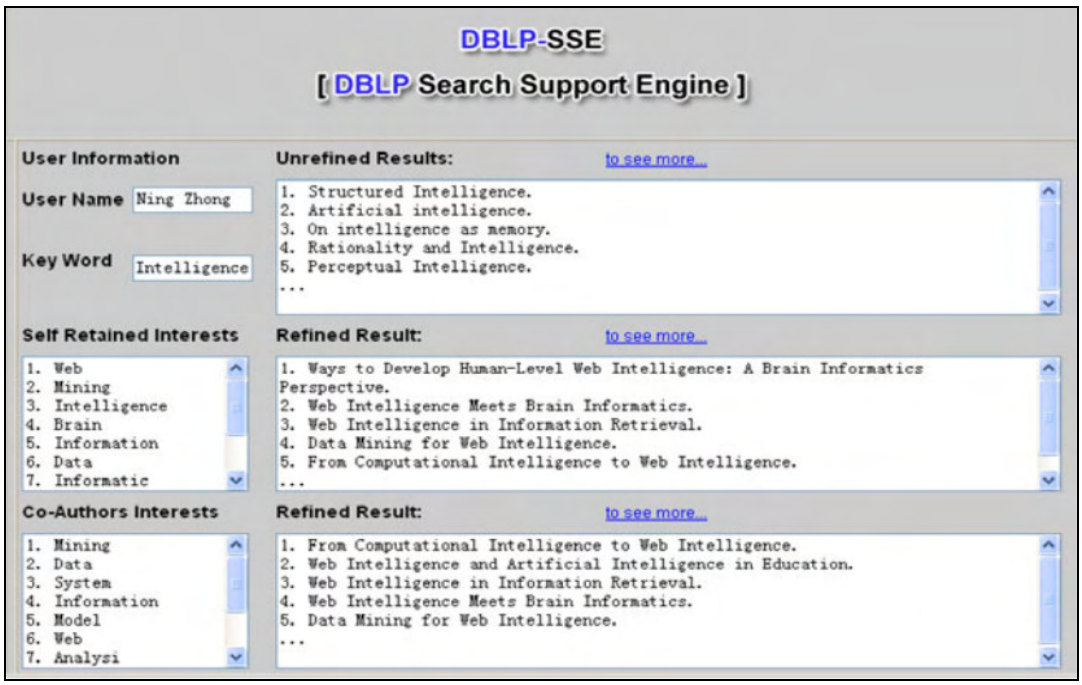



Figure 3. 6 : Recherche d'information personnalisée et sociale à partir de DBLP

Une des particularités de ces travaux est qu'ils s'intéressent également à la dynamique des centres d'intérêts des auteurs selon les formules (8) et (9) par exemple.

$$CI(t(i),n) = \sum_{j=1}^{n} y_{t(i),j} \qquad \textbf{(8)}$$

$$RI(t(i),n) = \sum_{j=1}^{n} y_{t(i),j} * AT_{t(i)}^{-b} \qquad \textbf{(9)}$$

Pour un auteur donné, si *t(i)* est un centre d'intérêt (extrait à partir des termes figurant dans les titres des publications), et $y_{t(i),j}$ est le nombre de publications relatives au centre d'intérêt t(i) pendant l'intervalle de temps j (1 an par exemple), *CI(t(i),n)* représente le poids cumulé (*Cumulative Interest*) du centre d'intérêt *t(i)* pendant *n* intervalles de temps consécutifs (10 ans par exemple). Ce poids est un poids global du centre d'intérêt sur un intervalle de temps qui peut ne pas refléter une distribution uniforme de ce centre d'intérêt pendant cet intervalle de temps. En fait, sur le moyen ou long terme, les auteurs sont parfois amenés à changer de domaine (ou centre) d'intérêts. Les domaines d'intérêts d'un auteur en début de carrière ne sont pas toujours les mêmes en fin de sa carrière par exemple. Pour prendre en compte cette évolution de centres d'intérêts (Zeng et al., 10) ont évalué la pertinence de l'usage de fonctions de rétention cognitives qui s'appuient sur des phénomènes d'oubli déjà observés en psychologie cognitive (Anderson et Schooler, 91). Selon cette évaluation, pour un intervalle de temps *j*, $y_{t(i),j} * AT_{t(i)}^{-b}$ représente la rétention totale *t(i)* pendant l'intervalle *j*. Ainsi, *RI(t(i),n)* représente le poids de rétention (*Retention Interest*) du centre d'intérêt i pendant n intervalles de temps consécutifs. Ce poids est plus important pour les centres les plus récents d'un

auteur. Les paramètres A et b déterminés expérimentalement valent 0.855 et 1.295 respectivement (Zeng et al., 09).

En plus des centres d'intérêts individuels des auteurs, ces centres d'intérêts en provenance de leur réseau social (*Group Interests, GI(t(i), u)*) sont calculés par agrégation des meilleurs centres d'intérêts de leur co-auteurs suivant la formule (10).

$$GI(t(i),u) = \sum_{c=1}^{m} E(t(i),u,c) \qquad \textbf{(10)}$$

$$E(t(i),u,c) = 1 \text{ si } t(i) \in I^{topN}{}_c \quad \text{et} \quad E(t(i),u,c) = 0 \text{ si } t(i) \notin I^{topN}{}_c$$

*m* représente le nombre de co-auteurs de l'auteur *u*, $I^{topN}{}_c$ représente les N premiers centres d'intérêts individuels du co-auteur *c*.

Même si l'évaluation de cette solution par des retours utilisateurs suivant l'interface de la figure 3.6 ne porte que sur un nombre très réduit d'utilisateurs, ils montrent que 100% des auteurs trouvent que la personnalisation (avec le profil individuel ou social) est meilleure que les résultats non personnalisés. 16,7% des participants trouvent que la personnalisation s'appuyant uniquement sur le réseau social fournit les meilleurs résultats. Ceci est intéressant dans la mesure où en cas d'absence d'information dans le profil de l'auteur, son profil social serait réellement utile et pertinent.

### 3.2.4 Synthèse

Quel que soit le type de système : extension du filtrage collaboratif avec l'usage des réseaux sociaux, systèmes de recommandations sociaux ou systèmes de recherche d'information sociale, toutes les évaluations expérimentales démontrent la pertinence de leurs résultats. Il est important de noter que le type de réseau social utilisé (sémantique du lien entre deux individus) et la manière d'exploiter le réseau social de l'utilisateur varient en fonction de chaque domaine d'application. Un comparatif des exemples de travaux présentés dans les sections précédentes figure sur le tableau 3.2.

Il ressort principalement de ce comparatif que le choix des sections du graphe représentant le réseau social de l'utilisateur, et des métriques de sélection des nœuds dans ces sections, afin d'extraire les informations dans le réseau social de l'utilisateur sont soit très basiques (tous les individus situés à distance 1 ou un sous ensemble de ces derniers), soit très complexes (calcul sur les grands graphes entiers) à mettre en œuvre. Ceci pourrait s'expliquer par le fait que ces travaux visent principalement à montrer l'efficacité de l'intégration des réseaux sociaux dans chacun des systèmes étudiés, sans toutefois mettre un accent important sur l'optimalité de l'intégration de ces réseaux. Ceci nous amène à nous intéresser à une nouvelle problématique qui consiste à rechercher les moyens les plus efficaces d'intégrer les réseaux sociaux dans les profils utilisateurs afin d'optimiser la qualité des résultats de ces systèmes. Pour ce faire, dans un premier temps il est important d'étudier les éléments existants et potentiellement réutilisables des travaux portant sur l'analyse des réseaux sociaux.

| | **Extension du filtrage collaboratif** | **Systèmes de recommandation sociaux** | **Recherche d'information sociale** |
|---|---|---|---|
| **Exemple** | (Massa et Avesani., 09) | (Guy et al., 09)(Guy et al., 10)(Cabanac, 11) | (Carmel et al., 09)(Ren et al., 10) |
| **Sémantique du lien** | Confiance | - Co-auteur et participation à des événements communs (Cabanac, 11).<br>- Similarité, Familiarité, et Similarité+Familiarité (Guy et al., 09)(Guy et al., 10) | - Co-auteur (Ren et al., 10)<br>- Similarité, Familiarité, et Similarité+Familiarité (Carmel et al., 09) |
| **Pondération des liens** | Oui | - Non (Cabanac, 11).<br>- Oui (Guy et al., 09)(Guy et al., 10) | - Non (Ren et al., 10).<br>- Oui (Carmel et al., 09) |
| **Section du graphe exploité** | Individus situés à distance 1 | - Graphes entiers, tous les individus (Cabanac, 11).<br>- Individus situés à distance 1 (Guy et al., 09)(Guy et al., 10). | - Individus situés à distance 1 (Ren et al., 10)<br>- Individus situés à distance 1 |
| **Métrique exploité pour la sélection des nœuds** | Coefficient de similarité de *pearson* supérieur à un seuil défini. | - Proximité, connectivité, probabilité supérieur à un seuil défini (Cabanac, 11).<br>- 30 individus ayant les liens les plus importants avec l'utilisateur (Guy et al., 09)(Guy et al., 10). | - Tous les individus à distance 1<br>- 30 individus ayant les liens les plus importants avec l'utilisateur (Carmel et al., 09) |
| **Inconvénients** | Section du graphe et sélection des nœuds pas forcément optimales ; Difficulté de collecte des données explicites de confiance entre utilisateurs. | - Complexité de calculs élevée pour des calculs sur des graphes entiers, sélection des nœuds pas forcément optimale (Cabanac et al., 11).<br>- Section du graphe et sélection des nœuds pas forcément optimales (Guy et al., 09)(Guy et al., 10). | Section du graphe et sélection des nœuds pas forcément optimales ((Ren et al., 10) (Carmel et al., 09). |

Tableau 3. 2 : Comparatif de travaux de filtrage social de l'information

## 3.3 Eléments sur l'analyse de réseaux sociaux

### 3.3.1 Préambule

Un réseau social est un ensemble d'identités sociales, telles que des individus ou encore des organisations, reliées entre eux par des liens créés lors d'interactions sociales (Wasserman et al., 94). Deux termes sont importants dans cette définition :

(i) les individus ou organisations qui sont généralement caractérisés par des propriétés (ou attributs) qui constituent pour nous les informations de leur profil.

(ii) Les liens entre les individus ou organisations qui constituent l'élément majeur donnant un sens au réseau social (interactions entre amis, collègues, contacts téléphoniques, etc.).

L'analyse des réseaux sociaux est menée dans le domaine des sciences sociales depuis les années 1930 (Breslin et Decker, 07). Cette analyse vise d'une part, à identifier les structures sociales distinctes dans les réseaux, et d'autre part, à expliquer le comportement des individus au sein de ces structures sociales, au moyen des études ethnographiques, de modèles mathématiques (théorie des graphes) ou d'éléments de la sociométrie. L'accessibilité de plus en plus grandissante des données sociales des utilisateurs avec l'explosion du Web 2.0 a ouvert la voie à des expérimentations sociales ou automatisées beaucoup plus importantes. Dans cette partie, nous nous intéressons principalement aux éléments intéressants de l'analyse des réseaux sociaux sur lesquels pourront s'appuyer nos travaux sur une exploitation des liens sociaux de l'utilisateur nécessaires à l'enrichissement de son profil. Nous présentons tout d'abord les concepts majeurs du domaine de la sociologie, puis, les points essentiels concernant les éléments mathématiques et informatiques sur lesquels nos travaux s'appuient.

## 3.3.2 Éléments sociologiques

Avant l'avènement du Web et particulièrement du Web social, l'essentiel des travaux sur l'analyse des réseaux sociaux a été mené dans les sciences sociales. Les différentes problématiques abordées dans ces études sont très vastes et sortent du cadre de cette thèse. Nous présentons ici uniquement les éléments que nous jugeons importants pour nos travaux : les analyses égocentrées, les analyses socio-centrées, la force des liens faibles, les trous structuraux et le capital social.

### *3.3.2.1 Les analyses égocentrées*

L'analyse d'un réseau social peut être centrée sur un utilisateur (analyse égocentrée) ou sur le réseau social entier (analyse socio centrée). Un réseau égocentrique (ou réseau personnel) utilisé dans l'analyse égocentrée désigne un sous-réseau constitué uniquement par un individu et l'ensemble de ses liens directs dans le graphe entier. La motivation principale vers l'étude des réseaux égocentriques est l'accès beaucoup plus facile aux données (par des enquêtes ethnographiques) comparé à l'accès beaucoup plus compliqué aux données d'un réseau social entier. Dans un réseau égocentrique, l'individu central du réseau est appelé *égo* et chacune de ses relations est un *alter.* L'étude des réseaux égocentriques est habituellement effectuée lorsque le profil de l'égo est connu, mais pas celui de ses *alters*. Ces études s'appuient donc sur les égos pour fournir de l'information sur leurs *alters*. Un réseau « boule de neige » indique que les *alters* identifiés dans une étude égocentrique deviennent eux-mêmes des égos et peuvent à leur tour identifier des *alters* supplémentaires. Toutefois l'étude des réseaux « boule de neige » implique également des contraintes logistiques très importantes. L'étude des réseaux égocentriques a donné lieu à des nombreux travaux importants en sociologie parmi lesquels la théorie sur la « force des liens faibles » que nous présentons dans la section 3.3.2.3.

### *3.3.2.2 Les analyses sociocentrées*

Les analyses sociocentrées exploitent l'intégralité d'un réseau social. La difficulté naturelle de ces analyses est l'accessibilité aux données du réseau social entier. Toutefois lorsque cette dernière est plus ou moins rendue possible, ces analyses égocentrées sont très utiles pour l'identification de structures sociales (communautés par exemple) ainsi que de leurs relations dans un réseau social. Les éléments mathématiques de la théorie des

graphes ainsi que les mesures de centralité des acteurs sont très souvent utilisés dans ces analyses (cf. section 3.3.3).

### *3.3.2.3 La force des liens faibles*

La force des liens faibles a été développée par le sociologue économiste Mark Granovetter (Granovetter, 73) qui considère que les individus n'ont pas de préférences immuables dans le temps, mais que leurs choix sont influencés par les personnes qui les entourent. De manière générale, les liens faibles d'un individu sont ses connaissances avec qui il a juste des contacts brefs et occasionnels, alors que ses liens forts sont des personnes proches qu'il rencontre fréquemment et avec qui il entretient des échanges très réguliers. Pour Granovetter, la force d'un lien est une combinaison (probablement linéaire) de la quantité de temps, de l'intensité émotionnelle, de l'intimité (confiance mutuelle), et des services réciproques qui caractérisent ce lien. Il démontre de manière empirique par des analyses égocentrées que les liens faibles d'un individu sont les plus susceptibles de lui apporter de nouvelles informations par rapport à ses liens forts, d'où l'importance des liens faibles. Paradoxalement, les liens forts ne transmettent généralement qu'une information redondante, peu utile, alors que les liens faibles fournissent des informations plus rares, plus inédites, plus difficiles d'accès, plus originales. Ce paradoxe apparent est très utile, il est par exemple exploité pour expliquer la diffusion de l'innovation, qui s'effectue de la périphérie vers le centre du réseau, ou encore, pour certains aspects du marché du travail, notamment la recherche d'emploi, beaucoup plus aisée par le biais des liens faibles que par les institutions prévues à cet effet, comme attestent plusieurs enquêtes de terrain. Par exemple dans une enquête de Granovetter effectuée auprès de 304 cadres techniciens de Boston, 56% ont trouvé leur emploi grâce à leur réseau personnel qui est donc le moyen le plus efficace pour trouver un emploi. Les emplois les plus satisfaisants dans cette enquête ont été obtenus non pas grâce à l'entourage proche (famille, amis, etc.), c'est-à-dire les liens forts, mais grâce à des collègues ou anciens collègues de travail fréquentés plus rarement. Même si la thèse de Granovetter est relativisée par d'autres sociologues comme Nan Lin (Lin, 86), elle reste celle qui est la plus approuvée dans les enquêtes de terrain.

### *3.3.2.4 Les trous structuraux*

Défini par le sociologue Ronald Burt (Burt, 92), un trou structural désigne l'espace vide entre deux individus dans un groupe : il s'agit d'une absence de relation. Cette absence de relation entre deux personnes permet à une tierce personne de se placer en intermédiaire et donc de tirer avantage de la situation. Ses avantages sont de trois sortes : un accès plus rapide à l'information (l'information ne suit plus les voies formelles et hiérarchiques de diffusion), une information de meilleure qualité (l'information est non redondante de part et d'autre du trou), un contrôle sur la diffusion de l'information (l'intermédiaire peut choisir quand, et à qui, il diffuse l'information). Cette notion est assez équivalente à celle des liens faibles de Granovetter (Gervasoni, 12) car plus les trous structuraux sont nombreux dans le réseau d'un individu, plus l'individu a des opportunités d'accès à des informations nouvelles non redondantes aux informations dont il a déjà connaissance de par ses liens forts.

### *3.3.2.5 Le capital social*

Dans le cadre de la sociologie des réseaux sociaux, on peut définir le capital social comme une forme de ressource sociale qui est produite par la taille du réseau personnel, par le volume des ressources du réseau et

qui génère des chances d'accès aux ressources par les individus du réseau (Gervasoni, 12). En plus bref, le capital social peut se définir comme le degré de facilité d'accès à des informations par les individus en s'appuyant sur leurs relations sociales. L'objet social à partir duquel se mesure le capital social est l'une des principales questions qui se pose (Borgatti et al., 98). Les mesures de capital social pourraient également prendre en compte les normes, procédures et cultures des sociétés. Ici, on s'intéresse juste aux aspects liés à la structure des réseaux sociaux. Certains sociologues estiment que le capital social doit se mesurer au niveau des individus (Putnam, 95)(Fukuyama, 95) alors que d'autres ne l'interprètent qu'à partir des structures internes (Burt, 92)(Lin, 86)(Brass, 92)(Ancona, 90) ou externes (Cohen 90)(Everett et Borgatti, 99) des groupes d'individus (Tableau 3.3).

Ces différentes conceptions/formes du capital social sont subdivisées suivant les quatre dimensions A, B, C, D du tableau 3.3 (Borgatti et al., 98).

***La dimension A*** concerne l'analyse du fonctionnement interne de l'humain. Elle est liée aux aspects psychologiques des individus plutôt qu'à leur réseau social (c'est pourquoi cette case est vide).

| | ***Type de focus*** | |
|---|---|---|
| ***Type d'acteurs*** | **Interne** | **Externe** |
| **Individus** | **A)** | **B)** Burt(92) ; Burt(83) ; Lin (86)(Brass, 92) ; |
| **Groupes** | **D)** (Putnam, 95) ; (Fukuyama, 95) ; (Bourdieu, 86) | **C)** (Ancona, 90) ; (Cohen et al., 90) ; (Everett et Borgatti, 99) |

Tableau 3. 3 : Différentes conceptions/formes de capital social

***La dimension B*** correspond à la conception individuelle du capital social. Elle se mesure très souvent par l'analyse de l'égo dans les réseaux égocentriques. Certaines mesures standards ou de centralité (cf. section 3.3.6) peuvent influencer positivement ou négativement le capital social de l'égo tel que présenté dans le tableau 3.4. Ces mesures ne sont pas exhaustives, le lecteur intéressé par toutes les mesures peut se référer à (Borgatti et al., 98).

| Nom de la mesure | Description | Relation au capital social |
|---|---|---|
| Degré (Burt, 83) | Nombre total d'alters dans le réseau égocentrique | *Positif :* plus un individu a de relations, plus il a la chance que ces derniers détiennent une information dont il pourrait avoir besoin |
| Densité (Burt, 83) | La proportion de pairs d'alters connecté dans le réseau égocentrique | *Négatif :* si trop d'alters sont connectés entre eux, il y a trop de redondance d'information (cf trous structuraux) |
| Centralité de proximité non normalisé (Freeman, 79) | Distance totale théorique entre l'égo et tous les autres individus | *Négatif :* plus l'égo est distant des autres individus, moins il a de chance de recevoir rapidement des informations |
| Centralité d'intermédiarité (Freeman, 79) | Nombre de fois que l'égo se situe sur le chemin le plus court entre deux individus. | *Positif :* plus cette mesure est grande plus l'égo est incontournable dans les échanges avec les autres individus, donc plus sujet à recevoir des informations |

Tableau 3. 4 : Quelques mesures du capital individuel

***La dimension C*** correspond à la conception de l'interprétation des liens internes des groupes pour la mesure du capital social. Quelques mesures liées à l'approche d'analyse sociocentrée correspondant à cette conception, et les relations de ces mesures avec le capital social sont présentées sur le tableau 3.5.

| Nom de la mesure | Description | Relation au capital social |
|---|---|---|
| Densité (Harary, 69) | Proportion des membres du groupe qui sont liés dans le réseau social | *Positif :* pour les échanges d'informations. |
| Distance moyenne ou maximum (Freeman, 79) | Moyenne (ou maximum) de la distance théorique entre tous couples d'individus dans le groupe | *Négatif :* les distances plus faibles impliquent des communications plus rapides. |
| Homophily* (Marsden, 88)<br>* Nécessite les attributs sur les nœuds en plus de la structure du réseau | Degré avec lequel les membres d'un groupe sont liés à d'autres membres du groupe qui leur sont similaires. | *Négatif :* moins d'homophily signifie plus de possibilité d'acquérir des informations nouvelles. |

Tableau 3. 5 : Quelques mesures du capital social interne aux groupes

***La dimension D*** correspond à la conception de l'interprétation des liens externes des groupes pour la mesure du capital social (cas par exemple des réseaux d'entreprises). Il s'agit donc des analyses des relations entre groupes d'individus, chaque groupe étant considéré comme une entité indivisible. Les analyses sont similaires à la dimension B, seulement les nœuds du réseau représentent ici les groupes plutôt que les individus. Ceci impacte les mesures de centralités qui peuvent être exploitées dans ces cas (Everett et Borgatti, 99) (Tableau 3.6).

| Nom de la mesure de centralité | Description | Relation au capital social |
|---|---|---|
| Degré du groupe (Everett et Borgatti, 99) | Nombre d'individus extérieurs liés à au moins un membre du groupe | *Positif :* les membres externes ayant des liens (positifs) avec le groupe peuvent apporter des informations nouvelles. |
| Centralité de proximité du groupe (non normalisé) (Everett et Borgatti, 99). | Distance totale du groupe vers tous les autres individus. La distance du groupe vers un individu est très souvent considérée comme la distance minimale entre les membres du groupe et cet individu. | *Négatif :* plus la distance vers les individus externe est grande, moins l'information est susceptible d'arriver rapidement dans le groupe. |
| Centralité d'intermédiarité du groupe (Everett et Borgatti, 99) | Le nombre de fois que le chemin le plus court entre deux individus passe par un membre du groupe. | *Positif :* plus cette mesure est grande, plus les informations sont susceptibles de transiter par le groupe. |

Tableau 3. 6 : Quelques mesures du capital social externe aux groupes

## 3.3.3 Principaux enjeux de l'analyse des réseaux sociaux

Les éléments mathématiques de la théorie des graphes sont généralement exploités par les sociologues (mesures du capital social par exemple) et plusieurs autres disciplines s'intéressant à l'analyse des réseaux sociaux. Le réseau social est représenté sous forme de graphe ou sous forme matricielle. Les mesures de centralité ou la recherche de communautés à partir de la structure du graphe sont des éléments très exploités.

Les outils informatiques servent à l'implémentation de ces éléments au moyen des API (exemple de l'API Facebook ou de Google Opensocial) et offrent également les fonctionnalités de visualisation : Ucinet[30], Stocnet[31], Pajek[32], Tétralogie[33], Tulip[34], VizWiz[35], etc. Avec l'avènement du Web social, de plus en plus de données sociales sont disponibles sur Internet et sont accessibles par divers moyens en posant toutefois des problèmes de sécurité et de confidentialité des utilisateurs. Dans la suite de cette partie nous ferons le tour des ces éléments importants qui sont en relation avec la problématique de cette thèse : accès à l'information, sécurité et confidentialité de l'information, métriques caractéristiques des individus et des communautés, et détection de communautés.

### 3.3.4 Accès aux données des réseaux sociaux

Le premier challenge pour l'analyse des réseaux sociaux consiste à accéder aux données sur les individus et leurs relations. Ce problème ne se pose pas à priori dans le cadre des systèmes fermés tels que les intranets d'entreprises (réseaux téléphoniques par exemple), où ces données restent propriétaires et les analyses dans ces cas restent très souvent limitées pour des questions de confidentialité de la vie privée des utilisateurs. Nous nous intéressons ici à l'accès aux données sur les réseaux sociaux dans la vie réelle ou via les médias tels que le Web. Pendant longtemps, seules des enquêtes ethnographiques (questionnaires, interviews individuelles, focus group, etc.) permettaient la collecte des données sur les réseaux sociaux des utilisateurs (Coutant et Stenger,10)(Stutzman, 06) (Boyd, 07)(Dwyer et al., 08). Ces techniques privilégient le contact humain avec les utilisateurs qui donne une dimension très réaliste des analyses qui sont réalisées par la suite. Toutefois, elles possèdent des limites telles que la difficulté de travailler sur des échantillons importants, le recueil parfois trop dirigé de l'information, les possibles biais dans les informations recueillies, ... Grâce aux technologies du Web et du Web 2.0 en particulier, la collecte automatique des données sur les réseaux est également rendue possible. Nous classifions ces techniques automatisées en quatre catégories : les techniques de Web mining, les techniques de fils de discussion, les techniques du Web sémantique, et les techniques basées sur les API des Réseaux sociaux numériques (Rsn).

**Les techniques de *Web mining*** appliquées aux réseaux sociaux consistent à analyser les contenus des pages Web afin d'identifier divers types de relations (cooccurrences de termes, citations de liens hypertextes, citations de co-auteurs, etc.) (Mika, 05) (Matsuo et al., 06) (Jin et al., 07). Dans le cas précis des Rsn par exemple, (Alim et al., 09) réalisent un recueil de données sur des profils publics[36] des utilisateurs via la fouille des fichiers HTML des profils utilisateurs. Si cette méthode d'extraction fonctionne sur certains sites de Rsn dont les profils publics des utilisateurs contiennent assez d'information (*MySpace* par exemple), la plupart des sites de Rsn donne l'accès à très peu d'information (très souvent, seuls le nom, prénom et éventuellement la liste d'amis) dans le profil public de leurs utilisateurs (Facebook, FriendSter, ...). Les techniques s'appuyant sur

[30] http://www.analytictech.com/ucinet/
[31] http://www.gmw.rug.nl/~stocnet/StOCNET.htm
[32] http://vlado.fmf.uni-lj.si/pub/networks/pajek/
[33] http://atlas.irit.fr/PIE/Outils/Tetralogie.html
[34] http://tulip.labri.fr/TulipDrupal/
[35] http://vizwiz.org/
[36] Le profil public est la partie du profil accessible même aux internautes non inscrits sur le Rsn.

les cooccurrences de noms de personnes dans les pages Web font face à un problème crucial de gestion d'homonymies qui peut générer beaucoup d'inconsistance dans les analyses.

**Les méthodes utilisées dans les fils de discussion**, permettent d'extraire et d'analyser les contenus publiés par des utilisateurs sur des chats, forums, ou groupes de discussion (Reffay et Lancieri, 06) (Sidir et al., 06) (Dimitracopoulou et Bruillard, 06). Ces techniques disposent d'assez d'outils pour extraire des contenus très pertinents et réalistes à partir des échanges des utilisateurs. Cependant elles sont limitées à des environnements fermés dont l'accès est généralement restreint exclusivement aux propriétaires des plates-formes. Dans le cas des Rsn, en général, l'accès aux données de ce type d'applications n'est pas possible pour des développeurs tiers.

**Les méthodes du Web sémantique**[37] consistent à représenter les ressources disponibles sur le Web, de telle sorte qu'elles soient interprétables automatiquement par les machines. Ces méthodes consistent d'une part à créer des vocabulaires[38] de représentation des informations d'un domaine spécifique, et d'autre part à écrire des règles[39] qui vont permettre aux machines de déduire des informations à partir des documents (données) représentés dans les vocabulaires définis. Ainsi, plusieurs vocabulaires du Web sémantique ont été définis. Ils peuvent représenter des données issues des plates-formes de Rsn : *FOAF*[40] (description des personnes, de leurs liens, et de leurs activités), *SIOC*[41] (description des échanges entre personnes, post de blogs, forums, etc.), *SKOS*[42] (description des concepts associés à d'autres vocabulaires), etc. Bien que très peu de Rsn offrent le moyen d'exporter les données de leurs utilisateurs sous ces formats, ces vocabulaires sont de plus en plus utilisés sur le Web. *FOAF* par exemple compte parmi les vocabulaires du Web sémantique les plus utilisés. La plate-forme *Livejournal* permet d'extraire les informations de ses utilisateurs sous le format FOAF. (Golbeck et Rothstein, 08) les utilisent pour la détection des identités multiples sur le Web, ou la détection de communautés basées sur les centres d'intérêts. La plateforme *Wordpress* permet l'extraction des données utilisateurs au format *SIOC*. Certains auteurs fusionnent tous ces vocabulaires afin de réaliser des analyses plus approfondies et plus riches (Breslin et Decker, 07) (Uldis, 08). Cependant les médias sociaux tels que les Rsn très utilisés de nos jours (exemple de Facebook) ne fournissent pas le moyen d'exporter les données suivant ces vocabulaires.

**Les méthodes basées sur les API de réseaux sociaux numériques (Rsn)** s'appuient sur les API fournies par les Rsn aux développeurs tiers (*API Facebook* et *OpenSocial* de *Google* notamment) pour développer des applications Web qui seront embarquées dans le Rsn. L'API Facebook est spécifique à la plateforme Facebook, alors qu'*OpenSocial* se veut interopérable et est utilisé dans plusieurs plates-formes (Orkut, LinkedIn, MySpace, Viadeo, etc.). Toutefois, Facebook étant le précurseur de ce type d'API, il dispose jusqu'à aujourd'hui (et de loin) de la plateforme hébergeant le plus d'applications tierces[43]. D'autres modèles d'applications s'appuyant sur ces API permettent de développer des applications sur les graphes sociaux des

[37] Ce que certains auteurs appellent Web 3.0

[38] Sous forme de description RDF (Resource Description Framework)

[39] Avec des langages comme OWL (Web Ontology Language)

[40] Friend Of A Friend (http://www.foaf-project.org/)

[41] Semantically-Interlinked Online Communities (http://sioc-project.org/)

[42] Simple Knowledge Organization-System (http://www.w3.org/2004/02/skos/)

[43] http://www.facebook.com/press/info.php?statistics

Rsn, mais sont plutôt embarquées dans les sites des développeurs ou des entreprises (*Facebook Connect*, *Google Friend Connect*, *MySpace Data Availability*, etc.). Les applications développées par ces API fournissent le moyen d'accéder à plusieurs types d'informations chez l'utilisateur : son profil explicite (informations sur l'identité, professionnelles, académiques, centres d'intérêts, etc.), ses activités (statuts, liens, photos, tags, commentaires, vidéos, groupes, événements, pages, flux d'activités, etc.), et son graphe social. Par rapport aux méthodes présentées précédemment, la diversité et la quantité des informations pouvant être extraites à partir de ces API peuvent enrichir considérablement les informations nécessaires à des analyses. Toutefois, l'accès aux données via les API de réseaux sociaux numériques est de plus en plus réglementé avec l'usage de protocoles comme *OAuth* qui réduit les possibilités de collecte massive et surtout non autorisée de données.

Chacune des approches de collecte ethnographique ou automatisée de données sur les réseaux sociaux possède ses avantages et ses inconvénients. Pour tirer partie des avantages de chacune des méthodes, des approches interdisciplinaires peuvent se révéler très pertinentes (Tchuente et al., 11).

### 3.3.5 Sécurité des données dans les réseaux sociaux

Le problème de sécurité et de confidentialité des données utilisateurs dans les réseaux sociaux se pose surtout depuis l'avènement des réseaux sociaux numériques (Facebook, MySpace, etc.). Un très grand nombre d'utilisateurs des réseaux sociaux numériques ne sont pas conscients de l'exposition de leur vie privée, ni de qui peut accéder à leur profil (Dwyer, 2008) (Stenger et Coutant, 10). Les Rsn définissent en général des politiques de sécurité et de confidentialité des données utilisateurs. Pourtant, certains travaux comme (Bonneau et al., 09) démontrent les possibilités d'accès aux données sur les Rsn par des tiers. L'exécution des applications tierces (et de ce fait l'extraction des données des profils) étant transparente aux utilisateurs, ces derniers peuvent facilement être des cibles de diverses attaques ou d'atteintes de données personnelles. Il peut s'agir d'attaques telles que la recomposition d'un réseau à partir des fragments de données accessibles (Tianjun et Hsinchun, 08) ou des attaques sur la machine physique de l'utilisateur par scan des ports et exécution de scripts malicieux (Patakis et al., 09). A l'inverse, certains auteurs proposent plutôt des méthodes permettant aux utilisateurs de rajouter des couches de sécurité pour la protection de leurs données personnelles sur les Rsn (Baatarjav et al., 09) (Guha et al., 08) (Felt et al., 08).

Après ces aspects liés à l'accessibilité et la sécurité des données dans les réseaux sociaux numériques, nous nous intéressons aux mesures de centralité et aux travaux sur la détection de communautés qui concernent les réseaux sociaux dans un cadre beaucoup plus général.

### 3.3.6 Mesures de centralité des individus et des groupes

L'intérêt de l'analyse des réseaux sociaux consiste à caractériser les individus ou les groupes d'individus uniquement à partir de la structure du réseau social (l'ensemble des liens du graphe) sans avoir à prendre en compte les attributs des individus.

### *3.3.6.1 Centralités des individus*

Pour les individus, cette caractérisation est très souvent réalisée par l'analyse de leur position relativement aux autres individus dans le graphe non orienté (Lemieux et Ouimet, 04) ou orienté (Malek, 09) représentant le réseau social. Les positions des individus sont généralement évaluées à travers la notion de centralité qui permet de comparer la position plus ou moins centrale d'un sommet dans un graphe (Freeman, 79) et qui est souvent utilisée pour mesurer le capital social des individus (Burt, 83). Plusieurs mesures de centralité existent dans la littérature et ont été créées au fil du temps, nous présentons ici uniquement les trois mesures les plus exploitées dans le contexte des graphes non orientés (Freeman, 79) : centralité de degré, centralité de proximité, centralité d'intermédiarité.

**La centralité de degré (*degree centrality*)** est une mesure qui reflète l'activité relationnelle directe d'un acteur. Elle mesure le nombre de connexions directes de chaque acteur dans un graphe. Selon cette mesure, l'acteur qui occupe la position la plus centrale dans un graphe est celui qui détient le plus grand nombre de connexions directes dans le graphe. Dans un graphe non dirigé avec *n* individus, le degré de centralité d'un individu *i* noté ici $C_D(i)$ est le nombre de connexions directes de l'acteur noté *d(i),* normalisé par le nombre total de connexions directes *n-1*, formule (11).

$$C_D(i) = \frac{d(i)}{n-1} \qquad \textbf{(11)}$$

**La centralité de proximité (*closeness centrality*)** est une mesure qui repose sur la distance géodésique, c'est-à-dire la longueur du plus court chemin reliant deux individus. Selon cette mesure de centralité, un individu est d'autant plus central s'il peut interagir facilement avec les autres individus. Par conséquence, sa distance avec les autres individus doit être courte. Si *d(i,j)* est la distance la plus courte entre deux individus *i* et *j* d'un graphe non orienté (mesurée par le nombre de liens via le chemin le plus court), la centralité de proximité de l'individu i, noté ici $C_c(i)$ est la somme des distances les plus courtes *d(i,j)* entre i et tous les autres individus j du graphe, normalisée à partir du minimum (n-1) de cette somme (cas où i est directement lié à tous les autres individus j du graphe), formule (12).

$$C_C(i) = \frac{n-1}{\sum_{j=1}^{n} d(i,j)} \qquad \textbf{(12)}$$

**La centralité d'intermédiarité (*betwenness centrality*)** est une mesure de l'importance de la position intermédiaire occupée par les individus d'un graphe. Sur le plan conceptuel, (Freeman, 79) a défini cette mesure pour rendre compte de la capacité qu'ont les individus à assurer le rôle de coordination et de contrôle. L'hypothèse est que plus un acteur se trouve dans une position intermédiaire, c'est-à-dire plus il est dans une situation où les individus doivent passer par lui pour atteindre d'autres individus, plus il aura la capacité à contrôler la circulation de l'information entre ces individus. Par conséquent, si *i* est un individu et si *i* se localise sur le chemin de plusieurs interactions, alors i est important. Soit $P_{jk}$ le nombre de chemins les plus courts (appelés géodésiques) entre deux individus j et k. L'intermédiarité de l'individu *i* est définie par le nombre de chemins les plus courts entre j et k passant par i noté $P_{jk}(i)$ (avec $i \neq j$ et $i \neq k$), normalisée par le nombre total des chemins les plus courts entre toutes les paires d'acteurs qui n'incluent pas i, formule (13).

$$C_B(i) = \sum_{j<k} \frac{P_{jk}(i)}{P_{jk}} \qquad \textbf{(13)}$$

### *3.3.6.2 Centralités des groupes*

Ces mesures de centralité présentées précédemment ont été définies pour les individus. Toutefois, certaines analyses nécessitent naturellement la caractérisation de la position des groupes d'individus plutôt que des personnes prises individuellement. Dans un laboratoire de recherche, on peut parfois vouloir analyser des équipes de recherche plutôt que des chercheurs pris individuellement. Dans cet optique (Everett et Borgatti, 99) étendent les mesures de centralité pour caractériser plutôt les groupes d'individus dans un réseau social. Dans toutes les définitions qui suivent, on suppose le graphe G=<V, E> représente un réseau social, V étant l'ensemble des individus et E l'ensemble des liens entre ces individus.

**La centralité de degré d'un groupe d'individus (*group degree centrality*)** est définie comme étant le nombre d'individus à l'extérieur du groupe, connecté à au moins un membre du groupe dans le réseau social. Chaque individu extérieur au groupe connecté à au moins un membre du groupe est compté une seule fois (même s'il dispose de plusieurs liens avec les membres du groupe). Si C représente un groupe d'utilisateurs dans le réseau social, et N(C) représente le nombre d'individus de V∉ C (individus extérieurs à C), la centralité de degré de C vaut N(C), normalisé par la valeur maximale de N(C) qui est |V|-|C|, formule (14). Il est a noté que plus le groupe est de taille importante, plus il est probable que sa centralité de degré soit élevé. En effet, plus le groupe est de taille importante, moins il y'a d'individus à l'extérieur du groupe (donc |V|-|C| devient petit) et il est plus probable que ces individus extérieurs soit connectés au groupe (N(C) est donc relativement plus grand).

$$GC_D(C) = \frac{N(C)}{|V| - |C|} \qquad \textbf{(14)}$$

**La centralité de proximité d'un groupe (*group closeness centrality*)** est définie comme l'inverse normalisée de la somme des distances du groupe vers tous les individus extérieurs au groupe. La principale différence avec la centralité de proximité pour les individus réside dans le calcul des distances. Il n'existe qu'un seul moyen de calcul de distance entre deux individus, en général l'usage du chemin le plus court entre les deux individus dans le graphe. Par contre pour calculer la distance entre un groupe (ensemble d'individus) et un individu externe dans un graphe, plusieurs cas peuvent être considérés : la moyenne des distance entre les l'individu et chaque membre du groupe, le minimum des distances entre l'individu et chaque membre du groupe, le maximum des distances entre l'individu et chaque membre du groupe, la médiane des distances entre l'individu et chaque membre du groupe, etc. Ainsi si C désigne un groupe d'individus dans le réseau social, $D_x = \{d(x,c), c \in C\}$ $x \in V\text{- }C$, l'ensemble des distances entre un individu $x$ externe à C, et $d_f(x,C) = f(D_x)$ la distance entre l'individu x et groupe C (agrégation des distances dans $D_x$ suivant une fonction d'agrégation *f=moyenne, minimum, maximum, médiane, etc.,*), la centralité de proximité du groupe C est exprimée comme la somme des distances $d_f(x,C)$ entre le groupe C et chacun les individus x externe à C, normalisée par le maximum de cette somme (|V|-|C|) obtenue lorsque tous les individus x se situent à distance 1 de C, formule (15). La question du choix de la fonction *f* du calcul des distances dépend de la nature des données. Si les données sont par exemple telles qu'un groupe peut être considéré comme une unité indivisible, la distance minimum serait alors

la plus appropriée. Par exemple à considérer les groupes des informateurs de police dans un réseau criminel. Si on considère qu'aussi tôt qu'un informateur a une information, l'information est passée de manière instantanée à la police, dans ce cas il est raisonnable d'utiliser la distance minimum. En fait l'accès à une information externe au groupe est fonction de la distance la plus courte entre un informateur et cette information.

$$GC_C(C) = \frac{|V| - |C|}{\sum_{x \in V-C} d_f(x, C)} \qquad \textbf{(15)}$$

**La centralité d'intermédiarité d'un groupe (*group centrality betweeness*)** mesure la proportion de chemins les plus courts (géodésiques) connectant les pairs d'individus non membres du groupe. Si C est un groupe d'individus, et $g_{u,v}$ le nombre de géodésiques connectant *u* et *v*, et $g_{u,v}(C)$ le nombre de géodésiques entre u et v passant par au moins un membre de C, la centralité d'intermédiarité du groupe C est la proportion de géodésiques passant par C, normalisée par nombre de chemins les plus courts entre toutes les paires d'individus non membre de C (donné par 1/2(|V|-|C|)*(|V|-|C|-1)), formule (16). La signification de la centralité d'intermédiarité de groupe est la même que la centralité d'intermédiarité des individus.

$$GC_B(C) = \frac{2 * \sum_{u<v} \frac{g_{u,v}(C)}{g_{u,v}}}{(|V| - |C|) * (|V| - |C| - 1)} \qquad \textbf{(16)}$$

Ces mesures de centralité sont appliquées aux groupes d'individus dans le réseau social. Les groupes d'individus considérés ici peuvent exister à priori dans une organisation dans la vie réelle (équipe de recherche) où alors être eux-mêmes détectés à partir de la structure des liens dans le graphe. La détection des communautés est une autre problématique majeure dans l'analyse des réseaux sociaux. Nous la présentons sommairement dans la section suivante.

## 3.3.7 Détection de communautés dans les réseaux sociaux

La problématique de détection de communautés dans les graphes est de trouver le moyen optimal de rechercher des groupes de nœuds (communautés) tels que les nœuds dans chaque groupe soient fortement connectés entre eux, et faiblement connectés aux nœuds du graphe à l'extérieur du groupe. Cette recherche de communautés s'appuie uniquement sur la topologie ou la structure du graphe (liens entre les nœuds). La détection de communautés qui s'appuie sur la topologie du graphe est une technique complémentaire aux techniques de *clustering* capables de retrouver des communautés (*clusters*) à partir des analyses sémantiques sur les attributs des nœuds. La nature relationnelle entre nœuds dans un graphe apporte toutefois une proximité réelle entre les membres d'une communauté retrouvée. De plus, dans de nombreux cas, seule la topologie du graphe est entièrement connue, alors que les attributs des nœuds possèdent très souvent des données manquantes. Pour être plus optimal, certains travaux combinent l'usage de la topologie du graphe et les attributs des nœuds (Gomez et al., 11) lorsque les deux types d'information sont disponibles. Dans cette partie, nous nous intéressons à la détection de communautés au sens premier du terme (c'est-à-dire basée sur la topologie).

La problématique de détection de communautés dans les graphes s'applique aux réseaux sociaux mais concerne également de nombreux autres types de réseaux : réseaux biologiques (réseaux métaboliques entre gènes et protéines, réseaux de neurones, etc.), réseaux d'infrastructures (réseaux de transports, réseaux de distribution électriques, etc.), les réseaux d'information (réseau Internet de pages Web reliées par les liens hypertextes), réseaux linguistiques (réseaux de synonymies, réseaux sémantiques, etc.), etc. Cette multitude de domaines d'application suscite le développement de très nombreux algorithmes génériques de détection de communautés pouvant être exploités dans chaque domaine (Fortunato, 10)(Pons, 05). Les paragraphes suivant restent très sommaires et peu détaillés sur les problématiques liées à la détection de communautés. Ils sont cependant suffisants dans le cadre de nos contributions. La détection de communautés est un champ très large et le lecteur intéressé par plus de détails pourra se référer à (Pons, 05)(Schaefer, 07)(Fortunato, 10).

Les questions qui se posent généralement concernant ces algorithmes de détection de communautés concernent les sujets suivants :

- ***La gestion du recouvrement :*** un algorithme prenant en compte le recouvrement des communautés est un algorithme pour lequel un même nœud peut appartenir à deux communautés distinctes retrouvées (exemple : CFinder (Palla et al., 05)). Ce qui est très souvent le cas dans les réseaux sociaux. D'autres algorithmes par contre n'autorisent pas le recouvrement, réalisent des partitions du graphe (exemple : Infomap (Rosvall et al., 07)). Le choix du type d'algorithme à exploiter doit dépendre fortement de la nature des données.

- ***La fixation à priori de la taille minimale de communautés à rechercher :*** certains algorithmes prennent en entrée la taille minimale de chacune des communautés à rechercher (exemple : CFinder (Palla et al., 05)).

- ***Les types de données de validation des algorithmes :*** si un très grand nombre de chercheurs s'intéressent à la détection de communautés dans les graphes, beaucoup d'entre eux ne disposent pas toujours de données de graphes ou de réseaux sociaux réels pour la validation de leurs travaux. Ainsi plusieurs travaux s'appuient sur des procédés de génération automatique de graphes respectant des caractéristiques principales des types de réseaux qu'ils souhaitent exploiter. Pour les réseaux sociaux par exemple, les graphes sont en général caractérisés par des propriétés telles que : une distribution de degré suivant une loi de puissance (graphes sans échelle ou *scale free*)[44], les six degrés de séparation[45], etc. Ces graphes auto-générés possèdent des communautés bien connues (car créées automatiquement) qui doivent alors être retrouvées par les algorithmes à évaluer. Cependant des expérimentations telle que celle de (Navarro et al., 10) montrent que ce type de graphes auto-générés n'est pas toujours pertinent pour des évaluations comparées aux graphes réels.

- ***Les techniques de construction et d'évaluation des communautés :*** les questions de la construction et de l'évaluation du découpage du graphe en communautés sont abordées suivant plusieurs approches. Nous citons ici uniquement les principaux qui sont *la modularité, la marche aléatoire* et les *k-cliques*. La modularité est une fonction de mesure de la qualité d'un partitionnement, introduite initialement par Girvan et Newman, pour choisir une coupe privilégiée dans un dendrogramme issu d'un *clustering*

[44] Très peu d'individus ont un nombre très élevé de voisins, et un très grand d'individus ont peu de voisins

[45] Deux nœuds quelconque du graphe sont séparés par une distance d'au moins

hiérarchique (Gervin et al., 02). *La modularité* mesure le nombre d'arêtes à l'intérieur des communautés auquel on enlève ce même nombre obtenu sur un graphe aléatoire (i.e sans structure) de même taille mais gardant exactement la même distribution de degrés. L'existence de ce lien dans le graphe réel ne sera donc pas un argument pertinent pour considérer que ces 2 nœuds sont effectivement dans la même communauté. Au contraire, l'existence d'un lien entre 2 nœuds ayant peu de chances d'être liés dans le graphe aléatoire sera un argument fort pour les regrouper. Des algorithmes comme Louvain (Blondel, 08) s'appuient sur la modularité tout en l'optimisant. A la différence de la modularité, *la marche aléatoire* s'appuie sur les distances entre sommets (Gaume, 04). Une marche aléatoire courte, partant d'un sommet donné, tend à rester dans la (les) communauté(s) de ce sommet. Ainsi, la distance entre les résultats de deux marches aléatoires partant de deux sommets distincts, révèle efficacement l'appartenance commune ou non de ces sommets à une même communauté. Cette distance permet à cette méthode de partitionner le graphe par l'intermédiaire d'un algorithme de *clustering* hiérarchique. L'algorithme *Walktrap* (Pons, 07) s'appuie par exemple sur les marches aléatoires. A la différence de la modularité et des marches aléatoire, les k-cliques s'appuient sur la recherche de motifs locaux dans le graphe. Une communauté est définie comme une chaîne de k-cliques adjacentes. Une k-clique est un sous-ensemble de sommets tous adjacents les uns aux autres, et deux k-cliques sont adjacentes si elles partagent des sommets. L'avantage immédiat d'une telle approche est la détection de communautés avec recouvrement, un sommet pouvant appartenir à plusieurs k-cliques non forcément adjacentes. Au-delà de ces trois techniques qui restent assez génériques sur les types de graphes cibles, de nouvelles approches très spécifiques aux réseaux sociaux sont proposées. La mesure de *cohésion sociale* proposée par (Frigerri et al., 11) s'appuie par exemple sur des travaux en sociologie pour construire et évaluer la pertinence des communautés. Ainsi, *la cohésion sociale* construit et évalue les communautés par rapport au nombre de triades (trois sommets tous connectés) présentes dans la communauté. Même si l'idée semble intéressante, les évaluations à grande échelle de cet algorithme sont encore attendues.

- ***La prise en compte de la dynamique des réseaux :*** les réseaux sociaux en particulier sont des structures qui évoluent très souvent et parfois de manière très rapide. De nouveaux nœuds et liens apparaissent au fil du temps, en même temps que d'autres peuvent disparaître. Pour la plupart des algorithmes, l'évolution de la topologie du graphe implique la régénération des communautés. Ceci peut s'avérer très lourd en temps d'exécution et ressources exploités, surtout sur de très grands graphes. Pour pallier cet inconvénient de nouveaux algorithmes tels que iLCD (Cazabet et al., 10) proposent de détecter des communautés qui pourront dynamiquement se mettre à jour au fil de l'évolution de la structure du réseau (apparition de nouveaux liens, suppression de liens existants). Cette approche est justifiée (car elle réduit la complexité de régénération des communautés à chaque modification de la structure du réseau social), mais nécessite également des validations à plus grande échelle.

### 3.3.8 Synthèse

Dans cette partie sur les éléments de l'analyse des réseaux sociaux, nous avons ciblé et décrit les éléments sociologiques, mathématiques et informatiques susceptibles d'être exploités pour traiter la problématique de

dérivation d'un profil social de l'utilisateur le plus caractéristique possible de ce dernier. Les éléments sociologiques fournissent des idées et hypothèses (analyses égocentrées vs analyses sociocentrées, force de liens faibles, trous structuraux, capital social) qui ont déjà fait leurs preuves dans les sciences humaines et sociales, et qui pourraient se révéler très utiles dans un passage à l'échelle dans les analyses informatisés. Les éléments mathématiques (sociométriques) fournissent des formules (mesures de centralités notamment) pouvant servir de support aux éléments sociologiques. Les éléments informatiques (APIs, protocoles, logiciels, etc.) fournissent des moyens permettant d'accéder de manière sécurisée aux données et de réaliser des implémentations des algorithmes proposés.

## 3.4 Conclusion

Ce chapitre a été divisé en deux grandes parties quelque peu disjointes, mais très complémentaires pour notre contribution dans cette thèse.

La première partie a présenté les travaux et techniques relatives au filtrage social. Ces travaux et techniques s'appuient sur les réseaux sociaux de très différentes natures (réseaux de confiance, réseaux de co-auteurs, réseaux de familiarité, réseaux de similarité sociale, etc.). Malgré ces différences, ces travaux ont un objectif commun : démontrer que l'ajout de données sociales dans les mécanismes traditionnels de filtrage d'information améliore la qualité de ces mécanismes. Dès lors, ils s'intéressent plus aux évaluations des mécanismes de filtrage social qu'à l'optimisation de l'usage du réseau social (graphe social) de l'utilisateur. Très souvent seuls les individus ayant les liens les plus forts avec l'utilisateur sont sélectionnés et exploités pour l'extraction de données sociales. Certes, les résultats utilisant ces sélections démontrent déjà l'intérêt d'une telle approche. Toutefois, nous pensons que ces résultats pourraient être améliorés et optimisés si en amont des mécanismes de filtrage social, nous pouvons répondre à la question suivante : comment exploiter efficacement le graphe social de l'utilisateur pour dériver les informations de son profil ? Se focaliser uniquement sur les liens forts de l'utilisateur (comme le font la quasi-totalité des mécanismes de filtrage social) n'est qu'une approche pour répondre à cette question. D'autres approches peuvent être considérées en exploitant la richesse des travaux existants dans l'analyse des réseaux sociaux. C'est ce qui motive la deuxième partie présentée dans ce chapitre.

La deuxième partie a présenté les éléments sociologiques, mathématiques et informatiques de l'analyse des réseaux sociaux qui sont potentiellement utiles pour un usage optimal du réseau social de l'utilisateur dans les mécanismes de filtrage social de l'information. Les réseaux sociaux numériques étant des plateformes qui fournissent le plus de données permettant de nos jours des évaluations à grande échelle sur les réseaux, nous nous sommes dans un premier temps intéressés aux problématiques d'accès et de sécurisation des données dans ces nouvelles plateformes. Dans un cadre beaucoup plus général, nous nous sommes ensuite intéressés aux mesures de caractérisation des individus et des groupes d'individus, ainsi qu'aux algorithmes de détection de communautés dans les réseaux sociaux qui sont susceptibles d'être exploités pour améliorer les travaux existants concernant le filtrage social de l'information.

La présentation des limites des mécanismes de filtrage social de l'information et les éléments pertinents de l'analyse des réseaux présentés en seconde partie de ce chapitre motivent notre contribution qui est présentée dans le chapitre qui suit.

# 4 Chapitre 3 : Contribution : modèle générique et techniques de dérivation de profils utilisateurs sociaux

# 4.1 Introduction

Notre contribution vise à s'appuyer sur les travaux existants dans l'analyse des réseaux sociaux pour répondre aux deux problématiques énoncées dans le chapitre précédent : non généricité de modèles de profils utilisateurs sociaux et non usage optimal du graphe social de l'utilisateur pour dériver des informations pertinentes de son profil.

Pour ce faire, ce chapitre est structuré comme suit : nous présentons dans un premier temps les notions et concepts inhérents du modèle de profil « social » (intégrant le réseau social de l'utilisateur) de l'utilisateur que nous proposons pour adresser la problématique de non généricité des mécanismes de filtrage social. Ensuite, nous présentons l'algorithme de dérivation des informations du profil de l'utilisateur à partir du modèle proposé en exploitant les communautés du graphe social de l'utilisateur. Enfin, nous présentons les principales stratégies possibles d'évaluation de notre proposition. Nous terminons le chapitre par une conclusion qui récapitule les principaux éléments de notre contribution.

# 4.2 Modèle générique social de profil utilisateur

Il est important de définir un modèle générique et social de profil de l'utilisateur pour deux raisons :

***(i)*** premièrement afin d'avoir une vision unique de la nature des réseaux sociaux exploités dans les expérimentations actuelles, ou du moins, définir une famille de réseaux sociaux ayant des caractéristiques communes à partir desquelles des interprétations tirées par rapport à leur impact sur le profil de l'utilisateur puissent être comparables.

***(ii)*** Ensuite, afin de séparer les mécanismes d'adaptation (filtrage social) et les profils utilisateurs (sociaux) de telle sorte que le modèle défini puisse être réutilisable avec le plus de flexibilité possible quelque soit le mécanisme qui y sera associé. La définition de ce modèle générique social de profil utilisateur a pour but de répondre aux deux premières limites évoquées dans l'introduction de ce chapitre. Dans un premier temps nous

définissons les types de réseaux sociaux sur lesquels le modèle que nous proposerons par la suite sera applicable, avant de présenter les contours et concepts du modèle lui-même.

### 4.2.1 Définition du réseau social

Tel que présenté dans l'état de l'art, les réseaux sociaux exploités dans les systèmes de filtrage social de l'information peuvent être de différentes natures (réseaux de confiance, réseaux de similarité, réseaux de familiarité, réseaux de co-auteurs, etc.). Si l'on considère ces réseaux représentés uniformément sous forme de graphes, leur différence se situe au niveau de la sémantique (signification) du lien entre deux nœuds (confiance, similarité, familiarité, co-auteur, etc.). Ce lien est en général déduit des interactions entre les utilisateurs. Ces interactions peuvent être directes (exemple de deux individus déclarés explicitement amis sur le réseau social numérique) ou indirectes (deux individus qui ont annoté ou créé une même ressource tel qu'un document).

Dans ce travail, nous considérons un réseau social dans le sens sociologique du terme. C'est-à-dire un réseau dans lequel la sémantique du lien entre deux individus implique que ces derniers se connaissent mutuellement dans la vie réelle. Dans le cas des systèmes d'information, les réseaux sociaux sont construits à partir des actions et interactions des utilisateurs. Ces actions n'impliquent pas forcément que les utilisateurs se connaissent dans la vie réelle (cas de réseaux de similarité). Afin de s'appuyer sur les travaux issus des sciences sociales, ***nous allons donc considérer comme réseau social dans un système d'information, les réseaux construits à partir des actions ou interactions entre utilisateurs qui impliquent de manière très probable que ces derniers se connaissent dans la vie réelle***. Dans un système d'information, les interactions considérées peuvent être des interactions directes (demande/acceptation d'amitié sur Facebook), mais aussi des interactions à priori indirectes mais qui laissent supposer une connaissance mutuelle entre les deux individus (deux auteurs qui coécrivent un article scientifique sont supposés se connaître…). Comme exemple de réseaux sociaux considérés ici, nous pouvons citer les réseaux sociaux numériques (Facebook, LinkedIn, Viadeo, etc.), les réseaux de co-auteurs d'articles scientifiques (DBLP, Mendeley, etc.), les réseaux d'abonnés (répertoires de contacts) chez un opérateur de téléphonie mobile, etc. Un contre-exemple est un réseau d'individus au comportement similaire (réseaux de similarité) tels que ceux exploités dans le filtrage collaboratif.

Nous considérons ce type de réseaux sociaux car ils peuvent tous être analysés et interprétés selon les résultats déjà obtenus dans les études ethnographiques ou empiriques dans le domaine des sciences sociales. Il est important de noter que les hypothèses d'influence du comportement d'un individu par son environnement social ont été étudiées en premier dans les sciences sociales (Gofmann, 59)(Granovetter, 73). Ce sont ces mêmes hypothèses qui motivent l'intérêt des chercheurs pour les systèmes de filtrage social de l'information, toutefois les mécanismes actuels de filtrage social exploitent très peu ou pas du tout ces hypothèses (à notre connaissance). Comme nous le verrons un peu plus tard dans ce document, ces travaux en sciences sociales représentent un des socles sur lequel est bâtie notre contribution. Pour se positionner sur les mêmes hypothèses et résultats que ces travaux, nous abordons donc les réseaux sociaux d'un point de vue sociologique.

## 4.2.2 Hypothèse du travail : vers un modèle de profil orienté communautés du réseau social

### *4.2.2.1 Rappel des objectifs*

Dans la section précédente, nous avons défini les types de réseaux sociaux sur lesquels s'appuient nos travaux. Partant de cette définition, nous allons montrer l'intérêt de notre contribution à partir des remarques/limites sur les techniques de filtrage social et en s'appuyant sur les éléments de l'analyse des réseaux sociaux présentés dans le chapitre précédent. Le but de notre contribution dans cette thèse étant de rechercher comment exploiter efficacement le graphe social de l'utilisateur pour dériver les éléments qui serviront à enrichir son profil, deux questions fondamentales se posent :

1. **Comment sélectionner de manière pertinente dans le graphe, les individus à partir desquels le profil de l'utilisateur sera dérivé ?**
2. **Comment analyser efficacement ces individus afin de dériver des éléments pertinents pour le profil de l'utilisateur ?**

Nous nous intéressons uniquement à la première question dans cette section et la section suivante. La seconde question sera abordée dans la section 4.3. Pour mieux y répondre en se positionnant par rapport à l'état de l'art, nous présentons d'abord les remarques/limites des travaux de la littérature du filtrage social.

### *4.2.2.2 Remarques sur les travaux relatifs au filtrage social*

Les travaux relatifs au filtrage social présentés en première partie du chapitre précédent s'appuient en général sur les individus directement connectés à l'utilisateur dans le graphe social (cf. Tableau récapitulatif 2.3 du chapitre précédent). Il peut s'agir par exemple de tous les individus qui sont directement connectés à l'utilisateur et qui lui sont également similaires (Massa et Avesani, 07), d'un sous-ensemble des individus possédant les liens les plus forts avec l'utilisateur lorsque le graphe est pondéré (30 premiers par exemple) (Guy et al., 08)(Carmel et al., 09), de tous ces individus sans distinction (Zeng et al., 09). Certains travaux beaucoup plus rares tels que (Cabanac et al., 11) proposent une approche qui sélectionne des individus dans le graphe entier en s'appuyant sur des mesures de centralité comme la centralité de proximité. Nous allons nous appuyer sur les éléments de l'analyse des réseaux présentés en deuxième partie du chapitre précédent pour commenter ces travaux.

### *4.2.2.3 Commentaires à la lumière des éléments de l'analyse des réseaux sociaux*

D'après les éléments sociologiques de l'analyse des réseaux sociaux, les deux approches présentées précédemment rentrent dans les deux catégories d'analyse de réseaux sociaux suivantes :

- ***Les analyses égocentrées :*** exemple de (Massa et Avesani, 07)(Guy et al., 08)(Carmel et al., 09)(Zeng et al., 09) qui se concentrent uniquement sur les individus situés à distance 1 (appelés *alters*) de l'utilisateur (appelé *égo*).
- ***Les analyses sociocentrées :*** exemple de (Cabanac., 11) qui analyse le graphe social tout entier.

Dans le contexte de nos travaux (systèmes d'adaptation de l'information aux utilisateurs), les analyses sociocentrées seraient très coûteuses en temps pour des systèmes nécessitant des profils construits en temps réels et régulièrement mis à jour (profils à court terme par exemple pour la personnalisation de requêtes dans un moteur de recherche). Les temps de calcul élevés de mesures de centralité des individus dans le graphe entier seraient trop importants et inacceptables pour ce type de systèmes. De plus, d'une part, la collecte de données sur les réseaux sociaux dans leur globalité n'est pas aisée, d'autre part des travaux fondateurs tels que ceux sur la force des liens faibles (Granovetter, 73) ou ceux sur les trous structuraux (Burt, 92) démontrent qu'un individu peut être caractérisé uniquement via l'analyse des réseaux égocentriques. Ceci permet de motiver l'intérêt dans un contexte comme le nôtre pour les réseaux égocentriques, d'autant plus que les temps de calcul dans ce type de réseau resteront beaucoup plus raisonnables pour l'analyse d'un égo ; Une dizaine voire centaines de nœuds uniquement. Ceci peut aussi justifier le fait que la majorité des travaux s'intéressent uniquement au voisinage proche (distance 1 le plus souvent) de l'utilisateur.

Cependant, même s'il s'avère plus judicieux de se concentrer sur les réseaux égocentriques, la question sur la manière de les exploiter se pose aussi. Certains auteurs considèrent tous les alters, d'autres uniquement des sous-ensembles jugés plus pertinents (ceux ayant les liens les plus forts par exemple). Il paraît évident que pour un égo, tous les alters de son réseau égocentrique n'ont pas la même importance, donc ne devraient pas en toute logique être considérés de la même manière dans l'analyse de l'égo. Les travaux qui s'appuient uniquement sur les liens forts de l'utilisateur sont certes intéressants, mais il est également clair que beaucoup d'informations potentiellement utiles pour l'utilisateur en provenance de ses liens faibles sont ignorés dans ces travaux (cf. section 3.3.2.3 sur la force des liens faibles de (Granovetter, 73)). En effet, les liens faibles de l'utilisateur sont considérés comme vecteur de nouvelles informations pour l'égo. De la même manière, pour le profil de l'utilisateur (égo dans ce cas), ses liens faibles seraient également pertinents pour dériver des informations nouvelles potentiellement utiles pour lui. Au final, s'appuyer uniquement sur les liens forts de l'utilisateur ne suffit pas, il faudrait également prendre en compte les liens faibles pour être plus optimal. Ceci motive notre approche basée sur les communautés plutôt que sur les individus. Nous la présentons dans la section suivante.

#### *4.2.2.4 Relations entre alters comparées aux liens forts et liens faibles : vers une approche orientée communautés*

Dans la section précédente, nous avons montré que considérer tous les *alters* sans distinction ne considérer que les *alters* ayant les liens forts avec l'égo n'est pas optimal pour dériver des informations sur l'égo. En fait, il serait extrêmement difficile de définir un système de pondération adéquat permettant de prendre en compte à la fois les liens forts et les liens faibles de l'utilisateur pour dériver de l'information qui serait utile pour lui. De plus, certains graphes ne sont pas toujours pondérés.

***Afin de considérer à la fois les liens qui seraient forts ou faibles, nous considérons dans cette thèse une approche qui ne s'appuie pas sur les individus dans le réseau égocentrique, mais sur les communautés d'utilisateurs dans ce réseau.*** Au lieu de mettre en avant le poids des relations entre l'égo et ses alters ou de ne faire aucune distinction entre les alters, nous nous intéressons plutôt aux relations existantes entre les alters et les communautés qui peuvent en être extraites. Cette idée vient du fait que nous remarquons (dans les sciences sociales) que le capital social d'un individu (facilité qu'un individu aurait à accéder à une ressource, cf. section 3.3.2.5) peut se mesurer par son implication dans un réseau social (Burt, 92)(Lin, 86)(Brass, 92),

mais aussi par le capital social des communautés d'individus (Putnam, 95)(Fukuyama, 95)(Bourdieu, 86)(Ancona, 90)(Cohen et al.,90) avec lesquelles il est en relation. Des travaux tels que ceux réalisés par (Adams, 10) encouragent également l'usage des communautés (plutôt que des individus) autour des utilisateurs pour comprendre le fonctionnement du réseau d'un utilisateur sur les réseaux sociaux numériques. Pour mieux expliquer nos propos, considérons l'exemple qui suit :

Un utilisateur X est un fan de tennis, et cet intérêt pour le tennis se traduit par le fait qu'il est abonné à un club de tennis. Supposons qu'un système d'information ne connaît pas l'intérêt de X pour le tennis, mais connaît un certains nombre d'intérêts des individus du réseau égocentrique de X, ainsi que les relations entre ces individus. Il s'avère par exemple que dans la réalité, X rencontre de temps en temps lors de son passage au club de tennis les individus C1, C2, C3, C4 qui sont également membres du club de tennis. C1, C2, C3, C4 se connaissent entre eux (probable car ils sont dans le même club de tennis), mais ont chacun des liens faibles avec l'égo X. Ceci peut se traduire par le graphe présenté sur la figure 4.1 dans lequel d'autres alters de X (I1 … I10) sont également présentés. La question est de savoir comment retrouver que X possède comme centre d'intérêt important le tennis ?

Selon les travaux de la littérature analysés au chapitre 2 :

- ***Dans le cas où seuls les liens forts ou un sous-ensemble de liens forts sont utilisés :*** le centre d'intérêt *tennis* ne figurera pas dans le profil de X dérivé du réseau social, car tous les liens entre X et les membres du club de tennis sont faibles,
- ***Dans le cas où tous les alters sont tous considérés comme équivalents :*** le centre d'intérêt tennis figurera mais risque d'être noyé dans tout un grand nombre de centres d'intérêts dérivés des autres alters, surtout si l'utilisateur a un très grand nombre d'alters.

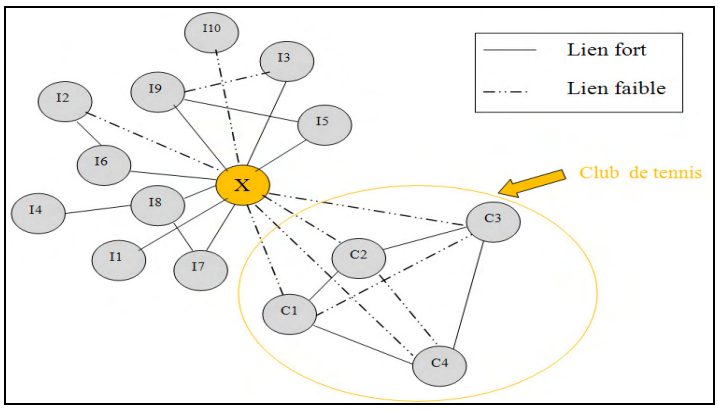



Figure 4. 1: Illustration de l'intérêt des communautés dans un réseau égocentrique

Dans la figure 4.1 il est par contre possible de retrouver que le tennis est un élément important du profil de l'utilisateur en s'appuyant sur les relations entre alters, qui pourra permettre d'identifier la communauté d'utilisateurs C1, C2, C3, C4.

De manière plus simple, il suffira par la suite de rechercher les points communs des profils de tous les membres de cette communauté et de constater qu'ils s'intéressent tous au tennis, et cet intérêt serait alors très certainement lié et important pour l'égo X. Ainsi, indépendamment de la force des liens, il serait possible des

retrouver des centres d'intérêts pertinents pour l'égo à partir des relations entre *alters* dans son réseau égocentrique.

Cette simple illustration traduit l'idée générale de notre contribution qui s'appuie sur l'hypothèse suivante : ***les communautés dans le réseau égocentrique de l'utilisateur sont plus caractéristiques de ce dernier par rapport aux alters considérés individuellement avec ou sans pondération des relations*** (Gofman, 59)(Granovetter, 73). Cette hypothèse constitue le fondement de notre approche et est à la base du modèle de profil social utilisateur proposé dans la section suivante.

## 4.2.3 Vers un modèle générique social de profil utilisateur à partir de réseaux k-égocentriques

Deux principales raisons motivent l'intérêt pour la modélisation d'un profil générique et social des utilisateurs pour les systèmes de filtrage social de l'information : le besoin de séparer la phase de construction du profil des mécanismes de filtrage pouvant utiliser ce profil dans un but de généricité, et l'intégration efficace du réseau égocentrique de l'utilisateur dans ce profil à modéliser (comme indiqué en fin de section précédente). Nous expliquons chacune de ces motivations avant de présenter le modèle générique et social de profil utilisateur qui en est déduit.

### *4.2.3.1 Besoin de généricité de profils dans le filtrage social*

Comme indiqué en synthèse des travaux de l'état de l'art et dans l'introduction de ce chapitre, les techniques actuelles de filtrage social ne font pas une séparation claire entre la modélisation du profil et son exploitation par les mécanismes associés. Comme indiqué dans le chapitre 1, le profil utilisateur doit être perçu comme une donnée ou une information qui peut être représentée indépendamment des mécanismes d'adaptation qui l'utilisent. Ceci rend génériques et par conséquent réutilisables les profils construits. Les travaux actuels de filtrage social ne s'intéressent en général qu'à démontrer l'intérêt de l'usage des données sociales, et ne respectent donc pas ce critère de généricité. Le réseau social de l'utilisateur et éventuellement son profil sont directement exploités par les mécanismes (situation présentée de manière simplifiée sur la figure 4.2A). Pour pallier cet inconvénient, notre objectif est de pouvoir définir un profil utilisateur générique intégrant son réseau social (modélisation « sociale »), c'est-à-dire indépendant d'un mécanisme précis de sorte que plusieurs types de mécanismes puissent l'exploiter (figure 4.2B).

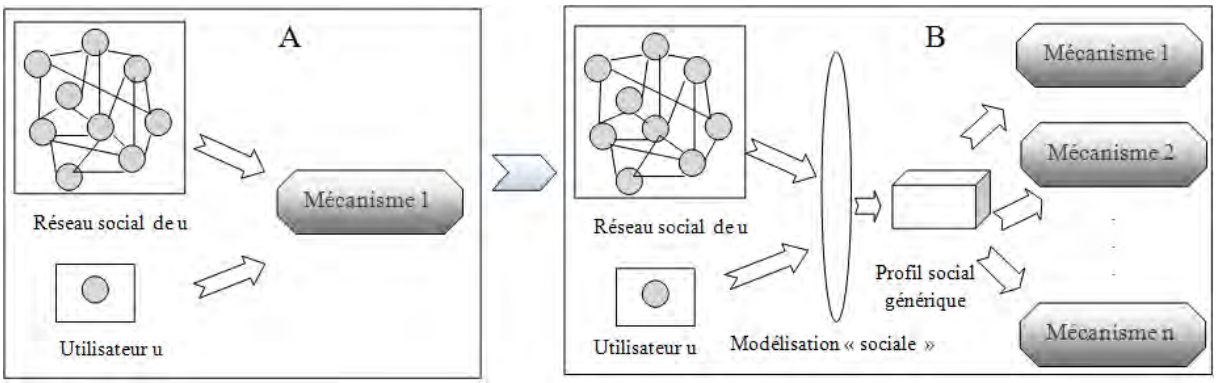



Figure 4. 2 : Usage du réseau social sans (A) et avec (B) modélisation du profil social de l'utilisateur

### *4.2.3.2 Réseaux égocentriques et k-égocentriques*

Pour concevoir le modèle du profil générique et social de l'utilisateur, nous nous appuyons sur l'hypothèse de la pertinence des communautés du réseau égocentrique de l'utilisateur. Cependant avant de présenter le modèle, il est important de bien cerner la notion de réseau égocentrique présenté dans la section 4.2.1. Dans la littérature, certains travaux considèrent comme réseau égocentrique le réseau constitué de l'égo et de tous ses alters. D'autres travaux considèrent en plus, les relations entre alters comme faisant partie du réseau égocentrique. Cette seconde définition du réseau égocentrique correspond mieux à celle qui permettrait d'évaluer notre hypothèse. Cependant selon notre hypothèse, ce ne sont pas les relations entre l'égo et ses alters qui sont utiles, mais uniquement les relations entre les alters. ***Nous définissons dans nos travaux, le réseau égocentrique d'un utilisateur, comme le réseau constitué des relations entre ses alters dans le réseau social entier***. Si l'on considère que le réseau social entier est représenté par un graphe G(V, E), V étant l'ensemble des individus et E étant l'ensemble des liens entre ces individus, le réseau égocentrique d'un utilisateur v noté $R_v$ est défini par le sous graphe $G_v'(V', E')$ tel que :

$$\mathbf{R_v} = G_v'(V', E') \subset G(V, E) \,/\, V'=\{ u \in V \,/\, d_G(u, v)=1\} \text{ et } E'=\{ (u, x) \in E, u \text{ et } x \in V'\} \quad \mathbf{(17)}$$

v est l'égo, $d_G(u, v)$ représente la distance dans G entre les individus u et v.

La figure 4.3 est un exemple de représentation graphique en communautés (réalisé manuellement) d'un réseau égocentrique correspondant à cette définition par le sociologue Dominique Cardon (Cardon, 05).

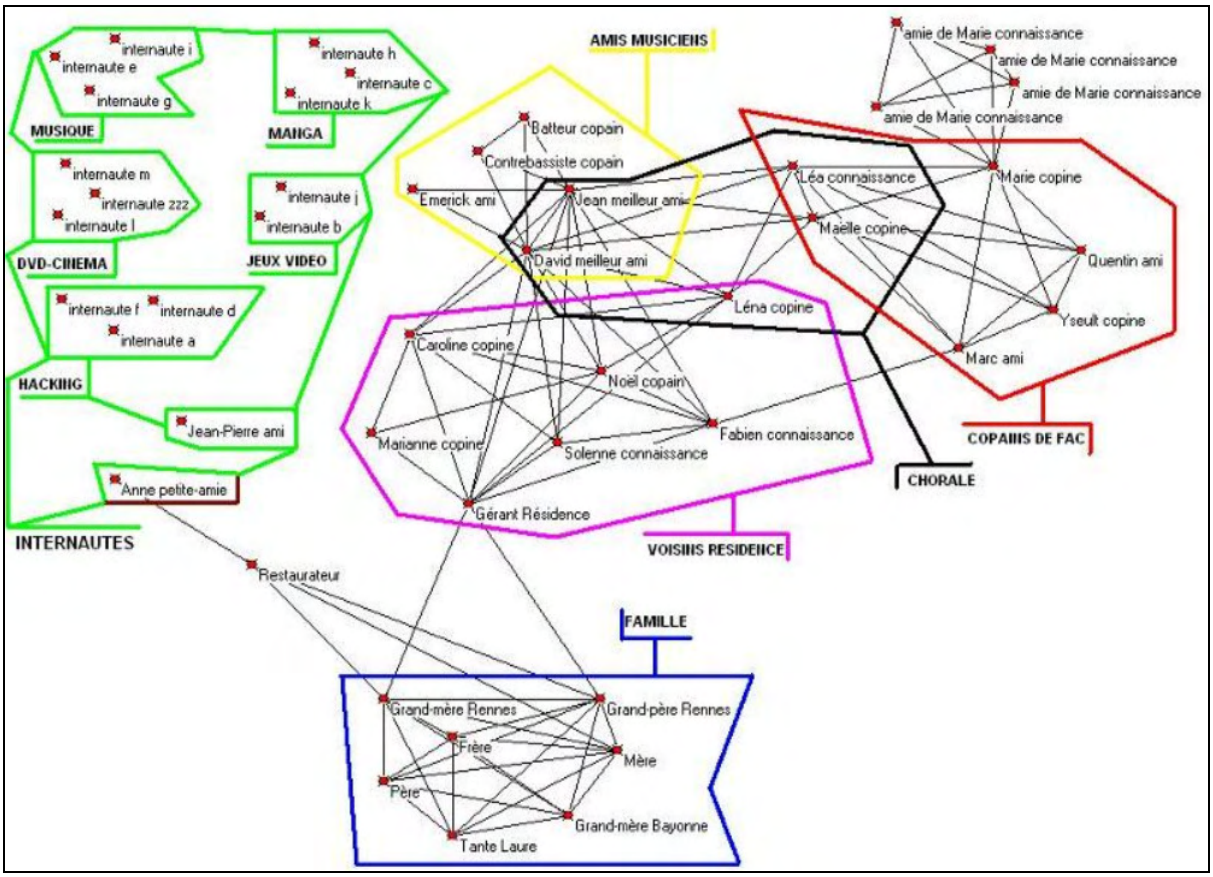



Figure 4. 3 : Exemple de représentation manuelle de communautés dans un réseau égocentrique

Cette représentation est un exemple typique de communautés d'un réseau égocentrique pouvant être exploité pour dériver le profil de l'égo selon notre hypothèse. Il faut tout de même noter que cette représentation n'est pas assez réaliste car dans la réalité il est très courant d'avoir des communautés qui se chevauchent dans un pareil réseau (Cazabet et al., 10). L'égo étant par définition lié à tous les alters, il n'est pas représenté dans le réseau égocentrique. Pour un graphe non pondéré cette définition est suffisante pour représenter l'information nécessaire à l'évaluation de notre hypothèse.

Pour un graphe pondéré, on pourrait également penser à des techniques de dérivation du profil social de l'utilisateur qui s'appuieraient à la fois sur la force des liens entre l'égo et les alters (et aussi entre alters) et sur les communautés du réseau égocentrique. Dans ces cas, afin de prendre en compte la force des liens entre l'égo et ses alters, l'égo peut être inclu dans le réseau égocentrique, que nous appelons ici ***réseau égocentrique étendu*** ($\mathbf{R_v^e}$). ***Un pareil réseau se défini comme une extension d'un réseau égocentrique en rajoutant l'égo et ses liens avec ses alters,*** formule (18).

$$R_v^e = G_v'(V' \cup \{v\}, E') \subset G(V, E) \,/\, V'=\{u \in V \,/\, d_G(u, v)=1\} \text{ et } E'=\{(u, x) \in E, u \text{ et } x \in V'\} \quad \textbf{(18)}$$

v est l'égo, $d_G(u, v)$ représente la distance dans G entre les individus u et v.

Ces deux définitions de réseau égocentrique et de réseau égocentrique étendu considèrent comme alter les individus à distance 1 de l'utilisateur dans le réseau social entier. Ceci correspond au cas idéal tel qu'on peut l'avoir dans le monde réel. Toutefois dans les systèmes d'information, les liens entre utilisateurs se créent au fil du temps et de leurs activités. Si on prend le cas de Facebook par exemple, un utilisateur qui crée son compte reconstitue son réseau social du monde réel au fil du temps. Il est donc courant de rencontrer des individus ayant très peu d'alters du monde réel à distance 1 dans le réseau social connu du système d'information. Il en découlera une difficulté à retrouver des communautés dans son réseau égocentrique qui aura une structure éparse et non cohésive. Si le réseau à distance 1 de l'utilisateur est très éparse, on peut penser aller un peu plus loin dans le graphe pour rechercher des communautés autour de l'utilisateur (Barrat et al., 10). Ainsi, pour un réseau égocentrique ayant peu d'alters, on pourrait inclure des alters potentiels de l'utilisateur situés à une distance k >=1(k∈ N) dans le graphe et évaluer la pertinence des communautés qui en découlent. ***Nous définissons ainsi des réseaux k-égocentriques d'un utilisateur (égo) comme étant le réseau des relations entre alters et alters potentiels de l'égo situés à une distance maximum k dans le réseau social entier***, formule (19).

$$R_v(k) = G_v'(V', E') \subset G(V, E) \,/\, V'=\{u \in V \,/\, d_G(u, v)<=k\} \text{ et } E'=\{(u, x) \in E, u \text{ et } x \in V'\} \quad \textbf{(19)}$$

Dans le cas des graphes pondérés, on pourra également distinguer les alters directs de l'égo et obtenir un ***réseau k-égocentrique étendu,*** formule (20).

$$R_v^e(k) = G_v'(V' \cup \{v\}, E') \subset G(V, E) \,/\, V'=\{u \in V \,/\, d_G(u, v)<=k\} \text{ et } E'=\{(u, x) \in E, u \text{ et } x \in V'\} \quad \textbf{(20)}$$

Afin de garder à la fois des communautés pertinentes et une complexité de calcul raisonnable, les valeurs de k devraient à priori rester assez faibles (2 ou 3 par exemple). Le modèle de profil proposé à partir de la notion de communautés dans les réseaux k-égocentriques est présenté dans la section qui suit.

## 4.2.4 Modèle proposé

Dans un but de généricité et de prise en compte efficace du réseau social de l'utilisateur, nous proposons le modèle de profil utilisateur dont la représentation UML est présentée sur la figure 4.4 (Tchuente et al., 12). De manière globale, ce modèle est constitué de deux dimensions (dimension utilisateur et dimension sociale) elles-mêmes constituées d'un ensemble d'attributs. Nous détaillons les composants de ce modèle dans les sections qui suivent.

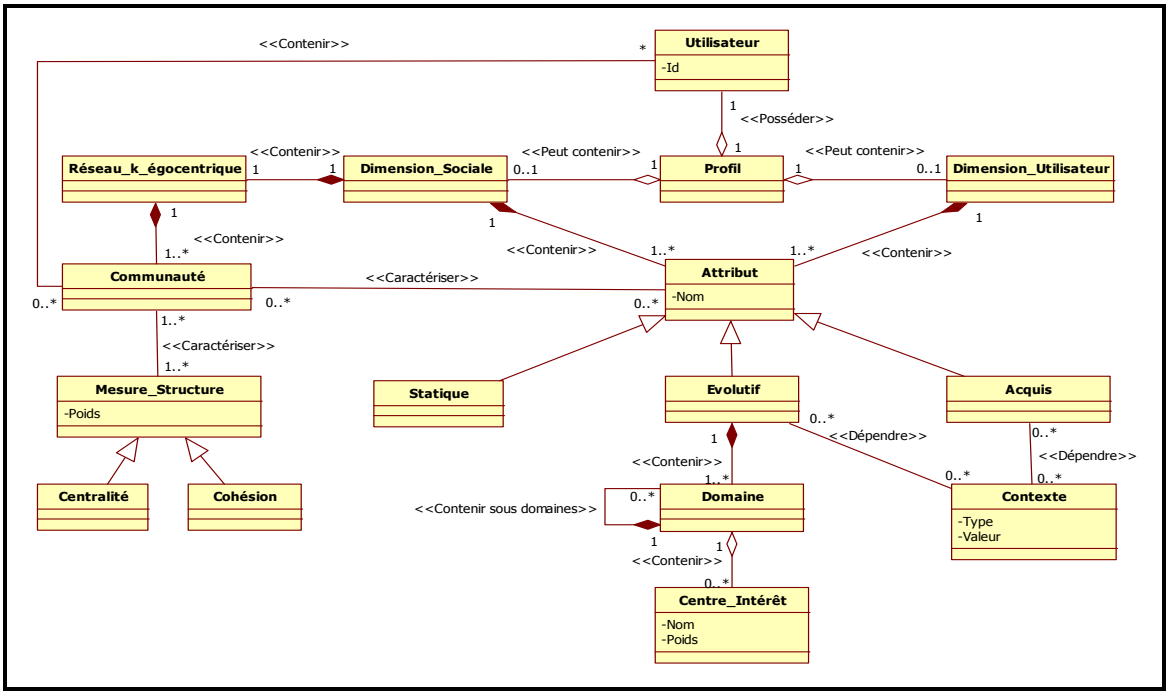



Figure 4. 4 : modèle générique social du profil utilisateur

### *4.2.4.1 La dimension sociale et la dimension utilisateur*

***La dimension utilisateur*** du profil est une dimension qui contiendra les éléments construits à partir des informations et interactions de l'utilisateur avec le système. Cette dimension peut être construite par les approches classiques de construction de profils présentés dans le chapitre 1. C'est la dimension la plus importante du profil qui doit être utilisée en priorité par un mécanisme d'adaptation de l'information. Toutefois, s'il arrive par exemple qu'il manque une information dans cette dimension, la dimension sociale pourra alors être exploitée par le mécanisme.

***La dimension sociale*** du profil est une dimension qui contiendra les éléments construits à partir des informations et interactions des communautés d'utilisateurs dans le réseau k-égocentrique de l'utilisateur qui est profilé (l'égo). L'information présente dans cette dimension est une information supplémentaire et/ou complémentaire aux informations de la dimension utilisateur qui pourra être exploitée en fonction des besoins des mécanismes d'adaptation de l'information à l'utilisateur. Il est à noter que nous parlons juste de la représentation de l'information dans cette dimension et pas de la manière dont elle pourra être exploitée. Ceci dépendra de chaque mécanisme : un mécanisme pourra par exemple recourir à l'information contenue dans la dimension sociale si la dimension utilisateur ne contient pas assez d'information. Un autre mécanisme pourra

aussi par exemple intégrer systématiquement les informations de cette dimension en plus de la dimension utilisateur dans son processus d'adaptation.

Comme indiqué dans la section précédente, nous construisons la dimension sociale du profil à partir des communautés présentes dans le réseau k-égocentrique de l'utilisateur. Chacune de ces communautés peut être caractérisée de deux manières :

- ***Sa structure interne ou externe vis-à-vis de la structure du réseau k-égocentrique*** **:** que nous appellerons « *structure de la communauté* » : la structure interne d'une communauté dépendra par exemple de la proportion de nœuds connectés dans la communauté (cohésion) et la structure externe dépendra par exemple de la centralité de la communauté dans le réseau k-égocentrique. Une communauté étant un groupe, sa centralité pourra être calculée par les métriques d'extension de centralité aux groupes d'utilisateurs dans les réseaux sociaux (cf. section 3.3.6.2, Everrett et Borgatti, 99). Nous reviendrons sur ces métriques de structure de communautés dans la section sur les algorithmes de dérivation de la dimension sociale. Ces métriques sont importantes car il va de soi que les communautés dans le réseau k-égocentrique d'un utilisateur n'ont pas la même importance. Une communauté très centrale et une communauté complètement isolée dans le réseau k-égocentrique n'ont certainement pas la même signification pour l'utilisateur. Pour le moment nous n'émettons pas d'hypothèse sur l'importance relative de ces métriques de structure de communautés dans le profil de l'utilisateur. Leur impact et importance seront évalués expérimentalement.

- ***Son profil*** **:** que nous appellerons « *sémantique de la communauté* » : à partir de la dimension utilisateur des profils des individus d'une communauté, la sémantique de la communauté peut être déduite en réunissant (agrégation) toutes les informations qu'ils contiennent. Il est important de noter que c'est la dimension utilisateur (et non la dimension sociale) du profil des alters qui permettra de dériver la dimension sociale du profil de l'égo.

La dimension sociale et la dimension utilisateur du profil sont toutes les deux représentées par un même ensemble d'attributs. Ceci les rend comparables pour une exploitation par les mécanismes d'adaptation de l'information ou pour des évaluations. Nous présentons les attributs des deux dimensions du profil dans la section qui suit.

### *4.2.4.2 Les attributs*

Nous représentons chacune des dimensions du profil comme étant constituée d'un ensemble d'attributs. Nous considérons trois principaux types d'attributs relatifs à la nature des communautés pouvant être extraits du réseau k-égocentrique de l'utilisateur : les attributs statiques, les attributs acquis et les attributs évolutifs.

#### *4.2.4.2.1 Les attributs statiques*

Les attributs statiques sont les attributs de l'utilisateur qui ne varient pas au fil du temps : nom, genre, année de naissance, couleur des yeux, etc. Du point de vue des communautés, ces attributs peuvent permettre de caractériser des communautés qui seront dites « statiques » dans le réseau k-égocentrique de l'utilisateur. Par exemple une communauté de personnes ayant toutes un même nom pourra être assimilée à une famille dans la vie réelle. Les attributs statiques sont généralement renseignés explicitement par les utilisateurs et sont très utiles pour des mécanismes tels que le filtrage à base de règles (Liang et al., 02). Ces attributs peuvent être

représentés de deux manières selon qu'ils sont dans la dimension utilisateur ($Att_s^{dimU}$) ou dans la dimension sociale ($Att_s^{dimS}$) du profil :

$$\begin{cases} Att_s^{dimU} = \{ (nomAttribut, ValeurAttribut) \} \\ Att_s^{dimS} = \{ (nomAttribut, ValeurAttribut, P^{social}) \} \end{cases} \quad \textbf{(21)}$$

Les attributs statiques de la dimension utilisateur sont représentés par un ensemble de couples (attribut, valeur). Par exemple pour un utilisateur de genre féminin on aura le couple *(Genre, Féminin)*. Les attributs statiques de la dimension sociale sont représentés par des triplets (attribut, valeur, $P^{social}$). Pour chacune des valeurs d'un attribut, sa pertinence pour l'utilisateur est calculée par une probabilité (ou poids) $P^{social}$ des valeurs de l'attribut dans les communautés du réseau égocentrique de l'utilisateur. Par exemple pour un utilisateur de genre inconnu, dans la dimension sociale de son profil, on pourra avoir les deux triplets *(Genre, Féminin, 0.1)* et *(Genre, Masculin, 0.8)* qui indique qu'il est plus probable (valeur 0.8) que l'utilisateur en question soit de genre masculin d'après son réseau social. Nous reviendrons sur le calcul de cette probabilité (ou poids) dans la section sur les algorithmes.

#### *4.2.4.2.2 Les attributs acquis*

*Les attributs acquis* sont les attributs que l'utilisateur acquiert au fil du temps et qui restent permanents une fois qu'ils sont acquis. C'est par exemple le cas d'un employeur ou d'un établissement fréquenté par l'utilisateur. Du point de vue des communautés, ce type d'attributs pourra permettre de caractériser des communautés qui seront dites « acquises » dans le réseau k-égocentrique de l'utilisateur (collègues, camarades de classes, etc.). Ces attributs peuvent être associés à des éléments de *contexte* (*lieu, temps,* etc.) donnant des informations sur l'environnement dans lequel ils ont été acquis. Les attributs acquis sont éventuellement calculés par projection sur des éléments de contexte. Etant donné une valeur d'élément de contexte noté $ValC_i$, les attributs acquis peuvent également être représentés de deux manières selon qu'ils sont dans la dimension utilisateur $Att_a^{dimU}(ValC_i)$ ou dans la dimension sociale $Att_a^{dimS}(ValC_i)$ du profil :

$$\begin{cases} Att_a^{dimU}(ValC_i) = \{(nomAttribut, valeurAttribut(ValC_i))\} \\ Att_a^{dimS}(ValC_i) = \{(nomAttribut, ValeurAttribut(ValC_i), p^{social}(ValC_i))\} \end{cases} \quad \textbf{(22)}$$

Si l'on suppose par exemple comme valeur d'élément de contexte temporel l'intervalle d'année [2009-2012] d'un attribut acquis *employeur*, pour un individu, ces attributs peuvent par exemple prendre les valeurs :

$Att_a^{dimU}$ [2009-2012] = *{(Employeur, IRIT)}*

$Att_a^{dimS}$ [2009-2012] = *{(Employeur, IRIT, 0.8), (Employeur, La Poste, 0.5)}*

Et l'on pourrait par exemple en déduire que même si l'utilisateur est employé par *IRIT*, il est probable de par la dimension sociale de son profil qu'il ait également été en relation avec l'employeur *La Poste.*

#### *4.2.4.2.3 Les attributs évolutifs*

Les attributs évolutifs représentent les centres d'intérêts de l'utilisateur qui évoluent dans le temps au fur et à mesure de ses interactions avec le système. Comme nous l'avons vu dans le chapitre 1, les centres d'intérêts des utilisateurs peuvent être représentés de plusieurs manières (vecteur de termes pondérés, réseaux sémantiques, ontologies, etc.). Afin de rester assez générique dans notre modèle, nous avons choisi une

représentation hiérarchique des centres d'intérêts suivant une taxonomie préexistante de concepts. Cette représentation a l'avantage (comme on le verra dans la suite) de structurer les vecteurs de termes pondérés du profil suivant une hiérarchie pouvant permettre de désambigüiser certains centres d'intérêts suivant des domaines d'intérêts distincts, et surtout de permettre aux mécanismes d'exploiter le profil de l'utilisateur suivant plusieurs niveaux de granularité. De plus, une taxonomie est une structure de données plus facile à mettre en place par des experts d'un domaine par rapport aux ontologies multi-relationnelles (Brut et al., 10)(Daoud et al., 08). De même que les attributs acquis, les attributs évolutifs sont éventuellement calculés par projection sur des éléments de contexte et suivant un domaine (feuille) existant dans la taxonomie. Pour le profil évolutif, au lieu de parler d'attribut, on parlera plutôt ici de *domaine* de la taxonomie, au lieu de parler de valeur de l'attribut, on parlera de *centre d'intérêt (Interet)* dans un domaine. Etant donnée une valeur d'élément de contexte noté $ValC_i$, les attributs évolutifs peuvent également être représentés de deux manières selon qu'ils sont dans la dimension utilisateur $Att_e^{dimU}(ValC_i)$ ou dans la dimension sociale $Att_e^{dimS}(ValC_i)$ du profil :

$$\begin{cases} \text{Att}_e^{\text{dimU}}(ValC_{i,}) = \{(nomDomaine,\ valeurInteret(ValC_{i,}),\ P\ (ValC_{i,}))\} \\ \text{Att}_e^{\text{dimS}}\ (ValC_i) = \{(nomDomaine,\ valeurInteret(ValC_{i,}), P^{social}(ValC_{i,})\} \end{cases} \quad \textbf{(23)}$$

Comme dans le cas des attributs acquis, les attributs évolutifs peuvent être évalués en fonction d'une valeur d'élément de contexte noté $ValC_i$.

La figure 4.5 présente un exemple de profil avec des attributs évolutifs (centres d'intérêts) hiérarchisés suivant une taxonomie prédéfinie à un seul niveau (la racine « Adaptation de l'information à l'utilisateur » et ses trois sous-domaines « modélisation de l'utilisateur », « modélisation des contenus » et « techniques de filtrage »). Ce profil correspond à la valeur de contexte temporel qui est l'intervalle d'années [2009, 2012]. Les centres d'intérêt sont calculés à partir des domaines feuilles de la taxonomie et sont par la suite remontés jusqu'à la racine. Par exemple, le domaine modélisation de l'utilisateur comporte trois centres d'intérêts (Social, 0.3), (Graphes 0.2), et (TextMining, 0.1). A partir des centres d'intérêts, les domaines sont également pondérés par sommation des poids des centres d'intérêts qu'ils contiennent (0.3+0.2+0.1=0.6 pour « modélisation de l'utilisateur » par exemple). Le profil peut ainsi être évalué à des niveaux de granularité plus ou moins fins permettant une flexibilité importante dans son exploitation par les mécanismes d'adaptation. Par exemple pour l'utilisateur dont le profil est présenté sur la figure 4.5, parmi les domaines préexistants de la taxonomie, il est plus intéressé par la « modélisation de l'utilisateur » (0.6) que par la « modélisation des contenus » (0.2) et les « techniques de filtrage » (0.2). D'autre part, la représentation hiérarchique peut permettre de désambigüiser les centres d'intérêts dans un profil. Sur l'exemple ci-dessous, à la racine on constate que l'utilisateur est intéressé au « Social » (0.4), mais en allant plus loin dans l'arborescence le « Social » concernant la « modélisation de l'utilisateur » (0.3) est différent (et plus important pour l'utilisateur) du « Social » concernant les « techniques de filtrage » (0.1).

L'algorithme de propagation des valeurs des centres d'intérêts d'un domaine contenant des sous-domaines consiste tout simplement à remonter les centres d'intérêts des sous-domaines en conservant leur poids, ou en faisant la somme des poids si le centre d'intérêt est présent dans plusieurs domaines fils. Dans l'exemple de la figure 4.5, à la racine le poids 0.4 de « Social » est calculé en sommant les poids 0.3 et 0.1 des sous-domaines « Modélisation Utilisateur » et « Techniques de filtrage » respectivement.

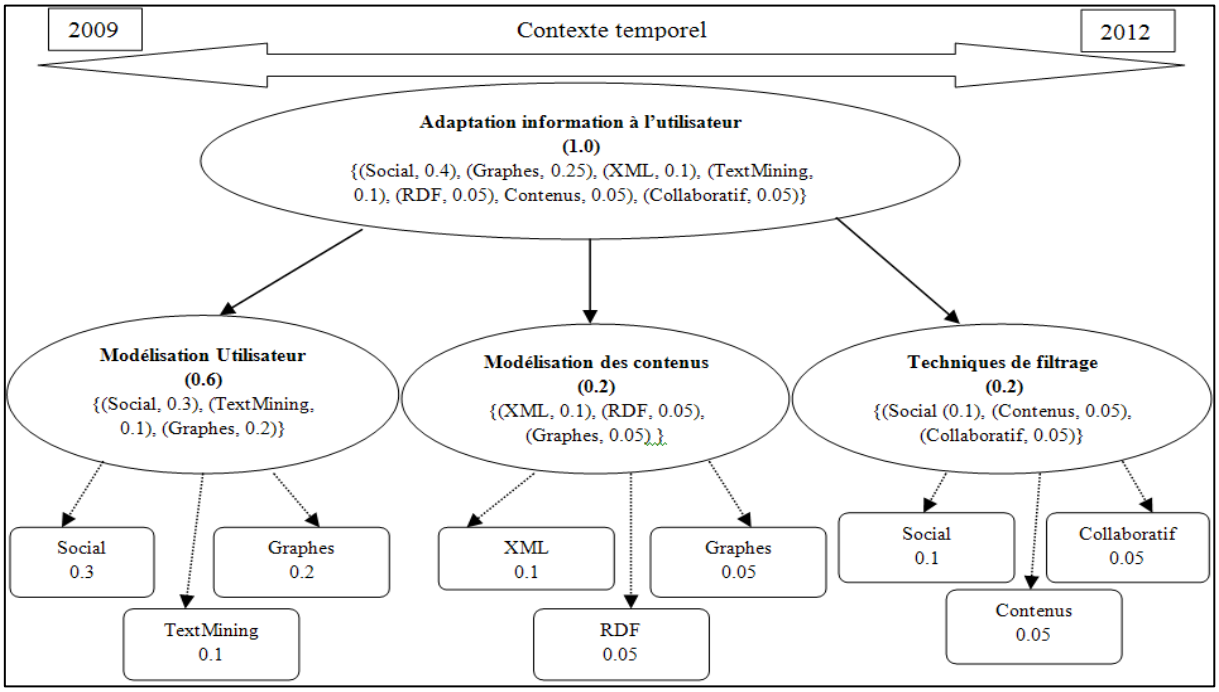



Figure 4. 5 : Exemple de profil évolutif hiérarchisé suivant une taxonomie

## 4.3 Processus et algorithme de dérivation de la dimension sociale du profil à partir des communautés du réseau k-égocentrique

Dans le modèle de profil présenté dans la section précédente, il est possible de dériver plusieurs instances d'attributs acquis et évolutifs en fonction des éléments contextuels. Toutefois, prendre en compte des éléments contextuels permet de restreindre la sélection de données en entrée des algorithmes de calcul du profil. Les algorithmes de calcul restent inchangés que l'on considère ou non des éléments contextuels. Dans la suite, nous nous intéresserons donc uniquement aux algorithmes de calcul des attributs des éléments du profil sans prendre en compte la notion de contexte. Le lecteur intéressé par la manière avec laquelle nous gérons les éléments contextuels (notamment le contexte temporel) peut se référer à (Tchuente et al., 10)(Tchuente et al., 12) (cf. Annexe A.3).

Nous nous intéressons dans cette section au processus de calcul des poids (probabilités) des attributs de la dimension sociale du profil. Les attributs de la dimension utilisateur sont en général renseignés de manière explicite par l'utilisateur (attributs statiques et acquis) ou construits à partir des activités de l'utilisateur (les attributs ou centres d'intérêts sont pondérés par les mesures telles que *tf* ou *tf.idf*, cf. formule 1). Pour dériver les attributs de la dimension sociale du profil ainsi que leur pondération, nous allons présenter dans cette partie, le processus s'appuyant sur la notion de communautés dans un réseau k-égocentrique défini précédemment dans ce document. L'algorithme qui découle de ce processus (notée $CoSP_k$, *Communities Social Profiles from k-egocentrics networks*) sera également présenté.

### 4.3.1 Processus de dérivation de la dimension sociale basé sur les communautés du réseau k-égocentrique

Le but du processus est de déterminer pour chaque couple atttribut valeur ($A_x$, $V_x$), son poids $P_{AxVx}^{social}$ du triplet ($A_x$, $V_x$, $P_{AxVx}^{social}$) dans la dimension sociale du profil de l'utilisateur. x représente ici soit s (pour les attributs statiques), soit a (pour les attributs acquis), soit e (pour les attributs évolutifs). Le processus s'articule autour de quatre principales étapes successives qui vont permettre de dériver les attributs de la dimension sociale du profil :

- **détection de communautés dans le réseau k-égocentrique,**
- **profilage des communautés détectées,**
- **caractérisation des communautés,**
- **dérivation des attributs de la dimension sociale.**

Ce processus est illustré sur la figure 4.6. Nous présentons chacune des étapes dans ce qui suit.

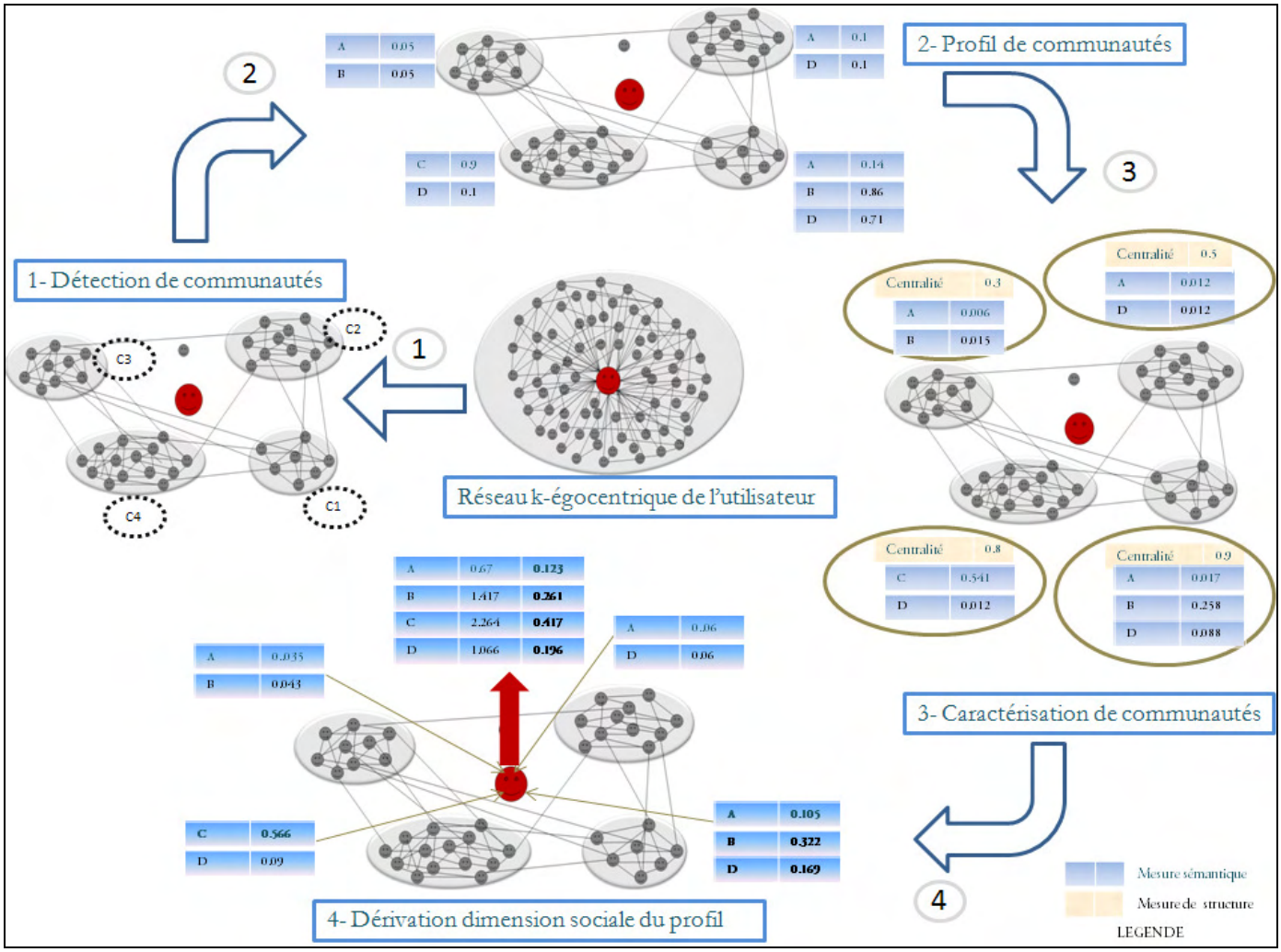



Figure 4. 6: Illustration du processus de dérivation de la dimension sociale à partir de communautés

### *4.3.1.1 Etape de détection de communautés dans le réseau k-égocentrique*

#### *4.3.1.1.1 Description*

Le point de départ du processus est le réseau k-égocentrique de l'utilisateur (étendu ou non). L'égo est représenté en rouge sur la figure, et ses alters en noir. Comme dans tout graphe, un algorithme de détection de communautés peut être appliqué au réseau k-égocentrique pour extraire des communautés d'utilisateurs. La question ici est de savoir quel serait le ou les algorithmes appropriés dans le cas des réseaux k-égocentriques. Comme indiqué dans la section 3.3.7 sur la détection de communautés dans les graphes, ces algorithmes diffèrent selon les principaux paramètres suivants : la gestion du recouvrement, la fixation à priori de la taille minimale de communautés à rechercher, les types de données de validation des algorithmes, les techniques de construction et d'évaluation des communautés, la prise en compte de la dynamique des réseaux. Le tableau 4.1 présente comment doivent être considérés ces critères dans le cas des réseaux k-égocentriques. Globalement, il ressort de ce tableau que les critères essentiels des algorithmes à considérer dans notre cas sont : conception purement orientée réseaux sociaux (sens sociologique) de l'algorithme, validation sur des données les plus réelles possible et par les utilisateurs eux-mêmes, prise en compte du recouvrement et de la dynamique des communautés.

| **Critère** | **Evaluation pour un réseau k-égocentrique** |
|---|---|
| Recouvrement de communautés (RC) | A prendre forcément en compte car dans la vie réelle les communautés autour d'un utilisateur se recouvrent très souvent. |
| Fixation taille minimale de communautés (FTMC) | La valeur doit être très faible compte tenu de la structure microscopique (petits graphes de quelques dizaines ou centaines de nœuds) des réseaux k-égocentriques. Le strict minimum (communauté de 2 individus) doit être pris en compte ici pour refléter toutes les éventualités de la vie réelle. |
| Types de données de validation (TDV) | Dans la mesure où l'approche se base sur les réseaux dans la vie réelle, il est important de faire des validations par des jugements des utilisateurs eux-mêmes à partir des réseaux sociaux (transposés dans des SI) susceptibles de représenter au mieux leur réseau dans la vie réelle. |
| Techniques de construction et évaluation des communautés (TCEC) | Considérer des techniques qui s'appuient sur des critères qui se rapprochent des sciences sociales telles que la cohésion sociale (Friggeri et al., 11). Toutefois, ce n'est pas la technique en elle-même qui est plus importante, c'est sa validation et sa pertinence par les utilisateurs qui est plus importante ici. |
| Prise en compte de la dynamique du réseau (DR) | Très important pour réduire la complexité des calculs qui peuvent être liés à la réexécution des algorithmes à chaque fois que le réseau k-égocentrique d'un utilisateur évolue. L'évolution du réseau doit automatiquement être prise en compte par l'algorithme |

Tableau 4. 1 : Critères d'évaluation des algorithmes de détection de communautés pour réseaux k-égocentriques

Pour rechercher le ou les algorithmes qui conviendraient ici, nous avons comparé chacun des critères énoncés précédemment sur des algorithmes ayant déjà été testés par les utilisateurs dans les réseaux sociaux numériques (Cazabet et al., 12)(Friggeri et al., 11). Le tableau 4.2 présente cette comparaison à partir de laquelle on peut déduire que l'algorithme iLCD (Cazabet et al., 10) serait le plus approuvé. Même si cet

algorithme n'est pas intégralement élaboré suivant des analyses sociologiques, il remplit parfaitement tous les autres critères. L'algorithme de cohésion sociale (Friggeri et al., 11) est par contre élaboré suivant des analyses sociologiques, mais son principal inconvénient est la non prise en compte automatique de la dynamique du réseau. L'algorithme *InfoMap* est très bien évalué par les utilisateurs, mais son principal inconvénient ici, est qu'il ne prend pas en compte le recouvrement des communautés. Enfin, l'algorithme *CFinder* bien qu'il gère très bien le recouvrement et la taille minimale des communautés (k-cliques) à détecter, il ne gère pas automatiquement la dynamique des communautés et les retours utilisateurs ne sont pas aussi satisfaisants que les autres algorithmes.

Dans nos travaux, nous avons donc choisi de réutiliser l'algorithme iLCD qui remplit le mieux les conditions nécessaires exprimées dans le tableau 4.2.

|  | RC | FTMC | TDV | TCEC | DR | **Solution adéquate** |
|---|---|---|---|---|---|---|
| iLCD (Cazabet et al.,10) |  |  |  |  |  |  |
| Cohésion sociale (Friggeri et al., 11) |  |  |  |  |  |  |
| InfoMap (RosVald et al., 07) |  |  |  |  |  |  |
| CFinder (Palla et al., 05) |  |  |  |  |  |  |

Tableau 4. 2: Comparaison de quelques algorithmes de détection de communautés évalués sur les réseaux sociaux numériques

#### *4.3.1.1.2 Exemple : communautés détectées sur l'illustration*

Sur l'illustration de la figure 4.6, l'algorithme de détection de communautés utilisé a par exemple détecté quatre communautés dans le réseau k-égocentrique de l'utilisateur en rouge.

Une fois cette étape réalisée, pour tout couple attribut-valeur $(A_x, V_x)$ nous allons calculer ses poids successifs *(P', P'', P''')* associés à chacune des trois étapes suivantes du processus jusqu'à son poids final $P^{social}{}_{AxVx}$ dans la dimension sociale du profil de l'utilisateur, tel qu'indiqué dans le tableau 4.3 ci-dessous :

| Couple $(A_x, V_x)$ de la dimension sociale | Poids qui sera associé à l'étape « profilage de communautés » | Poids qui seront associés à l'étape « caractérisation de communauté » | Poids qui sera associé à l'étape « dérivation de la dimension sociale » |
|---|---|---|---|
| | $P'_{AxVx}$ | $P''_{AxVx}$ et $P'''_{AxVx}$ | $P^{social}{}_{AxVx}$ |

Tableau 4. 3 : Poids successifs des couples attribut-valeur en fonction des étapes du processus

Chacune de ces étapes et la description du calcul des poids associés sont décrites dans les sections qui suivent.

### *4.3.1.2 Etape de profilage des communautés*

#### *4.3.1.2.1 Description*

Une fois les communautés détectées, le profil de chacune des communautés est construit par agrégation des valeurs de la dimension utilisateur des profils de tous les membres de la communauté. Considérons une communauté *C* comportant *m* utilisateurs, x désignant un type d'attribut x∈ {statique (s), acquis (a), évolutif(e)}, et un couple attribut-valeur $(A_x, V_x)$ appartenant à la dimension utilisateur du profil d'au moins un utilisateur de cette communauté. Le profil de la communauté C est calculé suivant la fonction $f_x^{profil}$ définie par[46] :

$$f_x^{profil} : (A_x, V_x) \in C \mapsto (A_x, V_x, P'_{AxVx}(C))$$

$$P'_{AxVx}(C) = \frac{\sum_{u=1}^{m} M_x(A_x, V_x, Att_x^{\dim U}(u))}{m} \quad \textbf{(24)}$$

Si (x=a ou x=s) $M_x(A_x, V_x, Att_x^{\dim U}(u)) = \begin{cases} 0 \text{ si } (A_x, V_x) \notin Att_x^{\dim U}(u) \\ 1 \text{ sinon} \end{cases}$

Si (x=e) $M_x(A_x, V_x, Att_x^{\dim U}(u)) = \begin{cases} P_{AxVx}(u) \text{ si } (A_x, V_x) \in Att_x^{\dim U}(u) \\ 0 \text{ sinon} \end{cases}$

$P'_{AxVx}(C)$ est déterminé comme la moyenne (ou valeur médiane) de la valeur $V_x$ pour l'attribut $A_x$ dans les dimensions utilisateur des profils des membres $(u)$ de la communauté C.

Pour les attributs statiques et acquis (x=s ou x=a), les poids associés sont soit 0 ou 1 selon qu'ils apparaissent ou pas dans la dimension utilisateur d'un profil. Par conséquent, $P'_{AxVx}(C)$ pour ces attributs est le nombre moyen d'apparitions du couple $(A_x, V_x)$ dans les dimensions utilisateur des profils des membres de la communauté.

Pour les attributs évolutifs (x=e), les poids associés valent 0 s'ils n'apparaissent pas dans la dimension utilisateur d'un profil, et sont compris dans ]0, 1] dans le cas contraire. Par conséquent, $P'_{AxVx}(C)$ est la moyenne des poids associés au couple $(A_x, V_x)$ dans les dimensions utilisateur des profils des membres de la communauté.

#### *4.3.1.2.2 Exemple : calcul du profil d'une communauté*

Par exemple, supposons que la communauté C1 (figure 4.6) contient 7 utilisateurs u1,.., u7, dont les dimensions utilisateurs des profils ainsi que le calcul du profil de la communauté sont présentés dans le tableau 4.4.

[46] Pour une communauté C, lorsqu'on note $(A_x, V_x) \in C$, ceci indique simplement que le couple attribut valeur $(A_x, V_x)$ est un élément du profil de la communauté C. Lorsqu'on note $(A_x, V_x, P) \in C$, ceci indique que le couple attribut valeur $(A_x, V_x)$ est pondéré par le poids P dans le profil de la communauté C.

| Attribut | U1 | U2 | U3 | U4 | U5 | U6 | U7 | $\mathbf{P'_{AxVx}}$ |
|---|---|---|---|---|---|---|---|---|
| Genre[47] *(statique)* | M | F | F | F | F | F | F | (Genre, M) = 1/7=0.14, (Genre, M) noté A (figure 4.6)<br>(Genre, F) = (1+1+1+1+1+1)/7= 0.86, (Genre, F) noté B (figure 4.6) |
| Langage programmation (LP) *(évolutif)* | Java (0.9) | Java (0.8) | Java (0.7) | Java (0.9) | Java (0.8) | | Java (0.9) | (LP, Java)= (0.9+0.8+0.7+0.9+0.8+0.9)/7=0.71, (LP, Java) noté D (figure 4.6) |

Tableau 4. 4: Exemple de calcul de profil de communauté

De la même manière, les profils des communautés C2, C3 et C4 sont calculés et présentés dans la figure 4.6.

### *4.3.1.3 Etape de caractérisation (sémantico-structurelle) des communautés*

#### *4.3.1.3.1 Description*

Deux types de caractérisation sont recherchés dans cette étape pour chaque communauté : la caractérisation sémantique et la caractérisation structurelle. Ces deux caractérisations sont ensuite fusionnées pour avoir une caractérisation unique (sémantico-structurelle). Nous présentons tour à tour dans cette section la caractérisation sémantique, la caractérisation structurelle, et la caractérisation sémantico-structurelle.

#### *4.3.1.3.2 Caractérisation sémantique*

***La caractérisation sémantique*** d'une communauté utilise les profils de toutes les communautés et consiste à rechercher sa spécificité par rapport aux autres communautés. En d'autres termes, il s'agit de chercher ce qui différencie chaque communauté relativement aux autres communautés. Cette différence représentera mieux l'affinité intrinsèque entre les membres de cette communauté dans le réseau k-égocentrique. Pour réaliser cette caractérisation, une analogie peut être faite avec la mesure de pondération tf.idf utilisée dans les systèmes documentaires pour déterminer la pertinence des termes dans un document précis dans un corpus de documents (Salton, 83). Si on considère qu'il existe *m* communautés ($C_1, C_2, \ldots, C_m$) dans le réseau k-égocentrique de l'utilisateur et que $(A_x, V_x, P'_{AxVx}(C_i))$ est un élément du profil de la communauté $C_i$, cet élément du profil sera caractérisé sous la forme $(A_x, V_x, P''_{AxVx}(C_i))$ dans laquelle $P''_{AxVx}(C_i)$ est déterminé par la fonction $f_x^{semantique}$:

$$f_x^{semantique} : (A_x, V_x, P'_{AxVx}(C_i)) \in C_i \mapsto (A_x, V_x, P''_{AxVx}(C_i))$$

$$P''_{AxVx}(C_i) = P'_{AxVx}(C_i) * \log \frac{m}{\sum_{j=1}^{m} presence(A_x, V_x, C_j)} \qquad \textbf{(25)}$$

$$\text{Avec } presence(A_x, V_x, C_j) = \begin{cases} 1 \ si \ (A_x, V_x) \in P'_{AxVx}(C_j) \\ 0 \ sinon \end{cases}$$

[47] Il est important de noter qu'un attribut statique tel que le genre (masculin ou féminin) n'est pas à priori prédictible dans la vie réelle (un individu est soit de genre masculin ou féminin). Le poids des valeurs indiquées ici (P') pour les valeurs de ces attributs représentent des probabilités qui nous serviront uniquement à déterminer l'algorithme qui prédit le mieux le genre (connu) d'un individu.

Par analogie à la mesure tf.idf (formule 1), $P'_{AxVx}(C_i)$ représente la fréquence du couple $(A_x, V_x)$ dans la communauté $C_i$ (le « tf » du tf.idf), et $\log \frac{m}{\sum_{j=1}^{m} presence(A_x, V_x, C_j)}$ représente la fréquence inverse du couple $(A_x, V_x)$ dans l'ensemble des profils de toutes les communautés (« idf » de tf.idf).

#### *4.3.1.3.3 Exemple : caractérisation sémantique d'une communauté*

Dans l'illustration de la figure 4.6, la caractérisation de la communauté C1 est, par exemple, réalisée de la manière suivante :

| (Attribut,Valeur) | *Fréquence* $P'_{AxVx}(C1)$ | Fréquence inverse *AxVx (C1)* | $\boldsymbol{P''_{AxVx}(C1)}$ |
|---|---|---|---|
| A | 0.14 | log (4/(1+1+0+1)) | 0.017 |
| B | 0.86 | log (4/(1+0+1+0)) | 0.258 |
| D | 0.71 | log (4/(1+1+0+1)) | 0.088 |

Tableau 4. 5 : Exemple de caractérisation sémantique d'une communauté

Il ressort de cette caractérisation que la communauté C1 est principalement caractérisée par B qui devrait alors avoir un poids conséquent dans la dimension sociale du profil selon notre hypothèse. Bien que le poids de D soit important dans le profil de C1, D n'est pas très caractéristique de C1, car il est également très présent dans d'autres communautés. Dans un cas extrême, si un couple attribut-valeur est présent dans toutes les communautés, sa caractérisation vaudra 0, et par conséquent il n'apparaîtra pas dans la dimension sociale du profil. Cet effet radical d'exclusion induit par la mesure tf.idf peut s'avérer inefficace dans certains cas (communautés très recouvrantes par exemple). Par exemple, si le poids de ce couple attribut-valeur (présent dans toutes les communautés) était important dans une communauté, alors que très faible mais existant dans d'autres communautés, sa caractérisation va toujours valoir 0. En fonction du domaine d'application et de la nature des activités des utilisateurs, la mesure de caractérisation sémantique pourra être adaptée pour gérer ce type d'effet. Des techniques telles que *tf.idf reduction* et *LSA (Latent Semantic Analysis)* peuvent également être exploitées (Isokazi et al., 02), de même que les techniques basées sur les machines à vecteurs de support (SVM) (Billsus et al., 99), ou plus simplement on pourrait considérer que les profils des communautés sont déjà caractérisés, c'est-à-dire considérer que $P''_{AxVx}(C_i) = P'_{AxVx}(C_i)$, ...

#### *4.3.1.3.4 Caractérisation structurelle*

***La caractérisation structurelle*** s'appuie uniquement sur la structure du graphe pour caractériser les communautés. Le fait qu'une communauté soit complètement isolée ou centrale dans le réseau k-égocentrique de l'utilisateur peut également être porteuse d'une information vis-à-vis de l'égo, bien qu'on ne puisse pas à priori indiquer la nature ou la qualité de cette information (elle sera déterminée expérimentalement). Ainsi, étant donné tout couple $(A_x, V_x)$ dans le profil d'une communauté $C_i$, ce couple se caractérise de manière structurelle par le triplet $(A_x, V_x, Struct(C_i))$ suivant la fonction $f_x^{structurelle}$:

$$f_x^{structurelle}: (A_x, V_x) \in C_i \mapsto (A_x, V_x, Struct(C_i))$$

$$Struct(C_i)) \in \{GC_D(C_i), GC_c(C_i), GC_B(C_i), Coh(C_i), \ldots\} \quad \textbf{(26)}$$

*Struct*($C_i$) représente la mesure de structure de la communauté $C_i$. *Struct*($C_i$) peut par exemple être une des mesures de centralité de groupes dans un réseau social tel que présenté dans la section 3.3.6.2 que nous rappelons dans le tableau 4.6. Le choix de la mesure de centralité ou de cohésion la plus pertinente pour *Struct*($C_i$) sera déterminée expérimentalement.

| Centralité de degré (voir formule 14) | Centralité de proximité (voir formule 15) | Centralité d'intermédiarité (voir formule 16) |
|---|---|---|
| $GC_D(C_i) = \frac{N(C_i)}{\lvert V \rvert - \lvert C_i \rvert}$ | $GC_C(C_i) = \frac{\lvert V \rvert - \lvert C_i \rvert}{\sum_{x \in V-C} d_f(x, C_i)}$ | $GC_B(C_i) = \frac{2 * \sum_{u<v} \frac{g_{u,v}(C_i)}{g_{u,v}}}{(\lvert V \rvert - \lvert C_i \rvert) * (\lvert V \rvert - \lvert C_i \rvert - 1)}$ |

Tableau 4.6: Exemples de mesures de structure de communautés

Au-delà des mesures de centralité, d'autres mesures de structure peuvent également être importantes. Nous définissions par exemple la mesure de cohésion (ou densité) d'une communauté $C_i$ (noté Coh($C_i$) comportant n utilisateurs comme étant le ratio entre le nombre de liens existant dans la communauté (noté $N_L(C_i)$) par le maximum possible de ce nombre de liens (n (n-1)/2) :

$$Coh(C_i) = \frac{2 * N_L(C_i)}{n(n-1)} \qquad \textbf{(27)}$$

D'autres mesures de structure de communautés peuvent être définies en exploitant par exemple la pondération des liens entre les individus dans un réseau k-égocentrique étendu. Mais pour l'instant, dans nos travaux nous ne nous intéresserons qu'aux mesures de centralité et à celle de cohésion définies ci-dessus.

Il faut remarquer qu'une mesure de structure est associée à une communauté entière. En conséquence, la même mesure de structure est utilisée pour tous les couples (Attribut, Valeur) de la communauté.

#### *4.3.1.3.5 Exemple : caractérisation structurelle d'une communauté*

Sur l'illustration de la figure 4.6, les valeurs associées (à titre indicatif sur cet exemple) à une mesure de structure (centralité) sont présentées pour chacune des communautés (cadres en jaune). Les communautés C1, C2, C3, et C4 ont respectivement comme valeur de mesure de structure 0.9, 0.5, 0.3 et 0.8.

#### *4.3.1.3.6 Caractérisation sémantico-structurelle*

Après la caractérisation sémantique et structurelle d'une communauté, tout triplet ($A_x$, $V_x$, $P'_{AxVx}(C_i)$) dans le profil d'une communauté $C_i$, est associé à une mesure de caractérisation sémantique $P''_{AxVx}(C_i)$ et une mesure de caractérisation structurelle *Struct*($C_i$) :

$$f_x^{caracterisation} : \quad (A_x, V_x, P'_{AxVx}(C_i)) \in C_i \;\mapsto\; \begin{cases} (A_x, V_x, P''_{AxVx}(C_i)) & \text{par } f_x^{semantique} \\ (A_x, V_x, Struct(C_i)) & \text{par } f_x^{structurelle} \end{cases}$$

Pour obtenir finalement une caractérisation unique (sémantico-structurelle), une fusion des deux caractérisations précédente est réalisée.

La fusion des mesures sémantiques et structurelles se fait dans chaque communauté pour chaque couple ($A_x$, $V_x$) auquel on associe un poids $P'''_{AxVx}(C_i)$ en fonction d'un paramètre $\alpha \in [0, 1]$ suivant la fonction par $f_x^{fusion}$ :

$$f_x^{fusion}: \quad \left.\begin{matrix}(A_x, V_x, P''_{AxVx}(C_i)) \\ (A_x, V_x, Struct(C_i))\end{matrix}\right\} \in C_i \quad \mapsto \quad (A_x, V_x, P'''_{AxVx}(C_i))$$

$$P'''_{AxVx}(C_i) = \alpha Struct(C_i) + (1-\alpha) P''_{AxVx}(C_i) \qquad \textbf{(28)}$$

La valeur optimale du paramètre α sera recherchée expérimentalement pour juger de la pertinence ou non de l'usage de mesures de structures.

#### *4.3.1.3.7 Exemple : caractérisation sémantico-structurelle d'une communauté*

Sur l'illustration de la figure 4.6, la fusion des mesures est représentée par les cercles verts encadrant les mesures de structures et de sémantique pour chaque communauté. Sur cette figure, les calculs sont réalisés avec α=0.1 (cf. Annexe B pour le détail des calculs réalisés sur la figure).

### *4.3.1.4 Etape de dérivation de la dimension sociale*

#### *4.3.1.4.1 Description*

La dérivation de la dimension sociale est la dernière phase du processus. Une fois les éléments des profils de chaque communauté représentés sous la forme $(A_x, V_x, P'''_{AxVx}(C_i))$, les profils des communautés sont combinés pour dériver les éléments de la dimension sociale. Si on considère qu'on dispose de *m* communautés $\{C_1, C_2, \ldots, C_m\}=C$, pour tout couple $(A_x, V_x)$ présent dans au moins une des communautés de C après les fusions des mesures sémantiques et structurelles, la dérivation de la dimension sociale est la fonction $f_x^{derivation}$ :

$$f_x^{derivation}: \quad (A_x, V_x, P'''_{AxVx}) \in C \mapsto \quad (A_x, V_x, P_{AxVx}^{social})$$

$$P_{AxVx}^{social} = Combinaison(\{P'''_{AxVx}(C_i)\}_{C_i \in C}) \qquad \textbf{(29)}$$

D'après notre hypothèse de travail, pour déterminer le poids final $P_{AxVx}^{social}$ de chaque couple attribut-valeur dans la dimension sociale du profil de l'utilisateur, les éléments les plus caractéristiques (après la fusion des mesures de structure et de sémantique) de chaque communauté seront les plus pertinents dans cette dimension du profil. Une manière radicale de le faire consisterait par exemple pour chaque couple attribut-valeur, de considérer uniquement la communauté qui est la plus caractérisée par ce couple attribut-valeur (le poids du couple attribut-valeur dans cette communauté pourra alors être reporté dans la dimension sociale). Cependant, en pratique, il est clair qu'étant donné un couple attribut-valeur, il ne sera pas exclusivement caractéristique d'une seule communauté (c'est à dire un poids $P'''_{AxVx}$ important dans une communauté, et très faible voire nul dans les autres communautés). Ce qui est certain, c'est que les valeurs de $P'''_{AxVx}$ seront généralement distribuées de manière uniforme ou non dans l'ensemble des communautés. Pour aller dans le sens de notre hypothèse, l'idée serait alors de réaliser une combinaison des scores des poids $P'''_{AxVx}$ de sorte que l'importance du poids final $P_{AxVx}^{social}$ soit relatif au fait que $P'''_{AxVx}$ est important dans au moins une communauté. Pour trouver la fonction de combinaison adéquate dans notre cas, nous avons fait une analogie avec les systèmes de fusion de résultats de plusieurs moteurs de recherche dans les systèmes de recherche d'information.

#### *4.3.1.4.2 Analogie pour choix d'une fonction de combinaison optimale*

En recherche d'information, un moteur de recherche est généralement utilisé pour renvoyer des résultats à une requête utilisateur. Si l'on dispose de plusieurs moteurs de recherche, il est également possible de soumettre une requête qui sera évaluée par l'ensemble des moteurs de recherche. Les résultats (de la requête) de chacun des moteurs de recherche de l'ensemble seront alors combinés pour renvoyer un ensemble unique de documents correspondant à la requête de l'utilisateur. (Fox et Shaw, 94) ont montré que de pareils moteurs de recherche combinés améliorent les résultats par rapport à l'usage d'un seul moteur de recherche. Cependant la question « comment combiner efficacement les résultats de plusieurs moteurs de recherche ? » est encore étudiée dans la littérature (Hubert et al., 07). En réalité, le choix de la fonction de combinaison dépend des objectifs attendus par la fusion de l'ensemble de systèmes de recherche (moteurs de recherche). La fonction de combinaison de base *CombSUM* (Fox et Shaw, 94) additionne les scores obtenus par les différents systèmes combinés pour un document donné. Cette fonction ne tient pas compte de la relative distribution des scores d'un document dans l'ensemble des systèmes de recherche, seule la somme de ces scores est exploitée. Par contre, la fonction *CombMNZ* qui est une extension de *CombSUM* se base à la fois sur les scores obtenus par différents systèmes pour un document restitué et sur le nombre de systèmes ayant retrouvé ce document :

$$Score\_CombMNZ_j = \left(\sum_{i=1}^{nbre\_systemes} Score_{ij}\right) * count_j \qquad \textbf{(30)}$$

*j* est un document, *i* est un système de recherche, *nbre_systemes* est le nombre total de systèmes, $Score_{ij}$ est le score calculé pour le document *j* par le système *i*, $count_j$ est le nombre de systèmes fusionnés qui ont retrouvé le document *j*. Par rapport à *combSUM*, *CombMNZ* intègre (en quelque sorte) la distribution des scores, mais de manière agrégée par $count_j$. L'importance de la distribution des scores pour chaque système n'est pas prise en compte et ceci peut pénaliser certains documents jugés très pertinents par un seul système comme on peut le voir sur l'exemple du tableau 4.7 (une requête soumise à deux systèmes de recherche combinés S1 et S2, retourne des documents parmi D1, D2, D3, D4). La liste de résultats du système S2 n'inclut pas le document D4 (N/A). Le document D2 est jugé pertinent par le système S1(0.7) et très peu pertinent dans le système S2 (0.1). Le document D3 n'est réellement jugé pertinent par aucun des systèmes et a des scores quasi moyens (0.5 et 0.4). Cependant, dans la fusion des deux systèmes via *CombMNZ*, le document D2 (1.6) est moins pertinent que le document D3 (2.0), pourtant aucun des systèmes n'a réellement bien classé le document D3 (0.5 et 0.4).

| Documents/Systèmes | **D1** | **D2** | **D3** | **D4** |
|---|---|---|---|---|
| **Système S1** | 1.0 | **0.7** | **0.5** | 0.2 |
| **Système S2** | 0.7 | **0.1** | **0.4** | N/A |
| ***Score_CombMNZ*** | (1+0.7)*2=3.4 | (0.7+0.1)*2=1.6 | (0.5+0.4)*2=2.0 | (0.2)*1=0.2 |
| **Ordre final de fusion** | D1 (3.4) ; **D3 (2.0) ; D2 (1.8)** ; D4 (0.2) | | | |

Tableau 4. 7 : Exemple de fusion de systèmes de recherche de documents via CombMNZ

Pour pallier le problème de ce type de classement induit par *CombMNZ* (Fox et Shaw, 94), (Hubert et al., 07) propose une extension de *CombMNZ*, *Lin_CombMNZ* qui va favoriser la pertinence d'un document dans la fusion des systèmes de recherche si au moins un des systèmes l'a jugé pertinent. En d'autres termes, il s'agit de considérer avec plus d'importance le meilleur score donné à un document dans la combinaison des systèmes.

Ainsi, lorsqu'un système retrouve un document donné dans les premiers documents, *Lin_CombMNZ* réalise une combinaison linéaire des scores en donnant plus d'importance à la contribution de ce système au rang final du document, par rapport aux autres systèmes :

$$ScoreLin\_CombMNZ_j = \sum_{i=1}^{nbre\_systemes / Score_{(i-1)j} < Score_{ij}} (Score_{ij} * i) \qquad \textbf{(31)}$$

*j* est un document, *i* est un système de recherche, $Score_{ij}$ est le score calculé pour le document *j* par le système *i*, *nbre_systemes* est le nombre total de systèmes. Cette fonction agit en deux temps pour trouver le score de fusion de chaque document. Dans un premier temps, les systèmes sont classés par ordre croissant en fonction du score qu'ils ont attribué au document ($Score_{(i-1)j} < Score_{ij}$). Dans un second temps (la combinaison linéaire), le score attribué par chaque système est multiplié par le rang *i* du système dans la classification ordonnée précédente. Ainsi, le score du système qui dispose du score le plus important pour le document *j* est multiplié par 5 (si on dispose de 5 systèmes), le score du second système attribuant le score le plus important pour le document j est multiplié par 4, …, le score du système attribuant le score le moins important pour le document j est multiplié par 1. L'exemple précédent est repris dans le tableau 4.8 suivant (avec la fonction de fusion *Lin_CombMNZ*).

| | **D1** | **D2** | **D3** | **D4** |
|---|---|---|---|---|
| **Système S1** | 1.0 | **0.7** | **0.5** | 0.2 |
| **Système S2** | 0.7 | **0.1** | **0.4** | N/A |
| **Classement des systèmes** | 1)S2, 2) S1 | 1)S2, 2) S1 | 1)S2, 2) S1 | 2)S1 |
| ***ScoreLin_CombMNZ*** | (1*0.7+2*1.0)=2.7 | (1*0.1+2*0.7)=1.5 | (1*0.4+2*0.5)=1.4 | (2*0.2)=0.4 |
| **Ordre final de fusion** | D1 (2.7) ; **D2 (1.5)** ; **D3 (1.4) ;** D4 (0,4) | | | |

Tableau 4. 8 : Exemple de fusion de systèmes de recherche de documents via Lin_CombMNZ

Contrairement à la fonction CombMNZ (Tableau 4.7), le document D2 est plus pertinent que le document D3 dans la fusion des deux systèmes S1 et S2. Ceci vient du fait qu'au moins un système (S1 ici) à lui tout seul a bien classé le document D2.

Les analogies entre ces fonctions de fusion de moteur de recherche en système d'information et notre phase de combinaison des profils (caractérisés et fusionnés) des communautés pour dériver la dimension sociale du profil de l'utilisateur sont les suivantes (figure 4.7) :

- Les documents et leur score sont assimilés respectivement aux couples attribut-valeur et à leur poids (P''') des éléments de profil de communautés.

- Les systèmes (moteurs) de recherche sont assimilés aux communautés du réseau k-égocentrique de l'utilisateur. De la même manière que les moteurs de recherche renvoient des documents, les communautés sont caractérisées par les couples attribut-valeur d'éléments de profil.

- La fonction de combinaison permettant de fusionner les résultats de moteurs de recherche est assimilée à la fonction de combinaison devant permettre de dériver la dimension sociale du profil à partir des profils de communautés caractérisés par les poids P''' (formule 28). Pour aller dans le sens de l'hypothèse de notre travail, la fonction de combinaison doit pouvoir renvoyer un score important pour un couple attribut-valeur dès que le poids de ce couple attribut-valeur est important dans une

seule communauté. La fonction de fusion *Lin_CombMNZ* comme on vient de le voir permet de répondre à ce besoin.

La figure 4.7 présente l'analogie décrite précédemment sur deux systèmes (donc deux communautés), mais le même processus se généralise à n systèmes (n communautés) comme les formules l'indiquent.

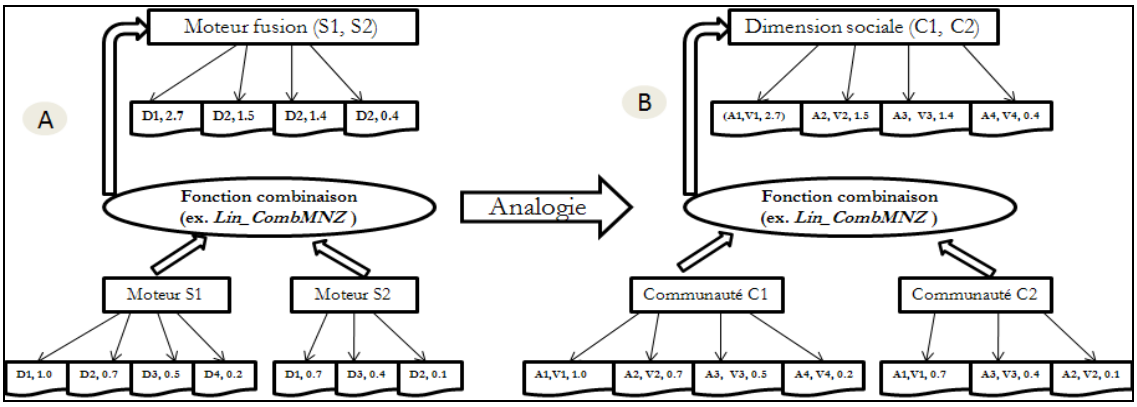



Figure 4. 7: analogie entre la fusion de moteurs de recherche en RI (A) et la dérivation de la dimension sociale à partir de communautés (B)

Donc dans la formule 29, la fonction de combinaison que nous utiliserons correspond à la fonction *Lin_CombMNZ* :

$$P_{AxVx}^{social} = Combinaison(\{P''_{AxVx}(C_i)\}_{C_i \in C}) = ScoreLin\_CombMNZ_{AxVx} = \sum_{j=1}^{nbre\_communautes / P'''_{AxVx}(C_{j-1}) < P'''_{AxVx}(C_j)} (P''_{AxVx}(C_j) * j) \quad \textbf{(32)}$$

$j$ est un indice de communauté, $C_j$ est une communauté, $P'''_{AxVx}(C_j)$ est le poids calculé pour le couple $(A_x, V_x)$ dans la communauté $C_j$, *nbre_communautes* est le nombre total de communautés. De la même manière que cette fonction est utilisée en recherche d'information, elle opère en deux temps. Dans un premier temps, les communautés sont classées par ordre croissant en fonction du poids (P''') qu'ils ont attribué au couple attribut-valeur $(A_x, V_x)$ $(P'''_{AxVx}(C_{j-1}) < P'''_{AxVx}(C_j))$. Dans un second temps (la combinaison linéaire), le poids attribué par chaque communauté est multiplié par le rang $j$ du système dans la classification ordonnée précédente.

Dans la mesure où les mesures tf.idf et les sommes faites dans la dérivation de $P_{AxVx}^{social}$ peuvent être supérieures à 1, le poids $P_{AxVx}^{social}$ peut être normalisé en le divisant par la somme de tous les $P_{AxVx}^{social}$ de tous les couples attribut-valeur de la dimension sociale du profil.

#### *4.3.1.4.3 Exemple : dérivation de la dimension sociale*

Dans l'illustration de la figure 4.6, les poids des couples attribut-valeur représentés par A, B, C, D sont calculés suivant *Lin_CombMNZ* de la même manière que dans l'exemple précédent (voir le détail des calculs en Annexe 2). Sur cette illustration, on peut remarquer que :

- Le couple attribut-valeur représenté par C a le poids le plus important dans la dimension sociale. Ceci va dans le sens de notre hypothèse dans la mesure où une seule communauté (C4) est caractérisée de manière importante par C.

- Le couple attribut-valeur représenté par B a le second poids le plus important. Même si une communauté (C1) est caractérisée de manière importante par B, le poids de B est moins important dans la dimension sociale par rapport à C, car B est également présent dans une autre communauté (C3).
- Le couple attribut-valeur représenté par A et D ont des poids moins important par rapport à B et C, car ils ne caractérisent (de manière importante) aucune des communautés, et de plus, ils sont présents dans plusieurs communautés (3 communautés dans ce cas).
- Dans la dimension sociale du profil, les calculs présentés peuvent aboutir à des poids supérieurs à 1 dans la dimension sociale du profil (dûs au tf.idf et à la combinaison linéaire). Pour normaliser, on peut tout simplement diviser chaque poids obtenu par la somme de tous les poids obtenus (si cette dernière est >=1). Sur l'illustration cette somme vaut 0.67+1.417+2.264+1.066=5.417 (cf. Annexe B). Donc les poids < 1 dans la troisième colonne de la dimension sociale sont les poids non normalisés obtenus (en deuxième colonne) divisés par le total 5.417.

Les étapes du processus de dérivation de la dimension sociale du profil de l'utilisateur sont décrites sous forme algorithmique dans la section suivante.

## 4.3.2 Algorithme de mise en œuvre du processus ($CoSP_k$)

L'algorithme $CoSP_k$ (*Community Social Profils from k-egocentric networks*) permettant de mettre en œuvre toutes les étapes du processus de dérivation de la dimension sociale du profil utilisateur est présenté ci-dessous :

**Algorithme $CoSP_k$**

**Début**

*//initialisation*
$R_v(k)$= réseau k-égocentrique de l'utilisateur v

*//Etape 1 : détection de communautés dans le réseau k-égocentrique, usage de l'algorithme iLCD*
$C$ = détection de communautés avec recouvrement dans le graphe induit par $R_v(k)$

*//Etape 2 : profilage des communautés*
**Pour** chaque communauté $C_i$ appartenant à $C$ **faire**
*AttributsValeurs*= Ensemble des couples attribut-valeur *(a, v)* appartenant à la dimension utilisateur du profil d'au moins un individu de $C_i$
**Pour** tout couple *(a, v)* appartenant à *AttributsValeurs* **faire**
*P'* = poids de *(a, v)* dans la communauté $C_i$ par agrégation des poids de *(a, v)* dans l'ensemble des dimensions utilisateur des profils des individus de $C_i$ *//mesure tf par exemple*
Mettre à jour le couple *(a, v)* dans Ci avec le triplet *(a, v, p')*
**Fin Pour**
**Fin pour**

*//Etape 3 : caractérisation sémantique et structurelle des communautés*

**Pour** chaque communauté $C_i$ appartenant à C **faire**
*struct* = mesure de structure de la communauté $C_i$ *//centralité de degré de communauté par exemple*
*AttributsValeursPoids*= Ensemble des triplets attribut-valeur (a, v, p') dans le profil de $C_i$
**Pour** tout triplet *(a, v, p')* appartenant à *AttributsValeursPoids* **faire**
*p''* = poids de caractérisation sémantique de *(a, v, p)* dans $C_i$ par rapport aux autres communautés dans C *//mesure tf –idf par exemple*
*p'''* = poids de fusion de la mesure sémantique *p'' associée à (a, v)* et de la mesure de structure *struct* associée à la communauté $C_i$ *//p'''=astruct+(1-a)p'' par exemple*
Mettre à jour le triplet *(a, v, p')* dans le profil de Ci avec le triplet *(a, v, p''')*
**Fin Pour**
**Fin pour**

*//Etape 4 : dérivation de la dimension sociale du profil*
**Pour** chaque couple (a, v) appartenant au profil d'au moins une communauté Ci de C
$p^{social}$ = combiner les différents poids *p'''* associés à (a, v) dans les différentes communautés de C *//fonction de combinaison Lin_CombMNZ par exemple*
Rajouter (a, v, $p^{social}$) dans la dimension social $Att_x^{dimS}(v)$ du profil de l'utilisateur V
**Fin pour**

**Fin**

Chacune des étapes du processus est notée en commentaire dans cet algorithme. En fonction de la valeur du paramètre k, on peut avoir plusieurs instances de cet algorithme ($CoSP_1$ pour k=1, $CoSP_2$ pour k=2, etc.).

Cet algorithme peut être paramétré dans le choix des algorithmes ou des mesures comme suit :

- Etape 1, l'algorithme de détection de communautés : nous avons choisi l'algorithme iLCD pour les raisons présentées plus haut.
- Etape 2, la mesure de profilage des communautés (mesure *tf* par exemple).
- Etape 3, la mesure de caractérisation sémantique (*tf.idf* par exemple).
- Etape 3, la mesure de caractérisation structurelle (centralité de degré par exemple).
- Etape 4, la fonction de combinaison des poids des éléments du profil dans les communautés : nous avons choisi *Lin_CombMNZ* pour les raisons présentées plus haut.

Le changement ou les variations de ces paramètres pourront influer sur la qualité des profils obtenus.

Il est à noter que pour les attributs évolutifs, les calculs présentés ici se font au niveau des feuilles de la taxonomie. A la fin du processus, il suffit de remonter les calculs réalisés par sommation jusqu'à la racine de la taxonomie comme expliqué plus haut (cf. figure 4.5).

### 4.3.3 Bilan

Nous avons présenté dans cette section le modèle, le processus (et l'algorithme $CoSP_k$) proposés pour la dérivation de la dimension sociale du profil de l'utilisateur. Ces propositions s'appuient principalement sur notre hypothèse de recherche selon laquelle les communautés autour d'un utilisateur sont plus significatives pour ce dernier comparé aux individus considérés un par un dans son réseau social. A la suite de notre

proposition présentée dans cette section, la question majeure qui se pose est donc celle de la comparaison de notre approche (communautaire) par rapport à celles existantes (individuelles). Ce type de comparaison n'étant pas actuellement évoqué dans la littérature, nous présentons dans la section qui suit des stratégies d'évaluation comparatives de notre approche par rapport aux travaux existant.

# 4.4 Stratégies d'évaluation de la proposition

Nous proposons trois stratégies d'évaluation de l'algorithme présenté dans la section précédente : évaluation automatisée par filtrage social, évaluation automatisée et comparative entre dimensions du profil, évaluation par confrontation à la perception humaine. Pour chacune de ces stratégies nous allons donner ses avantages et ses inconvénients.

## 4.4.1 Evaluation automatisée par filtrage social

L'évaluation automatisée par filtrage social est la stratégie d'évaluation la plus utilisée actuellement dans les travaux de la littérature. Les travaux actuels comme nous les avons présentés dans l'état de l'art, utilisent directement le réseau social de l'utilisateur dans les mécanismes de filtrage social et comparent les résultats obtenus avec les mécanismes n'intégrant pas le réseau social de l'utilisateur. Dans le cas où la modélisation du profil est clairement séparée des mécanismes de filtrage comme proposé dans cette thèse, il s'agira d'évaluer l'impact de la dimension sociale du profil construite par chacun des algorithmes proposés, sur les mécanismes de filtrage social de l'information. Ce type d'évaluation peut être semi-automatique lorsque les utilisateurs sont explicitement impliqués dans l'apport des jugements de pertinence sur les résultats des mécanismes de filtrage proposés.

### *4.4.1.1 Avantages*

Le principal avantage de cette stratégie d'évaluation est qu'elle valide les attentes finales des utilisateurs en besoin informationnel vu qu'elle se situe à la fin du processus de filtrage de l'information.

### *4.4.1.2 Inconvénients*

Cette stratégie de validation peut avoir deux inconvénients majeurs : la confusion sur l'objet de l'évaluation et l'importance des ressources à mobiliser pour la validation.

La confusion sur l'objet de validation peut être due au fait de la séparation entre modélisation de l'utilisateur et mécanismes de filtrage de l'information. Dans la modélisation de l'utilisateur on dispose des algorithmes de construction de la dimension sociale (et utilisateur) du profil. Dans les mécanismes de filtrage on peut également disposer des algorithmes d'exploitation de cette dimension sociale (conjointement avec la dimension utilisateur dans la plupart des cas). On peut penser à de nombreuses approches d'intégration de la dimension utilisateur et de la dimension sociale du profil de l'utilisateur dans un mécanisme de filtrage de l'information. La question qui se pose alors lorsqu'on évalue directement les mécanismes de filtrage est de savoir s'il s'agit d'évaluer les techniques de filtrage ou les techniques de construction du profil social. Si les

deux sont évalués en même temps, ceci rend beaucoup plus lourd le processus d'évaluation qu'il conviendrait alors de séparer suivant deux validations distinctes de la même manière que la modélisation de l'utilisateur soit séparée des mécanismes de filtrage de l'information.

Le second inconvénient sur l'importance des ressources à mobiliser pour ce type d'évaluation est un corollaire du premier inconvénient. Si on évalue à la fois les techniques de filtrage et les techniques de construction du profil, l'évaluation est beaucoup plus complexe. Si l'on dispose de *n* techniques de construction du profil et de *m* techniques d'usage des dimensions du profil dans les mécanismes de filtrage, il faudra évaluer *n*m* possibilités de filtrage d'information. Ce qui est bien sûr très lourd à mettre en œuvre surtout lorsque l'évaluation est semi-automatisée et implique donc des jugements de pertinence des utilisateurs sur chacune des possibilités.

Pour pallier les deux inconvénients de la stratégie précédente, il est important de séparer l'évaluation des profils utilisateurs de l'évaluation des mécanismes de filtrage. Si l'on dispose de *n* techniques de construction du profil et de *m* techniques d'usage des dimensions du profil dans les mécanismes de filtrage, il faudra donc évaluer *n+m* possibilités dans ce cas. Dans cette hypothèse, l'évaluation des *n* techniques de construction du profil permettra de choisir une seule technique qui sera alors la seule à être exploitée suivant les *m* techniques de filtrage de l'information.

## 4.4.2 Evaluation automatisée et comparative entre dimensions du profil

Cette stratégie d'évaluation consiste à évaluer uniquement les techniques de dérivation de la dimension sociale du profil en les comparant à la dimension utilisateur du profil lorsqu'elle existe. Cette stratégie d'évaluation s'appuie sur l'hypothèse selon laquelle la dimension utilisateur du profil contient les éléments pertinents du profil de l'utilisateur. Pour être considéré comme pertinent, plusieurs algorithmes de dérivation de la dimension sociale seront comparés afin de rechercher celui qui produit la dimension sociale la plus proche possible de la dimension utilisateur. Pour réaliser ce type d'évaluation, il faudrait donc considérer des dimensions utilisateur qui se rapprochent le plus du profil réel des utilisateurs : profil renseigné explicitement par l'utilisateur lui-même, ou ne considérer que les utilisateurs ayant un volume d'activités très important permettant de construire une dimension utilisateur consistante de leur profil. Si les profils sont représentés de manière vectorielle comme dans cette thèse, la comparaison pourra consister à évaluer l'angle (cosinus) entre les vecteurs de la dimension sociale et les vecteurs de la dimension utilisateur du profil de l'utilisateur. D'autres mesures telles que la précision, le rappel pourront également être utilisées pour évaluer le degré de prédiction de la dimension utilisateur par chaque algorithme de dérivation de la dimension sociale.

### *4.4.2.1 Avantages*

Cette stratégie a l'avantage de permettre l'évaluation des algorithmes de dérivation de la dimension sociale indépendamment des mécanismes de filtrage d'information. On suppose logiquement que l'algorithme permettant de construire un meilleur profil utilisateur, sera la plus pertinente pour les mécanismes de filtrage de l'information. De plus, le fait que cette évaluation soit automatique peut permettre de faire des évaluations sur des jeux de données très importants.

#### *4.4.2.2 Inconvénients*

Cette stratégie ne comporte pas d'inconvénient majeur, à condition que les dimensions utilisateur utilisées dans l'évaluation soient réellement pertinentes. Ce qui peut tout de même être difficile à prouver dans certains cas.

En considérant les travaux de la littérature sur le filtrage social, nous avons remarqué au début et dans la section 4.2.3.1 de ce chapitre (cf. figure 4.2) que ces travaux ne séparent pas de manière explicite les phases de construction des profils utilisateurs « sociaux » des phases d'usage de ces profils par les mécanismes de filtrage social associés. De ce fait, la comparaison de notre approche par rapport à ces travaux ne peut se faire de manière directe, il est nécessaire de tirer de ces travaux des algorithmes de dérivation de la dimension sociale du profil de l'utilisateur à partir des mécanismes de filtrage social qu'ils proposent. Ces algorithmes basés sur les individus (que nous noterons $ISP_k$, *Individual Social Profile from k-egocentric networks*) pourront alors être exploités pour l'évaluation comparative de notre approche basée sur les communautés. Nous présentons ces algorithmes dans la section 4.4.4 de ce chapitre.

### 4.4.3 Evaluation par confrontation à la perception humaine

Comme la précédente stratégie, cette stratégie permet de n'évaluer que les algorithmes de la phase de modélisation du profil utilisateur, les algorithmes de dérivation de la dimension sociale dans notre cas. Par rapport à la stratégie précédente, il n'est pas nécessaire de disposer d'une dimension utilisateur des profils pour valider les dimensions sociales construites. Les dimensions sociales sont construites et présentées directement aux utilisateurs pour recueillir leurs jugements de pertinence. L'utilisateur devra donc dire dans quelle mesure le profil social construit correspond à ses centres d'intérêts personnels.

#### *4.4.3.1 Avantages*

Cette stratégie a l'avantage d'être potentiellement très fiable dans la mesure où c'est l'utilisateur final qui juge de la pertinence de chacune des méthodes évaluées. De plus, par rapport à la stratégie précédente, on a pas besoin de s'appuyer sur une dimension utilisateur du profil qu'il peut être difficile à acquérir de manière très pertinente dans certains cas.

#### *4.4.3.2 Inconvénients*

Le principal inconvénient de cette méthode peut être la difficulté à travailler sur des jeux de données importants compte tenu de l'implication explicite de chacun des utilisateurs dans le processus de validation. Les interfaces de présentation des profils construits doivent également être bien conçues pour faciliter la perception des profils par les utilisateurs.

Comme dans le cas de la stratégie précédente, on a besoin de définir des algorithmes basés sur les individus ($ISP_k$) que nous utiliserons pour comparer avec notre approche basée sur les communautés, dans la mesure où les travaux existant de la littérature ne les définissent pas de manière explicite. Nous définissons ces algorithmes dans la section qui suit.

## 4.4.4 Algorithmes basés sur les individus du réseau k-égocentrique ($ISP_k$) pour validation par comparaison de dimensions ou par confrontation à la perception humaine de l'approche proposée

Dans la littérature, nous avons constaté que les travaux actuels sur filtrage social s'appuient uniquement sur les individus considérés un par un dans le réseau social de l'utilisateur. En général il s'agit de l'ensemble (ou sous ensemble) des individus directement connectés à l'utilisateur dans le réseau social entier. Cette manière d'exploiter les individus du réseau social peut être étendue en considérant le réseau k-égocentrique de l'utilisateur comme on le verra dans cette section. Nous considérons deux principaux algorithmes basés sur les individus qui peuvent être déduits des travaux de filtrage social dans la littérature : un algorithme basé sur la structure et la sémantique du réseau k-égocentrique $ISP_k^{ss}$ et un algorithme trivial (basé uniquement sur la sémantique) du réseau k-égocentrique ($ISP_k^t$). La principale différence entre ces deux algorithmes est que dans l'algorithme $ISP_k^{ss}$ , les individus du réseau k-égocentrique seront distingués en fonction de la force de leur lien avec l'égo (Carmel et Avesani, 09)(Guy et al., 08)(Guy et al., 09) ou de leur centralité dans le réseau social (Cabanac, 11), alors que dans l'algorithme trivial ($ISP_k^t$), aucune distinction n'est faite chez les individus du réseau k-égocentrique de l'utilisateur (Zeng et al., 09). Nous présentons tour à tour chacun de ces algorithmes dans la suite de cette section. Avant de présenter chaque algorithme, les étapes du processus servant à le définir seront expliquées. Chaque individu du réseau égocentrique de l'utilisateur (égo) sera noté $I_i$ dans cette section.

### *4.4.4.1 Processus et algorithme basé sur la structure et la sémantique du réseau k-égocentrique de l'utilisateur ($ISP_k^{ss}$)*

Chaque individu sera distingué en fonction de la force de son lien avec l'égo (dans un réseau k-égocentrique étendu) ou de sa centralité dans le réseau k-égocentrique de ce dernier (dans un réseau k-égocentrique). De manière assez similaire à l'algorithme basée sur les communautés, nous distinguons trois étapes dans le processus qui permettra de dériver la dimension sociale du profil de l'utilisateur selon cet algorithme : profilage des individus, caractérisation des individus, et dérivation de la dimension sociale. Comme dans le cas de l'algorithme basée sur les communautés, les poids associés à chaque couple attribut-valeur ($A_x$, $V_x$) de la dimension sociale du profil de l'égo seront successivement P', P'' et P''', et $P^{social}$ après chacune de ces étapes.

#### *4.4.4.1.1 Profilage des individus*

Il suffit ici de considérer que pour chaque individu du réseau k-égocentrique, la dimension utilisateur de son profil selon le modèle de profil social proposé dans ce chapitre. Donc le poids $P_{AxVx}(I_i)$ de la dimension utilisateur du profil d'un individu est égal au poids $P'_{AxVx}(I_i)$ à l'issu de cette étape.

$$f_x^{profil} : (A_x, V_x, P_{AxVx}(I_i)) \in Att_x^{dimU}(i) \mapsto (A_x, V_x, P'_{AxVx}(I_i))$$

$$\text{avec } P_{AxVx}(I_i) = P'_{AxVx}(I_i) \qquad \textbf{(33)} \text{ cf. 24}$$

Dans le cas où les dimensions utilisateur des individus du réseau k-égocentrique de l'utilisateur sont déjà calculées, il n'y a donc rien à faire à cette étape, ces dimensions utilisateur sont simplement réutilisées.

#### *4.4.4.1.2 Caractérisation des individus*

La mesure de *caractérisation sémantique* est la même que celle utilisée pour l'algorithme basé sur les communautés. La dimension utilisateur de chaque profil d'individu du réseau égocentrique sera caractérisée par rapport à tous les autres individus du réseau k-égocentrique selon le principe de la mesure *tf.idf* par exemple:

$$f_x^{semantique} : (A_x, V_x, P'_{AxVx}(I_i)) \in Att_x^{dimU}(i) \mapsto (A_x, V_x, P''_{AxVx}(I_i))$$

$$P''_{AxVx}(I_i) = P'_{AxVx}(I_i) * \log \frac{m}{\sum_{j=1}^{m} presence(A_x, V_x, I_j)} \qquad \textbf{(34) cf. 25}$$

$$\text{Avec } presence(A_x, V_x, I_j) = \begin{cases} 1 \text{ si } (A_x, V_x) \in P'_{AxVx}(I_j) \\ 0 \text{ sinon} \end{cases}$$

La mesure de *caractérisation structurelle* (mesure *struct*) de chaque individu sera basée sur des mesures telles que la centralité (degré, proximité, intermédiarité, etc.) (Cabanac, 10) de l'individu dans le réseau k-égocentrique. Nous rappelons quelques unes de ces mesures dans le tableau 4.9 (voir formules 11, 12 et 13). Si l'on considère plutôt un réseau k-égocentrique étendu dans lequel l'égo ainsi que ses liens pondérés avec les individus du réseau sont inclus dans le graphe, la caractérisation structurelle de chaque individu *i* du réseau égocentrique peut tout simplement être le poids du lien entre l'égo *v* et cet individu *i (struct=$w_{v,i}$)* dans le graphe, tel que considéré dans plusieurs travaux de la littérature l'égo (Carmel et al., 09)(Guy et al., 08)(Guy et al., 09).

| Centralité de degré (voir formule 11) | Centralité de proximité (voir formule 12) | Centralité d'intermédiarité (voir formule 13) |
|---|---|---|
| $C_D(i) = \frac{d(i)}{n-1}$ | $C_C(i) = \frac{n-1}{\sum_{j=1}^{n} d(i,j)}$ | $C_B(i) = \sum_{j<k} \frac{P_{jk}(i)}{P_{jk}}$ |

Tableau 4. 9 : Quelques mesures de centralités des individus dans un réseau social

$$f_x^{structurelle} : (A_x, V_x) \in I_i \mapsto (A_x, V_x, Struct(I_i))$$

$$Struct(I_i)) \in \{C_D(i), C_c(i), C_B(i), w_{v,i}, \ldots\} \qquad \textbf{(35) cf. 26}$$

La fusion sémantique et structurelle se fait comme dans l'algorithme basé sur les communautés :

$$f_x^{fusion} : \left.\begin{matrix} (A_x, V_x, P''_{AxVx}(I_i)) \\ (A_x, V_x, Struct(I_i)) \end{matrix}\right\} \in I_i \mapsto (A_x, V_x, P'''_{AxVx}(I_i))$$

Toutefois le calcul de P''' peut se faire suivant deux cas : lorsqu'on est dans un réseau k-égocentrique ou lorsqu'on est dans un réseau k-égocentrique étendu (c'est-à-dire lorsqu'on peut avoir $Struct(I_i) = w_{v,i}$).

**Cas 1 : réseau k-égocentrique**

La fusion des caractéristiques sémantiques et structurelles se fait de la même manière que pour le processus basé sur les communautés en fonction d'un paramètre $\alpha \in [0, 1]$ :

$$P'''_{AxVx}(I_i) = \alpha Struct(I_i) + (1-\alpha) P''_{AxVx}(I_i) \quad \textbf{(36)}$$ **cf. 28**

**Cas 2 : réseau k-égocentrique étendu**

Si on dispose d'un réseau k-égocentrique étendu dans lequel $Struct(I_i) = w_{v,i}$ , P''' peut être déterminé indépendamment de $\alpha$ comme dans la plupart des travaux de la littérature en multipliant chaque élément du profil (sémantique) d'un individu par le poids du lien entre cet individu et l'égo.

$$P'''_{AxVx}(I_i) = Struct(I_i) * P''_{AxVx}(I_i) \quad \textbf{(37)}$$

#### *4.4.4.1.3 Dérivation de la dimension sociale*

Elle se fait également selon deux cas en fonction du type de fusion (P''') effectué à l'étape précédente.

**Cas 1 : réseau k-égocentrique**

Elle se fait par combinaison de la même manière que pour l'algorithme basé sur les communautés :

$$f_x^{derivation} : (A_x, V_x) \in I_i \mapsto (A_x, V_x, P_{AxVx}^{social})$$

$$P_{AxVx}^{social} = Combinaison(\{P'''_{AxVx}(I_i)\}_{I_i \in I}) = ScoreLin\_CombMNZ_{AxVx} = \sum_{j=1}^{nbre\_individus / P'''_{AxVx}(I_{j-1}) < P'''_{AxVx}(I_j)} (P''_{AxVx}(I_j) * j) \quad \textbf{(38)}$$ **cf. 32**

*nbre_individus* est le nombre d'individus dans le réseau k-égocentrique de l'utilisateur. Le raisonnement sur la combinaison des poids est le même, sauf qu'ici, il s'agit d'individus et non de communautés.

**Cas 2 : réseau k-égocentrique étendu**

Dans le cas où on dispose d'un réseau k-égocentrique étendu dans lequel $Struct(I_i) = w_{v,i}$ , $P_{AxVx}^{social}$ peut être déterminé comme dans la littérature en faisant juste la somme des poids caractérisés P''' dans les profils des individus du réseau k-égocentrique étendu.

$$P_{AxVx}^{social} = Combinaison(\{P'''_{AxVx}(I_i)\}_{I_i \in I}) = Somme_{AxVx} = \sum_{j=1}^{nbre\_individus} P'''_{AxVx}(I_j) \quad \textbf{(39)}$$

#### *4.4.4.1.4 Algorithme basé sur les individus $ISP_k^{ss}$*

De ces étapes, l'algorithme $ISP_k^{ss}$ est alors le suivant :

**Algorithme $ISP_k^{ss}$**

**Début**

*//initialisation*
$R_v(k)$= réseau k-égocentrique de l'utilisateur v
$R_v^e(k)$= réseau k-égocentrique étendu de l'utilisateur v *//uniquement si possible*

I=ensemble des individus de $R_v(k)$ ou $R_v^e(k)$

*//Etape 1 : construction du réseau k-égocentrique ou du réseau k-égocentrique étendu*
Considérer chaque individu $I_i \in R_v(k)$ comme une « communauté avec un seul membre »

*//Etape 2 : profilage des communautés*
Considérer la dimension utilisateur du profil de chaque individu $I_i \in R_v(k)$ comme le profil de la « communauté avec un seul membre »

*//Etape 3 : caractérisation sémantique et structurelle des individus*
**Pour** chaque individu $I_i$ appartenant à I **faire**
*struct* = mesure de structure de la communauté $I_i$ *//centralité de degré par exemple*
*AttributsValeursPoids*= Ensemble des triplets attribut-valeur (a, v, p') dans le profil de $I_i$
**Pour** tout triplet *(a, v, p')* appartenant à *AttributsValeursPoids* **faire**
*p"* = poids de caractérisation sémantique de *(a, v, p)* dans $I_i$ par rapport aux autres individus dans I *//mesure tf –idf par exemple*
**Si** on considère juste un réseau égocentrique $R_v(k)$ **alors**
p"'= poids de fusion de la mesure sémantique *p" associée à (a, v)* et de la mesure de la structure *struct* associée à la l'individu $I_i$ en fonction de α *//ex :* αstruct+(1-α)p"
**Sinon si** on considère un réseau égocentrique étendu $R_v^e(k)$ **alors**
p"'= poids de fusion de la mesure sémantique *p" associée à (a, v)* et de la mesure de structure (force du lien entre l'égo et $I_i$) *//ex :* struct*p"
**Fin si**
Mettre à jour le triplet *(a, v, p')* dans le profil de $I_i$ avec le triplet *(a, v, p"')*
**Fin Pour**
**Fin pour**

*//Etape 4 : dérivation de la dimension sociale du profil*
**Pour** chaque couple (a, v) appartenant au profil d'au moins un individu Ii de I **faire**
**Si** on considère juste un réseau égocentrique $R_v(k)$ **alors**
$p^{social}$ = combiner les différents poids *p"'* associés à (a, v) pour tous les individus de I
*//fonction de combinaison Lin_CombMNZ par exemple*
**Sinon si** on considère un réseau égocentrique étendu $R_v^e(k)$ **alors**
$p^{social}$ = faire somme des différents poids *p"'* associés à (a, v) pour tous les individus de I
**Fin si**
Rajouter (a, v, $p^{social}$) dans la dimension sociale $Att_x^{dimS}(v)$ du profil de l'utilisateur V
**Fin pour**

**Fin**

Cet algorithme englobe la plupart des techniques existantes dans la littérature et les étend en utilisant la même approche que celle basée sur les communautés.

### *4.4.4.2Algorithme trivial basé uniquement sur la sémantique du réseau k-égocentrique de l'utilisateur ($ISP_k^t$)*

Dans ce cas, on considère que tous les individus du réseau k-égocentrique de l'utilisateur ont la même importance pour ce dernier. Seule l'étape de dérivation de la dimension sociale est réalisée dans ce cas en agrégeant les dimensions utilisateurs des profils de tous les individus du réseau k-égocentrique.

#### *4.4.4.2.1 Dérivation de la dimension sociale*

Il s'agit de construire la dimension sociale du profil de l'utilisateur comme étant le profil « moyen » des dimensions utilisateurs des individus de son réseau k-égocentrique. Ce profil « moyen » peut par analogie être perçue comme le profil d'une communauté qui contient tous les individus du réseau k-égocentrique de l'utilisateur (étape 2 de l'algorithme $CoSP_k$). Si $R_v(k)$ est le réseau égocentrique de l'utilisateur v, et *C* est l'ensemble des *n* individus de ce réseau :

$$f_x^{derivation} : (A_x, V_x) \in C \mapsto \quad (A_x, V_x, P_{AxVx}{}^{social})$$

$$P^{social}{}_{AxVx} = \frac{\sum_{u=1}^{n} f_x(A_x, V_x, Att_x^{\dim U}(u))}{n} \qquad \textbf{(40)}$$

Si (x=a ou x=s) $f_x(A_x, V_x, Att_x^{\dim U}(u)) = \begin{cases} 0 \ si \ (A_x, V_x) \notin \ Att_x^{\dim U}(u) \\ 1 \ sinon \end{cases}$

Si (x=e) $f_x(A_x, V_x, Att_x^{\dim U}(u)) = \begin{cases} P_{AxVx}(u) \ si \ (A_x, V_x) \in \ Att_x^{\dim U}(u) \\ 0 \ sinon \end{cases}$

#### *4.4.4.2.2 Algorithme trivial basé sur les individus $ISP_k^t$*

L'algorithme trivial qui en est déduit est le suivant :

**Algorithme $ISP_k^t$**

**Début**

*//initialisation*
$R_v(k)$ = réseau k-égocentrique de l'utilisateur v
*C* = ensemble des individus de $R_v(k)$

*//Etape unique et finale : profilage de la communauté C*
**Pour** chaque couple (a, v) appartenant au profil d'au moins un individu *i* de C **faire**
$P_{av}{}^{social}$ = poids de *(a, v)* dans la communauté C par agrégation des poids de *(a, v)* dans la dimension utilisateur des profils de tous les individus i de C
Rajouter (a, v, $P_{av}{}^{social}$) dans la dimension social $Att_x^{dimS}(v)$ du profil de l'utilisateur v
**Fin Pour**

**Fin**

# 4.5 Conclusion

Afin de pallier les problèmes de généricité des profils utilisateurs sociaux nécessaires aux techniques de filtrage social, ainsi que l'inexistence de travaux portant explicitement sur l'optimisation des méthodes d'exploitation du réseau social de l'utilisateur pour dériver les informations de son profil, nous avons présenté notre contribution suivant trois grands axes : la définition d'un modèle générique et social de profil utilisateur, la définition d'un algorithme permettant de dériver la dimension sociale du modèle à partir des communautés du réseau k-égocentriques de l'utilisateur, et enfin la définition de stratégies d'évaluation d'algorithmes s'appuyant sur le modèle proposé.

Le modèle générique de profil social proposé s'appuie sur la notion de réseau k-égocentrique de l'utilisateur, qui représente le réseau constitué des relations entre les individus situés à une distance maximum $k$ de l'égo dans le réseau social entier. Les réseaux égocentriques ou réseaux 1-égocentrique (lorsque $k$=1) ont été principalement étudiés dans la littérature en sociologie et ont servi de support à plusieurs résultats (très exploités dans différents domaines de recherche) tel que le principe de la force de liens faibles du sociologue Mark Granovetter (Granovetter, 73). Notre principale hypothèse de travail s'inspire de ce principe en stipulant que la force des liens entre l'égo et les individus de son réseau k-égocentrique n'est pas forcément l'information la plus importante pour dériver des éléments du profil de l'égo, mais qu'il serait plutôt plus pertinent de s'appuyer sur la signification de communautés autour de ce dernier. Les communautés autour de l'utilisateur sont détectées uniquement à partir des relations entre les individus de son réseau k-égocentrique et ne prennent pas encore l'intensité (force ou faiblesse) des liens. Toutefois les affinités entre les individus de chacune de ces communautés représentent à notre sens des affinités qui se rapportent aussi très probablement à l'égo. Dans le modèle proposé, nous distinguons donc deux dimensions : une *dimension utilisateur* (qui sera construite à partir des activités de l'utilisateur) et une *dimension sociale* (qui sera construite à partir des activités des communautés dans le réseau social de l'utilisateur). Les deux dimensions du profil sont comparables et génériques. Les données contenues dans chaque dimension permettent de caractériser l'utilisateur et peuvent être exploitées de manière flexible selon les besoins de mécanismes de filtrage social de l'information.

A partir du modèle proposé, nous avons présenté le processus et l'algorithme de dérivation de la dimension sociale à partir des communautés du réseau k-égocentrique de l'utilisateur ($CoSP_k$). Cet algorithme se décompose en quatre principales étapes :

- *L'étape de détection de communautés :* nous avons présentés les critères de choix de l'algorithme de détection de communauté qui serait le plus approprié dans notre contexte d'étude. Une étude de plusieurs algorithmes de la littérature en fonction de ces critères nous a permis de choisir l'algorithme qui correspond le mieux à l'ensemble de ces critères.

- *L'étape de profilage des communautés :* nous avons présenté comment agréger les dimensions utilisateur des profils des membres d'une communauté pour construire le profil de communauté pour chacun des types d'attributs considérés dans le modèle (attributs statiques, attributs acquis et attributs évolutifs).

- *L'étape de caractérisation sémantico-structurelle :* nous avons présenté comment distinguer les caractéristiques sémantiques et structurelles de chaque communauté relativement aux autres

communautés. La caractérisation sémantique peut s'appuyer sur des mesures comme tf.idf, et la caractérisation structurelle peut s'appuyer sur des mesures de centralité de groupes (ou classes) dans un réseau social comme le degré, la proximité, l'intermédiairité, et la cohésion de groupes.

- *L'étape de dérivation de la dimension sociale :* nous avons présenté les critères de choix d'une fonction de combinaison optimale permettant de combiner les profils caractérisés des communautés, pour déduire la dimension sociale du profil de l'utilisateur.

Le troisième axe de notre contribution a consisté à présenter les stratégies de validation de notre proposition. Nous avons défini trois principales stratégies :

- *Evaluation automatisée par filtrage social :* qui consiste à intégrer l'approche communautaire proposée dans les mécanismes de filtrage social et à comparer les résultats obtenus avec les approches existantes basées plutôt sur les individus du réseau social de l'utilisateur. Nous avons expliqué que même si cette stratégie permet de valider les besoins finaux de l'utilisateur, elle serait très complexe à mettre en œuvre compte tenu de la multitude de mécanismes de filtrage social qu'on peut associer à un modèle social de profil utilisateur.

- *Evaluation automatisée et comparative entre dimensions du profil :* par rapport à la stratégie précédente, cette stratégie sépare l'évaluation des profils sociaux construits de l'évaluation des mécanismes de filtrage social de l'information associés à ces profils, en s'appuyant sur le fait que la meilleure technique de construction de profils sociaux d'utilisateurs sera également celle qui produira de meilleurs résultats pour les mécanismes de filtrage social de l'information. Il est donc question dans cette stratégie de s'appesantir surtout la recherche de l'algorithme le plus efficace pour la construction de la dimension sociale du profil de l'utilisateur. Ce type d'évaluation peut être automatisée en considérant des utilisateurs très actifs ayant une dimension utilisateur très riche en information, et à prédire les informations de cette dimension utilisateur par différents algorithmes de dérivation de la dimension sociale.

- *Evaluation par confrontation à la perception humaine :* comme la stratégie précédente, il s'agit également de séparer l'évaluation des profils sociaux de l'évaluation des mécanismes de filtrage social associés. Cependant, l'évaluation n'est pas automatisée ici, il s'agit plutôt de construire différentes dimensions sociales du profil des utilisateurs par plusieurs algorithmes, et à demander aux utilisateurs d'évaluer de manière explicite ces dimensions pour déduire le meilleur algorithme.

Pour comparer notre approche avec les approches existantes suivant les deux dernières stratégies présentées, il est cependant nécessaire de dériver des travaux de la littérature les algorithmes de dérivation de la dimension sociale à partir des individus du réseau social (k-égocentrique) de l'utilisateur. Ceci dans la mesure où ces travaux de la littérature ne séparent pas de manière explicite la phase de construction du profil de la phase d'usage de ces profils dans les mécanismes de filtrage social associés. Nous avons dérivé et présenté à partir des travaux existants deux types d'algorithmes basés sur les individus : un algorithme ($ISP_k^{ss}$) qui prend en compte la force des liens entre l'égo et les individus de son réseau (ou les mesures de centralités dans le réseau k-égocentrique), et un algorithme trivial ($ISP_k^{t}$) qui ne fait aucune distinction entre tous les membres du réseau k-égocentrique de l'utilisateur.

Dans le chapitre qui suit, nous allons présenter la double évaluation de notre approche ($CoSP_k$) comparativement à ces deux approches exploitées dans la littérature ($ISP_k^{ss}$ et $ISP_k^{t}$) suivant les stratégies d'évaluation automatisée et comparative entre dimensions du profil, et dans une certaine mesure l'évaluation par confrontation à la perception humaine.

# 5 Chapitre 4 : Expérimentations et évaluations de la contribution

# 5.1 Introduction

Dans le chapitre précédent, nous avons proposé des algorithmes de dérivation de la dimension sociale du modèle du profil utilisateur. Puis nous avons présenté des stratégies de validation possibles où nous avons montré qu'une évaluation des profils construits indépendamment des mécanismes de filtrage social est mieux adaptée. Dans ce cas, si l'on dispose d'un échantillon assez important d'utilisateurs volontaires, l'évaluation peut se faire par confrontation des profils construits à la perception humaine. Il est aussi possible pour des utilisateurs très actifs de réaliser une évaluation en mesurant la similarité entre la dimension utilisateur et la dimension sociale du profil utilisateur. Il est alors possible de réaliser des évaluations sur des échantillons de taille plus importante. Dans un cas idéal, ces deux approches d'évaluation peuvent être complémentaires ; C'est la démarche que nous adoptons, même si nous insistons plus sur l'évaluation par similarité des dimensions du profil, qui est plus souple à mettre en œuvre car automatisable et sans intervention de ressources humaines externes (Tchuente et al., 12) (Tchuente et al., 12 bis) (Tchuente et al., 13) (Tchuente et al., 13 bis).

Le but de l'évaluation est de démontrer de manière empirique la pertinence de notre hypothèse de travail que nous rappelons ici : ***les communautés dans le réseau égocentrique de l'utilisateur sont plus caractéristiques de ce dernier par rapport aux alters considérés individuellement avec ou sans pondération des relations.*** A partir des algorithmes présentés dans le chapitre précédent, démontrer la pertinence de cette hypothèse revient à démontrer la pertinence de l'algorithme basé sur les communautés du réseau k-égocentrique de l'utilisateur ($CoSP_k$) par rapport aux algorithmes basés sur les individus de ce réseau ($ISP_k^{ss}$ et $ISP_k^{t}$). Les évaluations comparatives de ces algorithmes peuvent être analysées suivant plusieurs paramètres tels que la valeur k des réseaux k-égocentriques, le choix de différentes mesures de structure (centralité ou cohésion par exemple), le paramètre $\alpha$ qui relativise l'importance des mesures de structures par rapport aux mesures sémantiques, la densité des réseaux k-égocentriques (que nous définirons par la suite), etc.

Afin de montrer de manière empirique la généricité de notre contribution, nous avons évalué les algorithmes proposés dans deux domaines d'application distincts : les réseaux sociaux numériques (*Facebook*) et les réseaux de co-auteurs (*DBLP*). Dans ce chapitre, nous allons présenter les spécificités de chaque environnement (accès aux données, méthodologie de construction des profils, caractéristiques des échantillons) ainsi que les résultats obtenus dans chacun de ces environnements. De manière globale, ces évaluations ont montré de manière significative la pertinence de notre hypothèse de travail en fonction de paramètres expérimentaux tels que ceux cités dans le paragraphe précédent.

# 5.2 Evaluation sur les réseaux sociaux numériques : cas de Facebook

L'essor du Web 2.0 est essentiellement dû aux réseaux sociaux numériques (Facebook, MySpace, LinkedIn, Viadeo, etc.) qui ont proposé aux internautes un moyen d'avoir de multiples possibilités d'interactions sociales via Internet. Ces environnements nous intéressent dans le cadre de notre évaluation pour deux raisons majeures :

- Ils donnent la possibilité aux internautes de transposer en ligne leur réseau social de la vie réelle. La notion de réseau social ici, va dans le sens des réseaux sociaux réels dans la vie réelle (comme en sociologie). Elle correspond à la définition du réseau social donnée dans ce document et notre hypothèse de travail peut ainsi y être appliquée.
- Ils fournissent aux internautes de nombreuses fonctionnalités (tags, mur, photos, liens, groupes, pages, événements, commentaires, applications de parties tierces, etc.) leur permettant de générer un maximum de traces d'activités et d'interactions. On dispose alors d'importantes quantités de données potentiellement utiles pour la construction des profils utilisateurs.

Le premier enjeu majeur dans notre évaluation concerne l'acquisition de ces données sur le réseau social (structure) et sur les activités des utilisateurs. Comme présenté dans l'état de l'art au chapitre 2 (section 3.3.4), ces environnements fournissent en général des API à des développeurs tiers leur permettant de proposer de nouvelles fonctionnalités à leurs utilisateurs en exploitant les masses des données produites par ces derniers. Une application tierce *Anniversaire* s'appuiera par exemple sur les dates de naissance des utilisateurs pour proposer un calendrier dans lequel les dates d'anniversaire sont indiquées. La question qui se pose dans notre cas particulier consiste à analyser l'accessibilité des données de l'utilisateur et de son réseau k-égocentrique (via les API) qui nous sont utiles pour évaluer les algorithmes proposés. Nous nous intéressons particulièrement au cas de Facebook qui est le réseau social numérique de loin le plus utilisé de nos jours d'une part, et qui dispose également d'une API qui est de loin la plus riche en terme de fonctionnalités et la plus utilisée par les développeurs d'autre part. Ainsi pour présenter notre première évaluation sur Facebook, nous allons dans un premier temps présenter l'accessibilité aux données utilisateurs qui nous sont utiles via l'API Facebook. Nous allons ensuite présenter la méthodologie de construction des deux dimensions du profil des utilisateurs à partir des données accessibles. L'échantillon de données utilisées pour l'évaluation ainsi que les résultats obtenus et leurs interprétations seront ensuite présentés tour à tour.

## 5.2.1 Accès aux données utilisateurs via l'API Facebook

### *5.2.1.1 Généralités sur le développement d'applications Facebook*

Facebook a été le premier site de réseau social numérique à proposer une API pour le développement de nouvelles fonctionnalités par des tiers. D'un point de vue technique (architectural), une application Web développée via l'API Facebook par un tiers est hébergée sur un serveur d'application Web (PHP par exemple) comme toute application Web traditionnelle, même si les interfaces de ces applications sont généralement affichées sur Facebook via le compte de l'utilisateur. En réalité, l'application hébergée sur un serveur d'application tiers interagit avec Facebook via l'API Facebook. Pour utiliser une application Facebook développé par un tiers, chaque utilisateur Facebook doit explicitement valider l'installation de cette application sur son profil (Tchuente et al., 09).

Par rapport aux applications Web traditionnelles, les applications Web développées par des tiers sur Facebook (ou sur les réseaux sociaux numériques en général) ont deux valeurs ajoutées principales. Premièrement, elles peuvent exploiter la structure du graphe social des utilisateurs pour diffuser largement et rapidement des informations. Ensuite, elles peuvent accéder à certaines données et traces d'activités publiées par les utilisateurs pour personnaliser par exemple les contenus qu'elles proposent aux utilisateurs. En ce qui nous

concerne, nous nous sommes intéressés à la qualité et la quantité des données accessibles sur les profils (comptes) des utilisateurs Facebook pouvant nous permettre d'évaluer les algorithmes présentés dans le chapitre précédent. Pour un utilisateur (*égo*) donné dans Facebook, il s'agit pour nous de rechercher trois catégories de données :

- ***Les données relatives à l'utilisateur :*** ce sont les données renseignées explicitement par l'utilisateur (sexe, date de naissance, cursus académiques, employeurs, etc.), et les données issues des activités de l'utilisateur (tags, commentaires, statuts, *likes*, liens publiés, groupes rejoints par l'utilisateur, événements dont l'utilisateur est un participant, les pages dont l'utilisateur est un fan, etc.). Ces données seront utiles pour construire la dimension utilisateur du profil de l'utilisateur.
- ***Les données de structure du réseau k-égocentrique de l'utilisateur* :** pour k=1, il s'agit des contacts directs de l'utilisateur ainsi que les relations entre ces contacts. Pour k>1, il s'agit des individus situés à distance <=k de l'utilisateur ainsi que des relations entre ces individus. Ces données seront utiles pour construire le graphe représentant le réseau k-égocentrique de l'utilisateur et y extraire les communautés par l'algorithme de détection de communautés.
- ***Les données relatives aux individus du réseau k-égocentrique de l'utilisateur :*** il s'agit des données renseignées explicitement par ces individus ou des données issues de leurs activités (mêmes données que les données relatives à l'égo pour lequel le réseau k-égocentrique est construit). Ces données seront utiles pour construire et caractériser les profils des communautés qui serviront à dériver la dimension sociale du profil de l'utilisateur.

Entre la première version de l'API Facebook (en 2006) et la version actuelle, plusieurs changements ont eu lieu sur les données accessibles par les applications tierces et sur la manière d'accéder à ces données. Dans un premier temps, lors de la validation de l'installation d'une application tierce sur son profil, l'utilisateur n'avait pas le moyen de paramétrer les données accessibles par cette application à partir de son profil (Tchuente et al., 09)(Tchuente et al., 11). La politique d'accès aux données définie par Facebook s'appliquait de manière unilatérale une fois que l'utilisateur avait installé l'application. Aujourd'hui, cette politique a évoluée et plusieurs « permissions » (via le protocole OAuth) ont été mises en place pour permettre à l'utilisateur un contrôle de l'accès aux données de son profil par des applications tierces. Certaines données restent tout de même accessibles par défaut (dès que l'utilisateur installe l'application sur son profil), et d'autres nécessitent l'accord explicite de l'utilisateur comme on peut le voir (en partie) sur le tableau 5.1.

| | **Données accédées par défaut** | | **Données accédées avec accord explicite de l'utilisateur** |
|---|---|---|---|
| **Utilisateur** | ***Réseau égocentrique :*** | *Accessible* | |
| | ***Attributs statiques :*** | *Accessible* (nom, sexe par exemple). | *Exemple :* intérêts explicitement renseignés par l'utilisateur, hobbies, émissions télés. |
| | ***Attributs Acquis :*** | *Pas accessible* | *Exemple :* historique des emplois, cursus académique, ville de résidence |
| | ***Attributs évolutifs:*** | *Pas Accessible* | *Exemple :* éléments publiés (statuts, liens, notes, photos, vidéos, groupes, pages, événements, etc.) |
| **Amis de l'utilisateur** | ***Réseau égocentrique :*** | *Pas Accessible* | *Pas Accessible* |
| | ***Attributs statiques :*** | *Accessible* (nom, sexe par exemple). | *Exemple :* intérêts explicitement renseignés par l'utilisateur, hobbies, émissions télés. |
| | ***Attributs acquis :*** | *Pas accessible* | *Exemple :* historique des emplois, cursus académique, *ville de résidence* |
| | ***Attributs évolutifs :*** | *Pas accessible* | *Exemple :* éléments publiés (statuts, liens, notes, photos, vidéos, groupes, pages, événements, etc.) |

Tableau 5. 1 : Accessibilité aux données du modèle à partir du profil d'un utilisateur d'une application tierce sur Facebook

Ce tableau présente dans le cas de Facebook, les données pouvant être exploitées par le modèle « social » de profil proposé dans le chapitre précédent d'une part, et leur accessibilité actuelle via l'API Facebook (par une application tierce).

Les deux principales conclusions à tirer de ce tableau sont les suivantes :

- *Facebook* dispose de données que l'on peut rendre compatibles avec celles du modèle de profil « social » proposé : notions de *réseau égocentrique* (via de requêtes FQL[48] par exemple (Tchuente et al., 11)), *attributs statiques* renseignés explicitement par les utilisateurs dans leur profil Facebook (nom, sexe par exemple), *attributs acquis* renseignés explicitement par les utilisateurs dans leur profil Facebook (historique des emplois, cursus académiques, villes de résidence), attributs évolutifs (qui peuvent être déduits des données publiées par des utilisateurs sur Facebook : tags, commentaires, photos, vidéos, liens, adhésion ou participation à des groupes, pages, événements, etc.). Il est à noter tout de même que les données plus sensibles telles que les mails et les messages publiés via les tchats ne sont pas accessibles via l'API Facebook.

- Avec un accord explicite de l'utilisateur il est possible d'accéder aux attributs statiques, acquis, et évolutifs des membres du réseau égocentrique de l'utilisateur. Donc il est possible d'évaluer le modèle proposé pour chaque utilisateur donnant les droits supplémentaires requis à l'application. Toutefois, le réseau égocentrique des amis de l'utilisateur n'étant pas accessibles, il n'est pas possible de réaliser une validation en cascade sur plusieurs égos à partir d'un seul utilisateur d'une application tierce.

Pour réaliser une première évaluation du modèle dans le cas de Facebook, nous avons donc développé une application tierce (via l'API Facebook) nommée « *egoaccess*[49] » dont le but est de permettre à des utilisateurs volontaires de nous donner les droits nécessaires d'accès à leur profil afin d'évaluer notre modèle. Les caractéristiques de l'échantillon de données ainsi que les analyses réalisées seront présentées dans les sections 5.2.3 et 5.2.4. Nous présentons tout d'abord la méthodologie de construction des éléments du profil par analyse de données textuelles publiées par les utilisateurs.

## 5.2.2 Méthodologie de construction des dimensions sociales et utilisateur, et processus de validation

Afin de valider notre proposition, nous utilisons les algorithmes de dérivation de la dimension sociale présentés dans le chapitre 3 ($CoSP_k$, $ISP_k^t$, $ISP_k^{ss}$) pour construire les dimensions sociales du profil d'un utilisateur à partir des activités de son réseau égocentrique dans Facebook. La dimension utilisateur du profil de l'utilisateur est construite à partir des activités de l'utilisateur dans Facebook. La validation est effectuée en recherchant l'algorithme de dérivation de la dimension sociale qui construit une dimension sociale se rapprochant le plus possible de la dimension utilisateur.

[48] Langage Facebook (Facebook Query Language) pour l'interrogation de sa base de données (c'est simplement une couche implémentée par Facebook au dessus du SQL standard).

[49] Disponible à l'adresse https://apps.facebook.com/egoaccess/

### *5.2.2.1 Construction des dimensions sociales du profil d'un utilisateur Facebook*

Nous présentons ici la méthodologie de construction des centres d'intérêts (attributs évolutifs) de la dimension sociale du profil des utilisateurs (le même principe est utilisé pour le traitement des attributs statiques et acquis). Le but ici est de présenter comment les données issues des activités des utilisateurs dans Facebook sont exploitées par les algorithmes de dérivation de la dimension sociale du profil de l'utilisateur ($CoSP_k$, $ISP_k^t$, $ISP_k^{ss}$).

Cette méthodologie se décompose en quatre étapes présentées sur la figure 5.1.

- L'étape 1 vise à préparer des données textuelles extraites dans Facebook pour construire les couples attribut-valeur (centres d'intérêts) qui seront exploités par chacun des algorithmes de dérivation de la dimension sociale du profil de l'utilisateur.
- L'étape 2 correspond à l'usage des algorithmes de dérivation de la dimension sociale en exploitant les scores sémantiques des centres d'intérêts et les scores de structure des individus ou communautés.
- L'étape 3 consiste à mettre à jour la dimension sociale du profil de l'utilisateur sur les feuilles de la taxonomie représentant cette dimension.
- L'étape 4 consiste à mettre à jour la taxonomie entière, des feuilles vers la racine.

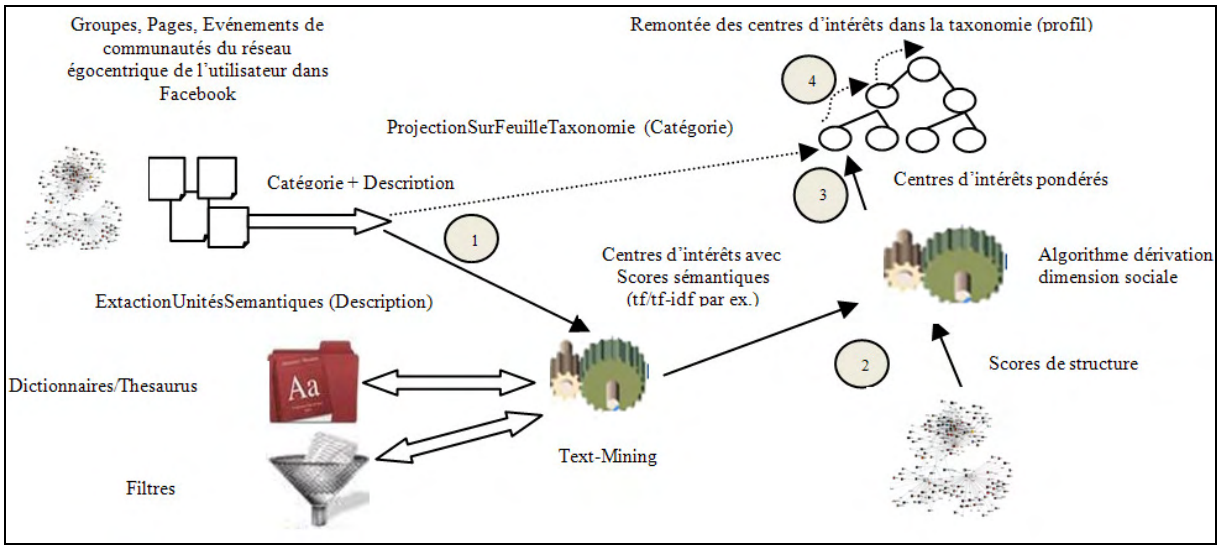



Figure 5. 1 : Méthodologie de construction des centres d'intérêts de la dimension sociale du profil d'un utilisateur Facebook

Ces étapes sont décrites avec plus de détails dans les sections suivantes.

#### *5.2.2.1.1 Etape 1 : projection sur la taxonomie et extraction des mots*

Comme indiqué dans le tableau 5.1, plusieurs éléments publiés (commentaires, liens, notes, photos, vidéos, tags, etc.) ou adhésions (pages, groupes, événements, etc.) de l'utilisateur dans Facebook peuvent être exploités pour construire les centres d'intérêts des attributs évolutifs de son profil (Tchuente et al., 10)(Tchuente et al.,11). Pour la validation du modèle « social » de profil proposé, nous nous sommes

intéressés uniquement à trois activités des utilisateurs : l'adhésion à des groupes, la connexion à des fans pages, et la participation à des événements. Ceci pour deux principales raisons :

- Les pages, groupes et événements créés dans Facebook sont affectés à des catégories préexistantes. La figure 5.2 (page officielle de la campagne électorale 2012 du politicien « Barack Obama ») est un exemple de page Facebook classée dans la catégorie « *politician* ». Nous exploitons la liste de toutes les catégories de pages, groupes, et événements pour créer une taxonomie de concepts permettant de représenter un profil à plusieurs niveaux de granularité tel que décrit dans le modèle que nous proposons. De plus, il est plus facile d'interpréter les connections d'un utilisateur vers des pages, groupes, et événements en termes de centres d'intérêts de ce dernier (si un utilisateur est connecté à un très grand nombre de pages de la catégorie *politique* par exemple, il est logique et simple d'interpréter que cet utilisateur est intéressé par la politique).

- Les pages, groupes et événements possèdent des titres ou descriptions textuelles qui contiennent des termes significatifs qui peuvent être interprétés comme des centres d'intérêts. Sur la page de la figure 5.2 par exemple, est annotée la description textuelle qui sera exploitée pour extraire les termes à intégrer comme centre d'intérêt dans la catégorie *politique* du profil d'un utilisateur fan de cette page.

Dans cette première étape deux actions principales sont réalisées (figure 5.1) :

- La catégorie de la page (groupe ou événement) est projetée sur la feuille correspondante dans la taxonomie représentant le profil de l'utilisateur. Par la suite (comme on le verra à l'étape 3), la feuille correspondante dans la taxonomie sera mise à jour avec les centres d'intérêts calculés à partir de la description textuelle de la page. Chaque catégorie correspond à un attribut du modèle de profil proposé dans ce document. Les valeurs correspondantes à cet attribut seront extraites à partir du texte descriptif de la page (groupe ou événement).


Figure 5. 2 : Exemple d'informations utilisées dans une page Facebook pour construire le profil de l'utilisateur

- La description textuelle de la page (groupe ou événement) est analysée pour extraire les mots dans le texte en utilisant des délimiteurs de mots (virgules, espaces, point virgule, etc.) (cf. Annexe A.2). Le nombre d'occurrences de chaque mot est calculé. Si on considère par exemple la phrase suivante dans la description de la figure 5.2 « *This page is run by Obama for America, President Obama's 2012*

*campaign.* », si les délimiteurs contiennent l'espace, la virgule et l'apostrophe, tous les mots extraits ont une seule occurrence, hormis le mot « Obama » qui a deux occurrences. Seuls les mots significatifs extraits via les délimiteurs de texte seront considérés comme les valeurs de l'attribut correspondant à la catégorie de la page (groupe ou événements). Pour ce faire, nous utilisons le principe de traitement de données textuelles de la plateforme dédiée *Tétralogie* (Dousset, 06). Des dictionnaires et des filtres sont exploités pour enrichir les analyses (cf. Annexe A.2). Les dictionnaires exploités dans notre cas, sont les dictionnaires de synonymes qui regroupent tous les synonymes d'un mot, qui seront alors considérés comme des occurrences de ce mot dans le texte analysé. Ces dictionnaires peuvent être construits par des experts d'un domaine (Tchuente et al., 10) ou automatiquement à partir de ressources sémantiques externes comme les ontologies (Haddadi et al., 09). Les filtres servent à exclure ou conserver certains mots dans les analyses. Seuls certains mots peuvent être retenus dans les analyses (filtres positifs) en particulier dans le cas d'une étude ciblée dans un domaine bien connu. D'autres mots vides de sens tels que les articles (filtres négatifs) sont également exclus des analyses (cf. Annexe A.2). Des listes de mots vides de la langue française et de la langue anglaise qui sont exploités dans la plateforme Tétralogie, sont réutilisées dans nos analyses. Les mots obtenus à la suite de l'application de filtres et de dictionnaires sont considérés comme les centres d'intérêts de l'utilisateur (valeurs associées à l'attribut ou catégorie analysée).

Afin de réduire les temps d'exécution, chaque catégorie de page (groupe ou événement) n'est traitée qu'une seule fois. Pour une catégorie donnée, tous les textes descriptifs de pages, groupes et événements correspondant à cette catégorie sont concaténés pour être analysés comme un seul texte.

#### *5.2.2.1.2 Etape 2 : usage des algorithmes de dérivation de la dimension sociale du profil de l'utilisateur*

Les couples attributs-valeurs (mots significatifs associés aux catégories) sont utilisés en entrée de chacun des algorithmes de dérivation de la dimension sociale du profil présenté dans la contribution. Ils sont pondérés (sémantiquement) suivant leur mesure *tf* et *tf.idf*. Pour obtenir la pondération sémantico-structurelle suivant ces algorithmes, les mesures de structure (centralité de degré par exemple) issues du réseau égocentrique de l'utilisateur sont combinées avec le poids sémantique de chaque centre d'intérêt. Chaque algorithme ($CoSP_k$, $ISP_k^t$, $ISP_k^{ss}$) est exploité pour construire une dimension sociale distincte du profil de l'utilisateur.

#### *5.2.2.1.3 Etape 3 : mise à jour des centres d'intérêts sur une feuille de la taxonomie*

La catégorie utilisée à l'étape 1 pour se positionner au niveau d'une feuille de la taxonomie est mise à jour avec la liste des centres d'intérêts pondérés de la dimension sociale à l'issue de l'étape 2.

#### *5.2.2.1.4 Etape 4 : propagation des centres d'intérêts des feuilles vers la racine de la taxonomie*

Une fois toutes les catégories analysées, les poids des centres d'intérêts sont remontés des feuilles vers la racine de la taxonomie de manière récursive. Les centres d'intérêts d'un nœud parent sont calculés en réalisant la moyenne des centres d'intérêts de ses nœuds fils dans la taxonomie (cf. figure 4.5).

### *5.2.2.2 Construction de la dimension utilisateur du profil d'un utilisateur Facebook*

Pour construire la dimension utilisateur du profil de l'utilisateur dans Facebook, nous utilisons les activités de l'utilisateur lui-même dans Facebook. La méthodologie de construction est présentée sur la figure 5.3. La méthodologie est également réalisée en quatre étapes comme pour la construction de la dimension sociale avec les particularités suivantes :

- Etape 1 : ce sont plutôt les connexions de l'utilisateur aux pages, groupes et événements qui sont exploités. La projection sur la taxonomie et la construction des couples attribut-valeur (centres d'intérêts) se font de la même manière que précédemment.

- Etape 2 : les centres d'intérêts de l'utilisateur sont pondérés uniquement par le score sémantique *tf*. Bien évidemment, les scores de structures et l'algorithme de dérivation de la dimension sociale n'ont pas lieu d'être ici.

- Etape 3 : mise à jour des feuilles de la taxonomie (idem que précédemment).

- Etape 4 : mise à jour de la taxonomie entière (idem que précédemment).

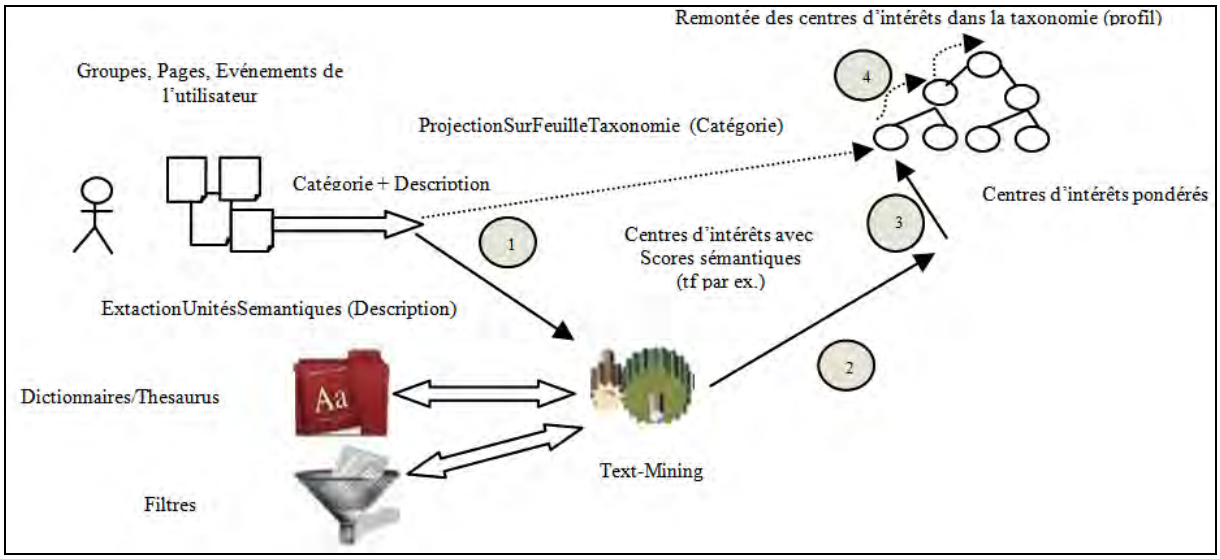



Figure 5. 3 : Méthodologie de construction des centres d'intérêts de la dimension utilisateur du profil de l'utilisateur dans Facebook

### *5.2.2.3 Processus de validation des dimensions sociales construites*

La principale stratégie de validation adoptée dans cette expérimentation sur Facebook est celle automatisée et comparative entre dimensions du profil (cf. section 4.4.2). Cette stratégie de validation a été choisie car elle est la plus simple à mettre en œuvre en termes de temps et de ressources humaines (pas besoin de retours de pertinence explicite de chacun des participants). La validation est effectuée en recherchant l'algorithme de dérivation de la dimension sociale qui construit une dimension sociale se rapprochant le plus possible de la dimension utilisateur (figure 5.4).

Ce processus se décompose en trois principales étapes :

- Etape 1 : construction de la dimension utilisateur du profil de l'utilisateur dans Facebook (cf. figure 5.1, section 5.2.2.1.1). Pour réaliser une évaluation la plus pertinente possible, seuls les utilisateurs très actifs (pour lesquels on aura des centres d'intérêts « réellement pertinents » dans leur dimension utilisateur) sont étudiés. Nous avons retenus dans cette expérimentation des utilisateurs connectés à au moins 250 pages, groupes ou événements.

- Etape 2 : construction de la dimension sociale du profil de l'utilisateur dans Facebook (cf. figure 5.1, section 5.2.2.1.1) à partir de son réseau égocentrique. Chacun des trois algorithmes présentés dans le chapitre 3 ($CoSP_k$, $ISP_k^t$, $ISP_k^{ss}$) est utilisé.

- Etape 3 : validation par comparaison de la dimension utilisateur avec chacune des dimensions sociales construites. Il s'agit ici de comparer des taxonomies ayant la même structure. Nous réalisons cette comparaison catégorie (attribut) par catégorie (attribut) (à des niveaux de granularité plus ou moins élevés suivant la hiérarchie définie par la taxonomie). Chaque catégorie (attribut) de la taxonomie étant présentée sous la forme d'un vecteur de centres d'intérêts pondérés (couples attribut-valeur pondérés dans le modèle proposé), nous utilisons le cosinus de similarité comme mesure de comparaison (cf. formule 2). L'algorithme qui produira une dimension sociale dont le cosinus de similarité avec la dimension utilisateur (pour les attributs statiques, acquis et évolutifs) est le plus élevé (l'angle entre le vecteur représentant dimension sociale et le vecteur représentant la dimension utilisateur est le plus petit) sera donc considéré comme l'algorithme le plus pertinent.

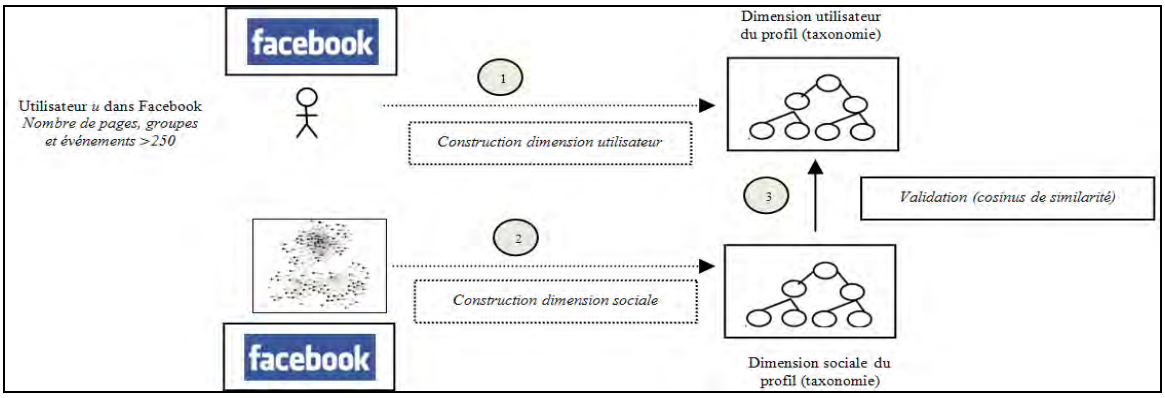



Figure 5. 4 : Processus de validation dans Facebook

Après avoir présenté les méthodologies de construction et de validation des profils, nous allons maintenant présenter les caractéristiques de l'échantillon d'utilisateurs analysés ainsi que les résultats comparatifs obtenus par les trois algorithmes de dérivation de la dimension sociale du profil de l'utilisateur.

## 5.2.1 Caractéristiques de l'échantillon de données étudié

Au total **64 utilisateurs** se sont portés volontaires et ont installé notre application en accordant les autorisations supplémentaires demandées pour l'extraction des activités des membres de leur réseau égocentrique. Cependant, seuls **15 utilisateurs** (**15 réseaux égocentriques**) ont été jugés suffisamment actifs (car connectés à au moins 250 pages, groupes ou événements) pour pouvoir mettre en œuvre la stratégie de validation décrite précédemment. Ces utilisateurs possèdent en **moyenne chacun 235 amis** dans leur réseau égocentrique, pour **un total de 3525 profils utilisateurs Facebook** distincts accédés et analysés dans cette expérimentation.

## 5.2.2 Résultats

### *5.2.2.1 Comparaison entre dimensions du profil par le cosinus de similarité*

Dans les mécanismes d'adaptation de l'information à l'utilisateur, ce sont le plus souvent les centres d'intérêts les plus pertinents pour un utilisateur qui sont considérés dans son profil. Dans cette expérimentation nous avons classifié en trois niveaux les centres d'intérêts contenus dans la dimension utilisateur du profil en fonction de leur poids : les 10 premiers centres d'intérêts (top 10, que nous considérons comme les plus importants), les 20 premiers centres d'intérêts (top 20), et les 50 premiers centres d'intérêts (top 50). Nous avons comparé les algorithmes de dérivation de la dimension sociale par rapport à chacun de ces niveaux de pertinence (top 10, top 20 et top 50) de la dimension sociale du profil de l'utilisateur afin de rechercher l'algorithme qui prédit le mieux les centres d'intérêts les plus pertinents pour l'utilisateur. La figure 5.5 présente les résultats obtenus pour chaque type d'attribut considéré dans le modèle proposé (attributs statiques, attributs acquis et attributs évolutifs) (Tchuente et al., 12 bis)(Tchuente et al., 13)

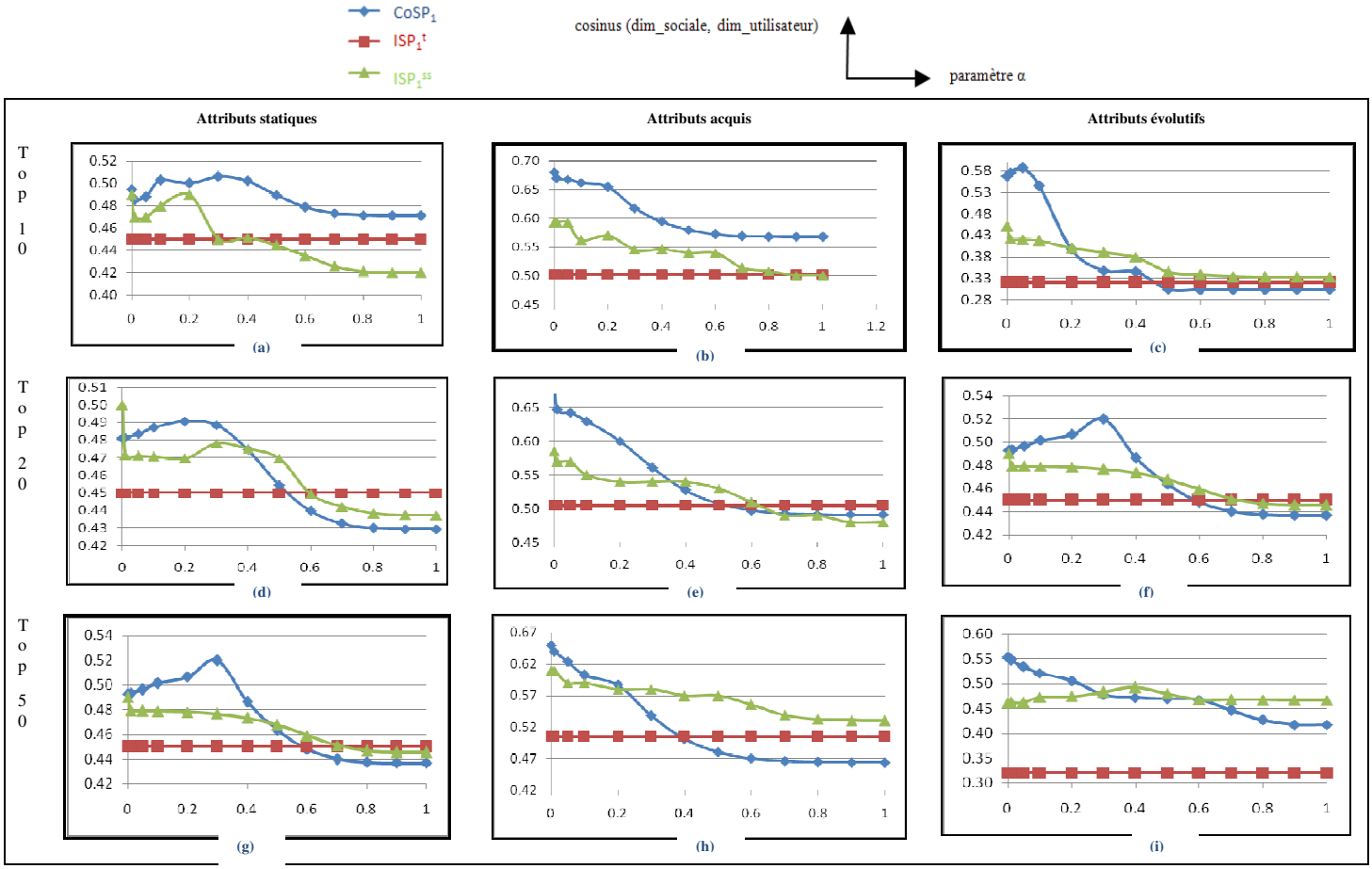



Figure 5. 5 : comparaison cosinus de similarité entre les dimensions utilisateurs et les dimensions sociales construites par les algorithmes $CoSP_1$ (Bleu), IBSP1($ISP_1^{ss}$), $ISP_1^t$ (rouge), en fonction du paramètre α (Tchuente et al., 13)

Sur l'axe x nous avons les valeurs du paramètre α (la mesure de structure utilisée ici est la centralité de degré), sur l'axe y les valeurs des cosinus (cf. formule 2) entre la dimension utilisateur et chacune des dimensions sociales.

Rappelons que dans cette expérimentation, k=1, donc les algorithmes évalués ici sont notés (en remplaçant k par 1 dans les noms des algorithmes) : $CoSP_1$ (algorithme basé sur les communautés), $ISP_1^{ss}$ (algorithme basé sur les individus qui exploite les mesures de structure) et $ISP_1^{t}$ (algorithme trivial basé sur les individus).

Il est à noter que la mesure de structure utilisée dans cette expérimentation est le degré de centralité des communautés (pour l'algorithme $CoSP_1$) et le degré de centralité des utilisateurs (pour l'algorithme $ISP_1^{ss}$) 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 basée sur les communautés, courbe bleue ($CoSP_1$) 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 ($CoSP_1$) fournit de meilleurs résultats que les autres algorithmes. Pour le top 10, cet algorithme est largement meilleur par rapport au top 20 et top 50 bien qu'il reste tout de même meilleur dans ces cas. Ceci implique donc que l'algorithme $CoSP_1$ 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 ($CoSP_1$ et $ISP_1^{ss}$) 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 structure 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 (valeurs de α entre [0, 1] notamment).

Il ressort de cette expérimentation que l'algorithme proposé basé sur les communautés ($CoSP_1$) est plus performante que les algorithmes basés sur les individus, surtout pour la prédiction des centres d'intérêts les plus pertinents de l'utilisateur.

#### *5.2.2.2 Confrontation à la perception humaine*

Les attributs évolutifs des profils étant représentés suivant une taxonomie, il est possible d'exploiter les profils construits suivant des niveaux de granularité plus ou moins élevés. La figure 5.6 présente par exemple, sous forme de nuages de mots, les profils (dimension utilisateur et dimension sociale construite par l'algorithme $CoSP_1$) du domaine (catégorie) « sport » d'un utilisateur ayant participé à l'expérimentation. A droite (figure 5.4b) sont présentés les éléments de la dimension sociale du profil construit par l'algorithme basé sur les communautés $CoSP_1$. Ce profil a été présenté à l'utilisateur en question (sans lui 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 5.6a. On peut constater que les centres d'intérêts les plus importants de la dimension utilisateur (figure 5.6a) se retrouvent également


Figure 5. 6 a- Dimension utilisateur catégorie « sport » du profil d'un utilisateur, b- Dimension sociale (par l'algorithme basée sur les communautés) du profil « sport » du même utilisateur.

dans la dimension sociale (figure 5.6b), cas par exemple de BasketBall, France, NBA, Limoges, Judo. Ceci montre par exemple que, même si l'on ne disposait pas de la dimension utilisateur du profil « sport » de cet utilisateur, elle peut être déduite à partir de son réseau social (algorithme basé sur les communautés). Cet exemple simple de validation par confrontation à la perception humaine montre également la pertinence potentielle de l'approche sociale de construction du profil utilisateur.

### *5.2.2.3Avantages*

Cette première expérimentation possède deux avantages principaux :

- elle montre la pertinence potentielle de l'algorithme de dérivation de la dimension sociale à base de communautés par rapport aux algorithmes basés sur les individus,
- elle est réalisée sur des données qui reflètent les données des réseaux sociaux tels qu'ils existent dans la vie réelle, et donc, correspond au type de données respectant notre hypothèse de travail. Les liens créés par des utilisateurs dans Facebook ont généralement le même sens que les liens entretenus entre ces derniers dans la vie réelle.

### *5.2.2.4Limites*

La principale limite de cette expérimentation est la quantité de données (ou réseaux k-égocentriques) étudiés. Même si les résultats montrent la pertinence de notre approche, le nombre assez limité de réseaux égocentriques analysés ne permet pas d'affirmer avec certitude cette pertinence. Cette limite étant naturellement due à la difficulté d'accès aux données utilisateur dans Facebook. La quantité limitée de réseaux analysés ne permet pas également de comparer plus en détails les trois algorithmes présentés. Par exemple, il peut être important d'analyser l'impact du nombre d'utilisateurs dans les réseaux k-égocentriques ou de la densité des réseaux égocentriques sur les résultats obtenus. Logiquement, on peut penser que plus un réseau k-égocentrique comportera d'utilisateurs et sera dense, plus l'approche basée sur les communautés fournira de meilleurs résultats. De plus il n'est pas possible avec l'API Facebook, d'accéder aux données des activités des utilisateurs situés à distance k (k>1), et de ce fait, seule l'expérimentation pour k=1 a été possible dans ce cas.

Ces limites nous ont poussées à réaliser une seconde expérimentation dans un environnement ouvert nous permettant d'accéder à plus de données (DBLP). Les détails de cette expérimentation sont présentés dans les sections qui suivent.

# 5.3 Evaluation sur les réseaux de co-auteurs d'articles scientifiques : cas de DBLP et Mendeley

Le but de cette seconde évaluation est de confirmer ou non les résultats obtenus dans l'évaluation préliminaire effectuée dans Facebook sur un jeu de données plus important. Dans ce qui suit, nous présentons tour à tour l'accès aux données, les méthodologies construction et de validation des dimensions des profils, les caractéristiques de l'échantillon de données, et les résultats obtenus.

## 5.3.1 Accès aux données dans DBLP

DBLP[50] est l'une des plus importantes librairies digitales qui référence de nombreux articles scientifiques publiés dans le domaine de l'informatique (conférences, journaux, séries, livres) dans le monde. Les données de ce site peuvent être accessibles au téléchargement[51] sous forme de fichier XML (DTD et contenus). Un descriptif sommaire des données contenues dans ce fichier est présenté dans (Ley et al., 09). Un exemple de représentation d'un article scientifique de ce fichier, est présenté dans la figure 5.7 où l'on peut aisément identifier le fait que l'article en question est publié dans un journal (attribut *key* de la balise <article>), l'auteur (balise <author>), le titre (balise <title>), les pages correspondantes à l'article dans le journal (balise <pages>), l'année de publication (balise <year>), le volume (balise <volume>), le numéro (balise <numero>), le nom du journal (balise <journal>), le DOI (balise <ee>), et l'URL associée (balise <url>).

Cependant, l'usage du fichier de données XML pouvant être assez lourd (plus de 1GB de nos jours), une API est également proposée par DBLP lorsqu'un développeur souhaite extraire uniquement certaines parties du fichier. Cette API est intéressante dans la mesure où elle offre la possibilité de renvoyer certaines vues (liste de publications d'un auteur donné, ou liste de co-auteurs d'un article donné par exemple)[52] [Ley et al., 09 bis]. Ces vues sont renvoyées en soumettant des requêtes sous forme d'URL (voir exemples sur le tableau 5.2).

```
<article key="journals/cacm/Szalay08"
     mdate="2008-11-03">
  <author>Alexander S. Szalay</author>
  <title>Jim Gray, astronomer.</title>
  <pages>58-65</pages>
  <year>2008</year>
  <volume>51</volume>
  <journal>Commun. ACM</journal>
  <number>11</number>
  <ee>http://doi.acm.org/10.1145/
     1400214.1400231</ee>
  <url>db/journals/cacm/
     cacm51.html#Szalay08</url>
</article>
```

Figure 5. 7: Exemple d'entrée du fichier XML de données de DBLP

[50] http://www.informatik.uni-trier.de/~ley/db/

[51] http://dblp.uni-trier.de/xml/

[52] http://dblp.uni-trier.de/xml/docu/dblpxmlreq.pdf

| Vue | Liste des publications d'un auteur donné | Caractéristiques d'une publication | Liste de co-auteurs d'un auteur donné |
|---|---|---|---|
| URL | http://dblp.uni-trier.de/rec/pers/*urlpt*/xk | http://dblp.uni-trier.de/rec/bibtex/journals/wias/key.xml | http://dblp.uni-trier.de/rec/pers/*urlpt*/xc |
| Remarque | Remplacer *urlpt* par l'identifiant qui est dérivé du nom entier de l'auteur (les règles de génération automatique de cet identifiant à partir du nom d'un auteur sont indiquées dans [Ley et al., 09 bis]) | Remplacer *key* par la clé d'une publication (*key* est l'attribut de la balise <article> d'une publication) | Remplacer *urlpt* par l'identifiant qui est dérivé du nom entier de l'auteur (les règles de génération automatique de cet identifiant à partir du nom d'un auteur sont indiquées dans [Ley et al., 09 bis]) |
| Exemple | Auteur : Dieudonné Tchuente<br><br>*urlpt* : t/Tchuente:Dieudonn=eacute=<br><br>voir fichier XML résultat figure 5.6.A | *Key : journals/wias/TchuenteCBPS12*<br><br>voir fichier XML résultat figure 5.6.B | Auteur : Dieudonné Tchuente<br><br>*urlpt* : t/Tchuente:Dieudonn=eacute=<br><br>voir fichier XML résultat figure 5.6.C |

Tableau 5. 2 : Exemples de vues renvoyées par l'API DBLP

A

```
- <coauthors author="Dieudonné Tchuente" urlpt="t/Tchuente:Dieudonn=eacute=">
    <author urlpt="b/Baptiste=Jessel:Nadine" count="3">Nadine Baptiste-Jessel</author>
    <author urlpt="c/Canut:C=_Marie=Fran=ccedil=oise" count="3">C. Marie-Françoise Canut</author>
    <author urlpt="h/Haddadi:Anass_El" count="1">Anass El Haddadi</author>
    <author urlpt="k/Kouamou:Georges_Edouard" count="1">Georges Edouard Kouamou</author>
    <author urlpt="p/P=eacute=ninou:Andr=eacute=" count="2">André Péninou</author>
    <author urlpt="s/Sedes:Florence" count="1">Florence Sedes</author>
  </coauthors>
```

B

```
- <dblpperson name="Dieudonné Tchuente">
    <dblpkey type="person record">homepages/01/6568</dblpkey>
    <dblpkey>journals/wias/TchuenteCBPS12</dblpkey>
    <dblpkey>conf/f-egc/TchuenteCB11</dblpkey>
    <dblpkey>conf/asunam/TchuenteCBPH10</dblpkey>
    <dblpkey>conf/icsea/KouamouT08</dblpkey>
  </dblpperson>
```

C

```
- <dblp>
  - <article key="journals/wias/TchuenteCBPS12" mdate="2012-05-10">
      <author>Dieudonné Tchuente</author>
      <author>C. Marie-Françoise Canut</author>
      <author>Nadine Baptiste-Jessel</author>
      <author>André Péninou</author>
      <author>Florence Sèdes</author>
    - <title>
        Visualizing the relevance of social ties in user profile modeling.
      </title>
      <pages>261-274</pages>
      <year>2012</year>
      <volume>10</volume>
      <journal>Web Intelligence and Agent Systems</journal>
      <number>2</number>
      <ee>http://dx.doi.org/10.3233/WIA-2012-0245</ee>
      <url>db/journals/wias/wias10.html#TchuenteCBPS12</url>
    </article>
  </dblp>
```

Figure 5. 8 : A) Liste de co-auteurs de l'auteur Dieudonné Tchuente. B) Liste de publications de l'auteur Dieudonné Tchuente, C) Exemple de description d'un article publié par l'auteur Dieudonné Tchuente.

Comme tout fichier XML de taille modeste, les fichiers XML renvoyés par ces types de vue peuvent être traités avec un parseur tel que SAX pour extraire les données [Ley et al., 09 bis].

Dans notre évaluation, nous exploitons la vue de co-auteurs (de manière récursive) pour construire le réseau k-égocentrique d'un auteur (égo). Pour k=1, il s'agira par exemple de considérer les relations de co-auteurs entre tous les co-auteurs d'un égo. A la différence de Facebook, les données sont complètement accessibles et il est possible d'accéder aux auteurs situés à distance k (k>=1) de l'égo. La nature du lien de co-auteurs entre deux auteurs sous- entend que ces deux derniers se connaissent à priori dans la vie réelle, ceci correspond bien à notre hypothèse de travail.

## 5.3.2 Méthodologie de construction des dimensions sociales et utilisateur des profils d'auteurs et processus de validation

Afin d'avoir des résultats les plus proches de la réalité, nous avons choisi d'utiliser deux sources de données distinctes pour construire les dimensions sociales et la dimension utilisateur du profil de chaque auteur analysé. Par dimensions sociales ici, nous entendons chacune des dimensions sociales construites par chacun des trois algorithmes à comparer.

### *5.3.2.1 Construction des dimensions sociales du profil d'un auteur*

La méthodologie est similaire à celle de Facebook (cf. figure 5.1), mais est plus simple dans la mesure où aucune taxonomie n'est utilisée (les domaines d'études sont beaucoup plus fermés par rapport à la multiplicité des domaines dans les activités des utilisateurs Facebook). Il n'est donc plus nécessaire de réaliser les actions de projection de catégories sur la taxonomie et de remontée des calculs dans la taxonomie (étape 1 et 4, cf. figure 5.1). Seules 3 étapes sont réalisées dans ce cas (figure 5.9).

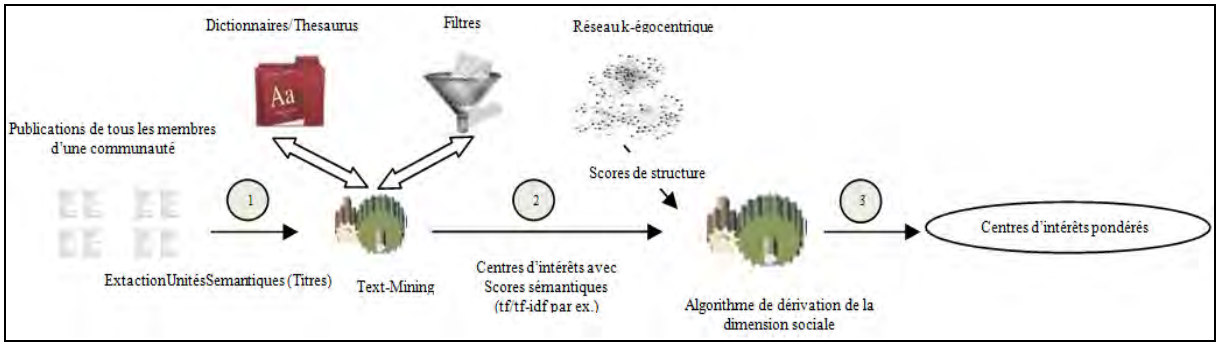



Figure 5. 9 : Construction de la dimension sociale des profils d'auteurs dans DBLP

- A la première étape, tous les titres des publications des membres d'une communauté sont utilisés pour extraire les unités sémantiques (cf. étape 1 figure 5.1, section 5.2.2.1.1).
- A la seconde étape, les unités sémantiques sont affinées avec les dictionnaires et filtres pour obtenir les centres d'intérêts pondérés avec la mesure *tf* dans cette évaluation (cf. étape 2 figure 5.1, section 5.2.2.1.2). Les scores de structure sont ensuite combinés avec les scores sémantiques selon les techniques exploitées par chacun des algorithmes de dérivation de la dimension sociale.

- Le profil de la communauté est obtenu en regroupant tous les centres d'intérêts pondérés de l'étape précédente au sein d'un vecteur unique de centres d'intérêts.

Il est à noter qu'une fois les profils des communautés (ou utilisateurs) construits, ils sont combinés par chacun des algorithmes pour obtenir la dimension sociale du profil.

### *5.3.2.2 Construction de la dimension utilisateur du profil d'un auteur*

Dans l'évaluation Facebook, la dimension utilisateur du profil d'un utilisateur était construite à partir des activités de ce dernier. La validation était alors réalisée par comparaison entre dimension utilisateur et dimensions sociales. Dans le cas de DBLP, nous souhaitons nous rapprocher le plus possible de la réalité en considérant comme dimension utilisateur, le profil des auteurs décrit par eux-mêmes. La validation ici se rapproche alors d'une validation par confrontation à la perception humaine.

Pour ce faire, nous nous sommes alors posé la question concernant l'accès aux profils d'auteurs d'articles scientifiques renseignés par ces derniers. Il s'avère que ces dernières années de nouveaux sites de réseaux sociaux numériques spécifiques aux auteurs d'articles scientifiques ont vu le jour (researchgate[53] ou mendeley[54] par exemple). Ces sites prévoient en général des rubriques permettant aux auteurs d'indiquer leurs centres d'intérêts. La figure 5.10 présente un exemple de centres d'intérêts d'un auteur (domaine de l'informatique) renseigné par lui-même sur le site mendeley.com.


Figure 5. 10 : Exemple de centres d'intérêts d'un auteur sur le site mendeley.com

Dans notre évaluation, nous exploitons donc le texte représentant les centres d'intérêts des auteurs sur le site mendeley.com pour dériver les centres d'intérêts en extrayant les unités sémantiques et en utilisant les dictionnaires et filtres (figure 5.11).

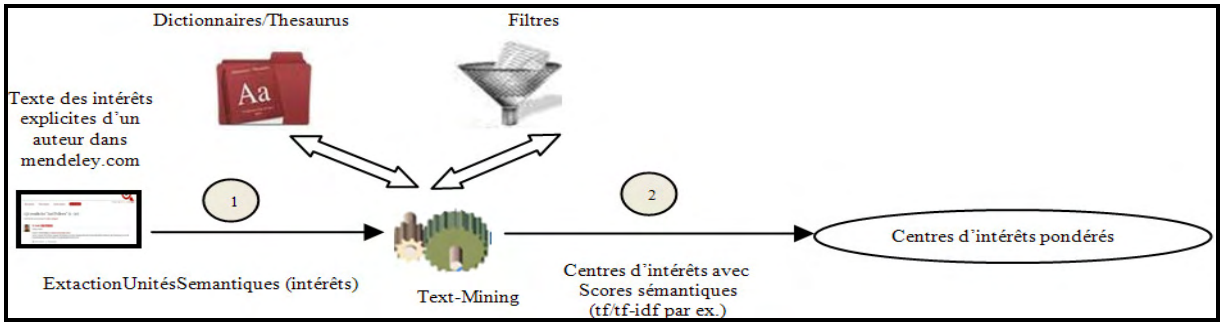



Figure 5. 11 : Construction de la dimension utilisateur du profil d'un auteur dans Mendeley

[53] *www.researchgate.net/*

[54] www.mendeley.com/

### *5.3.2.3 Intégration de sources de données pour validation*

La méthodologie de validation utilisée dans cette évaluation (figure 5.12) consiste à exploiter la dimension utilisateur du profil de chaque auteur construit dans Mendeley (étape 2) pour évaluer la pertinence (étape 4) de chacun des algorithmes de construction de la dimension sociale (étape 3) du profil du même auteur dans DBLP. Les étapes 2 et 3 correspondent respectivement aux phases de construction de la dimension utilisateur (cf. section 5.3.2.2) et de la construction de la dimension sociale (cf. section 5.3.2.1) du profil d'un auteur.

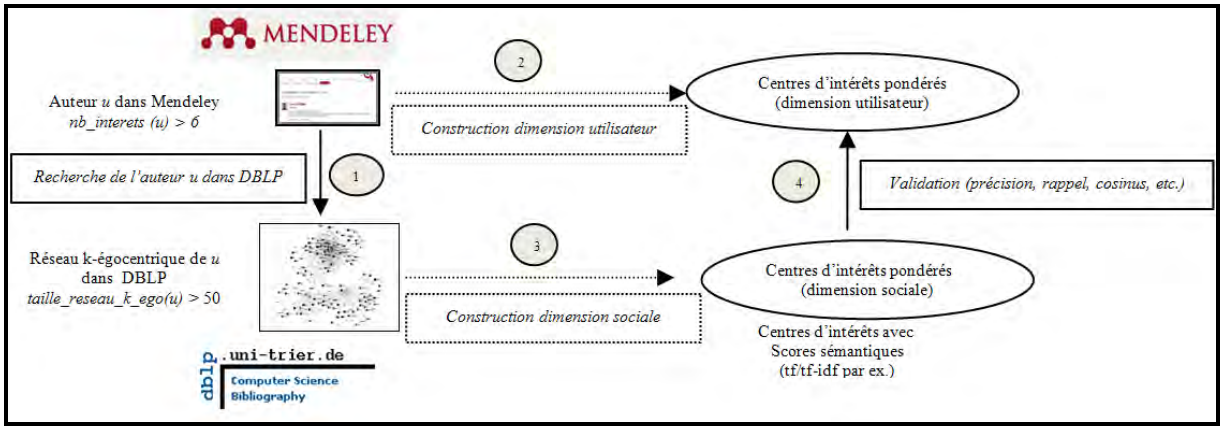



Figure 5. 12 : Processus de validation adoptée pour l'évaluation dans DBLP

La principale question qui se pose dans ce processus est de savoir comment intégrer les deux sources de données (Mendeley et DBLP) pour la validation (étape 1). Pour ce faire nous avons étudié deux approches :

- ***approche automatique :*** en utilisant les API des deux plateformes et en recherchant les utilisateurs possédant le même nom (en évitant les cas d'homonymie). Cependant, à la différence des données de DBLP, les données de Mendeley ne sont pas ouvertes comme dans le cas de Facebook (besoin des autorisations explicites d'auteurs pour accéder aux données). Le problème de l'accès sous autorisation rencontré dans l'évaluation Facebook se pose donc dans ce cas. La taille de l'échantillon à analyser risque d'être fortement réduite (car nécessité de trouver des auteurs volontaires et très actifs) comme dans le cas de l'évaluation Facebook,

- ***approche manuelle :*** en recherchant manuellement les auteurs de Mendeley qui existent dans DBLP (en évitant les cas d'homonymie). Dans la mesure où seuls, les centres d'intérêts des auteurs de Mendeley nous intéressent dans cette évaluation, nous avons plutôt opté pour cette solution manuelle avec laquelle il a été plus facile de retrouver un nombre assez important d'auteurs (en peu de temps) pour lesquels l'évaluation pouvait être réalisée en rajoutant deux contraintes supplémentaires :

    - L'auteur doit avoir indiqué un nombre important de centres d'intérêts dans Mendeley. En effet, il s'agit d'avoir une dimension utilisateur assez riche en centres d'intérêts qui seront prédits par chacune des dimensions sociales du profil du même auteur construit à partir de son réseau k-égocentrique dans DBLP. Nous avons choisi dans cette évaluation des auteurs de Mendeley disposant d'au moins *6 (nb_interets>6 sur la* figure 5.12*)* termes distincts (centres d'intérêts) dans la liste de leurs centres d'intérêts indiqués explicitement.

- L'auteur doit avoir un nombre important d'utilisateurs dans son réseau k-égocentrique dans DBLP. Il est évident que pour un auteur ayant très peu d'auteurs dans son réseau k-égocentrique, l'approche par communautés sera moins pertinente dans la mesure où l'algorithme de détection de communautés risquerait de renvoyer très peu ou pas de communautés. Nous avons choisi dans cette évaluation des auteurs qui dans DBLP ont au moins *50 (taille_reseau_k_ego>50 sur la* figure 5.12*)* auteurs dans leur réseau k-égocentrique.

La validation est réalisée en comparant le degré avec lequel chaque algorithme de dérivation de la dimension sociale (dans DBLP) prédit les centres d'intérêts explicitement indiqués par l'auteur dans Mendeley (étape 4). Le fait d'exploiter les profils « réels » indiqués par les auteurs eux-mêmes pour valider les dimensions sociales de leur profil construit dans DBLP permettra également de mieux juger de la pertinence des approches sociales, car le processus de validation est très proche de la validation par confrontation à la perception humaine (des auteurs). Pour l'étape de validation nous avons considéré trois mesures qui permettent de valider à la fois des taux de présence/absence (*précision*, *rappel*) des centres d'intérêts de la dimension sociale dans la dimension utilisateur, ainsi que la qualité des poids associés dans la dimension sociale (*cosinus de similarité*) pour prédire efficacement la dimension utilisateur.

Dans notre contexte d'évaluation, la précision d'un algorithme de dérivation de la dimension sociale est évaluée par le nombre de centres d'intérêts qu'il prédit dans la dimension utilisateur, divisé par le nombre total de centres d'intérêts qu'il renvoie dans la dimension sociale qu'il calcule.

$$\text{Pr}\,écision = \frac{nombre\_centres\_\text{int}erets\_predits\_dans\_\text{dim}ension\_utilisateur}{nombre\_centres\_\text{int}erets\_calcules\_dans\_\text{dim}ension\_sociale} \qquad \textbf{(41)}$$

Le rappel d'un algorithme de dérivation de la dimension sociale est évalué par le nombre de centres d'intérêts qu'il prédit dans la dimension utilisateur, divisé par le nombre de centres d'intérêt présents dans la dimension utilisateur.

$$Rappel = \frac{nombre\_centres\_\text{int}erets\_predits\_dans\_\text{dim}ension\_utilisateur}{nombre\_total\_centres\_\text{int}erets\_dans\_\text{dim}ension\_utilisateur} \qquad \textbf{(42)}$$

Pour le calcul de la précision et du rappel, nous nous intéressons uniquement aux centres d'intérêts les plus pertinents renvoyés par chaque algorithme de dérivation de la dimension sociale et non à l'ensemble de tous les centres renvoyés (car en général, seuls les centres d'intérêts les plus pertinents sont exploités dans les systèmes d'adaptation de l'information à l'utilisateur). Ainsi, si la dimension utilisateur d'un profil est constituée de *n* centres d'intérêts (vecteur de taille *n*), la précision et le rappel de chaque algorithme de dérivation de la dimension sociale seront calculés à partir du *top n+m (avec m>0)* premiers centres d'intérêts (vecteur de taille *n+m*) présent dans la dimension sociale.

Le cosinus de similarité entre la dimension sociale construite par un algorithme de dérivation de la dimension sociale et la dimension utilisateur du profil d'un utilisateur est le cosinus des vecteurs représentant chacune de ces dimensions (cf. formule 2).

Après avoir présenté le processus de validation qui a été utilisée dans cette évaluation, nous présentons dans les sections qui suivent, les caractéristiques de l'échantillon de données analysées, ainsi que les résultats obtenus.

## 5.3.3 Caractéristiques de l'échantillon de données étudié

Dans cette évaluation, nous avons fixé la valeur du paramètre k à 1. Donc on parlera dans la suite de réseau égocentrique (réseau de relations de co-auteur entre les co-auteurs d'un auteur donné) à chaque fois qu'on parlera du contexte précis de cette évaluation.

Comme indiqué précédemment (cf. figure 5.12), nous avons opté pour le croisement manuel des données des auteurs dans Mendeley (auteurs ayant plus de 6 centres d'intérêts indiqués explicitement) avec ceux du réseau égocentrique des mêmes auteurs dans DBLP (auteurs ayant au minimum 50 co-auteurs).

Nous avons limité notre recherche après avoir trouvé 105 auteurs répondant à ces deux critères. Le tableau 5.3 présente quelques statistiques descriptives sur ces 105 auteurs.

| Total de réseaux égocentriques étudiés | Moyenne du nombre de co-auteurs par auteur | Total d'auteurs utilisés dans les analyses | Maximum des nombres de co-auteurs | Minimum des nombres de co-auteurs | Moyenne de centres d'intérêts dans les dimensions utilisateur | Densité Moyenne des réseaux égocentriques |
|---|---|---|---|---|---|---|
| 105 | 98 | 10008 | 501 | 50 | 11 | 12% |

Tableau 5. 3 : Quelques statistiques descriptives sur l'échantillon de données dans DBLP

La moyenne du nombre de co-auteurs de l'échantillon analysé est de 98, pour un total de 10008 auteurs distincts utilisés dans les analyses. Le nombre maximum de co-auteurs pour un auteur est de 501, et le minimum est bien sûr de 50 (contrainte de départ). La figure 5.13 présente le nuage de points dans lequel chaque point représente un auteur pour lequel on a, en abscisse son nombre de co-auteurs et en ordonnée le total d'auteurs disposant de ce même nombre de co-auteurs. La courbe formée par ce nuage de points est similaire à la même courbe réalisée pour la totalité des auteurs dans DBLP (Zeng et al., 10). En réalité, la courbe entre auteurs et nombres de co-auteurs dans DBLP suit une loi de puissance (beaucoup d'auteurs ont peu de co-auteurs et très peu d'auteurs ont un grand nombre de co-auteurs). L'échantillon d'auteurs analysé dans cette évaluation est donc représentative (en terme de distribution suivant une loi de puissance) de tous les auteurs dans DBLP (Zeng et al., 10).

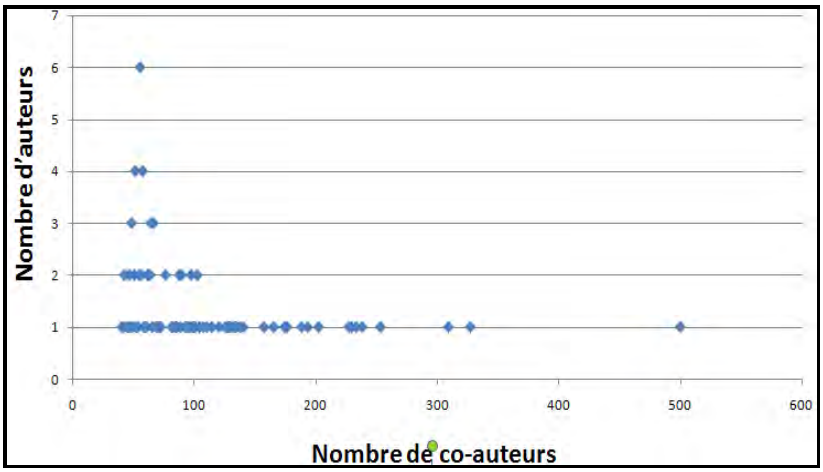



Figure 5. 13 : Nuage de points entre nombre d'auteurs et nombre de co-auteurs de l'échantillon d'auteurs

Le nombre moyen de centres d'intérêts dans les dimensions utilisateur des profils construits à partir de Mendeley est de **11**. Enfin, la densité moyenne des réseaux égocentriques est de **12%**.

Nous entendons par densité d'un réseau k-égocentrique, le nombre de liens existants dans le réseau k-égocentrique (noté $N_L$), par rapport au nombre total de liens possibles dans ce réseau (formule 38). Si un réseau égocentrique est constitué de n utilisateurs, le nombre total de liens possibles dans ce réseau est exprimé par n(n-1)/2, donc sa densité est exprimée par :

$$d = \frac{2 * N_L}{n(n-1)} \qquad \textbf{(38)}$$

Nous nous intéressons particulièrement à la densité des réseaux k-égocentriques car nous pensons que cette densité aura un impact sur la pertinence de l'algorithme basé sur les communautés. Nous supposons en effet que plus un réseau k-égocentrique sera dense, plus il sera plausible d'y détecter des communautés fortement significatives pour l'égo analysé. La figure 5.14 montre le nuage de points présentant pour chaque auteur de l'échantillon, la densité de son réseau égocentrique.

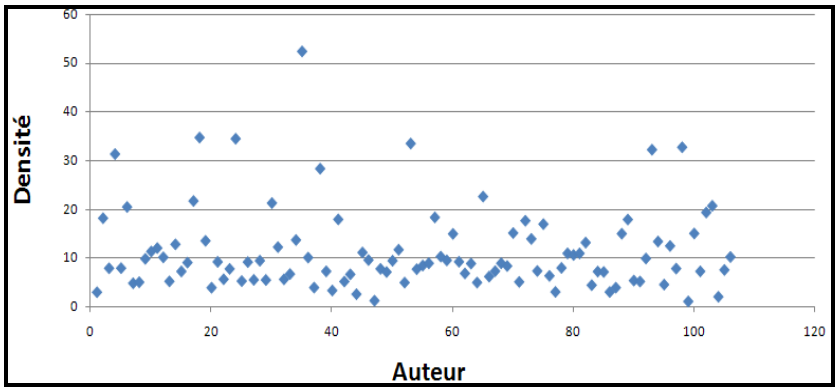



Figure 5. 14 : Nuage de points présentant la densité du réseau égocentrique de chaque auteur de l'échantillon

La moyenne de densité se situe à environ 12% et on remarque une égale distribution du nombre d'auteurs en deca et au dessus de 10%. Nous analyserons l'impact de la densité sur les résultats obtenus dans la section suivante.

## 5.3.4 Résultats

La mesure de structure exploitée dans cette évaluation est la centralité de degré (pour les communautés comme pour les utilisateurs).

Nous présentons les résultats d'une part par rapport au paramètre α, et d'autre part par rapport à la densité des réseaux égocentriques des auteurs (Tchuente et al., 13 bis).

### *5.3.4.1 Comparaisons relatives au paramètre de structure et à la densité*

#### *5.3.4.1.1 Sur tout l'échantillon de données*

La figure 5.15 présente le comparatif des dimensions sociales construites par les trois algorithmes étudiés en fonction du paramètre α. Cette comparaison concerne les dimensions utilisateurs et sociales de tous les auteurs (égos) de l'échantillon de donnée (105 réseaux égocentriques).

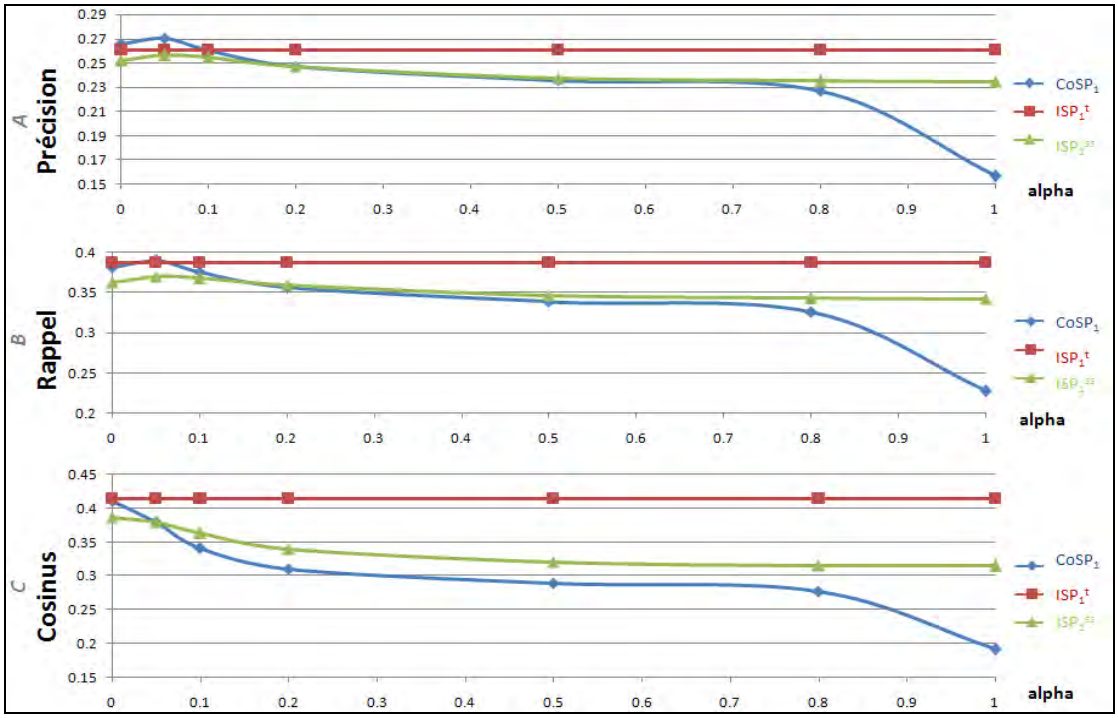



Figure 5. 15 : Comparatif de la pertinence des dimensions sociales construites par les algorithmes étudiés $CoSP_1$, $ISP_1^{ss}$, $ISP_1^{t}$, en fonction du paramètre α.

La figure 5.15A montre le comparatif par rapport à la précision, la figure 5.15B montre le comparatif par rapport au rappel, et la figure 5.15C montre le comparatif par rapport au cosinus de similarité.

A la différence de l'évaluation préliminaire dans Facebook, l'algorithme basé sur les individus en exploitant les mesures de structures ($ISP_1^{ss}$) produit plutôt de moins bons résultats par rapport à l'algorithme trivial basé sur les individus ($ISP_1^{t}$). L'algorithme basé sur les communautés n'est que très légèrement plus pertinent que les autres algorithmes (en terme de précision notamment). Pour approfondir les analyses, nous avons réalisé ce même comparatif en prenant en compte la densité des réseaux égocentriques.

#### *5.3.4.1.2 Sur les auteurs de densité supérieure à 10% (plus de 90% de l'échantillon)*

La figure 5.16 présente le comparatif des dimensions sociales construites par les trois algorithmes étudiés en fonction du paramètre α, pour les auteurs dont la densité du réseau égocentrique est supérieure à 10% (plus de 90% de l'échantillon de données).

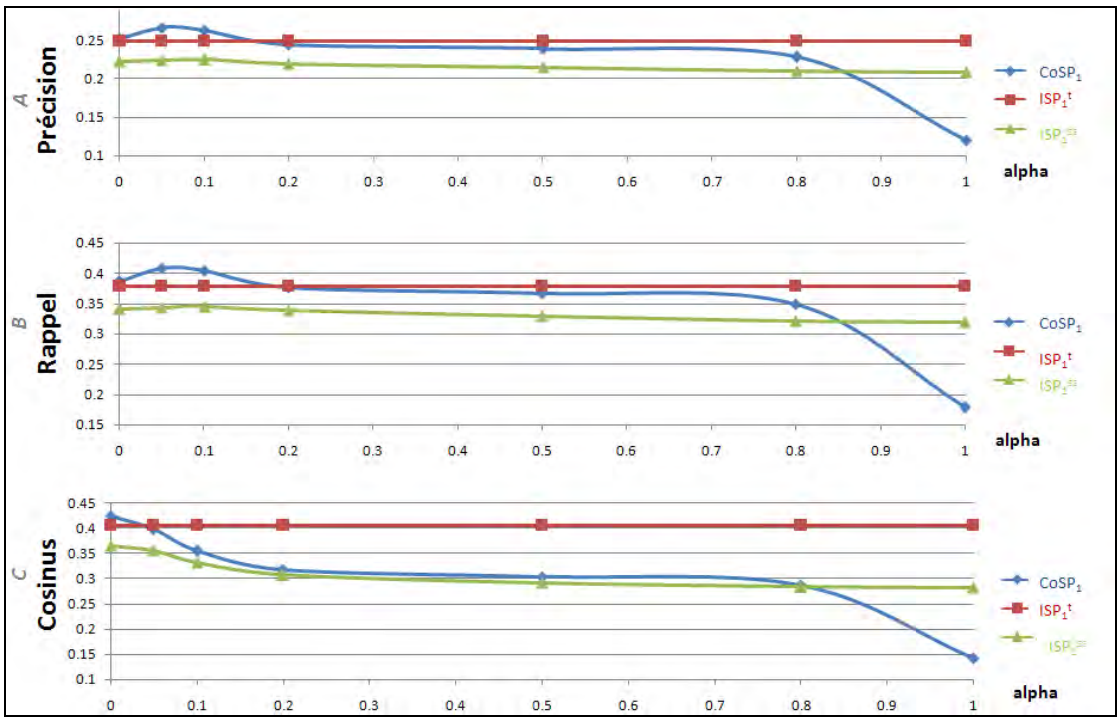



Figure 5. 16 : Comparatif de la pertinence des dimensions sociales construites par les algorithmes étudiés $CoSP_1$, $ISP_1^{ss}$, $ISP_1^{t}$ (rouge), en fonction du paramètre α, pour les auteurs dont la densité du réseau égocentrique est supérieure à 10%.

La figure 5.16A montre le comparatif par rapport à la précision, la figure 5.16B montre le comparatif par rapport au rappel, et la figure 5.16C montre le comparatif par rapport au cosinus de similarité.

Par rapport à l'observation précédente sur tout l'échantillon, nous remarquons principalement que ces courbes sont plus semblables à celles obtenues dans l'évaluation préliminaire réalisée dans Facebook. L'algorithme basé sur les communautés ($CoSP_1$) produit les dimensions sociales les plus pertinentes pour de faibles valeurs du paramètre α ($\alpha \in [0, 0.1]$), ceci qu'elle que soit la mesure comparative (précision, rappel, cosinus).

#### *5.3.4.1.3 Sur les auteurs de densité supérieure à 20% (plus de 50% de l'échantillon)*

Pour aller plus loin dans les analyses, nous avons réalisé le même comparatif que précédemment en considérant uniquement les auteurs dont la densité du réseau égocentrique est supérieure à 20% (un peu plus de la moitié de l'échantillon de données). Les comparatifs sont présentés sur la figure 5.17.

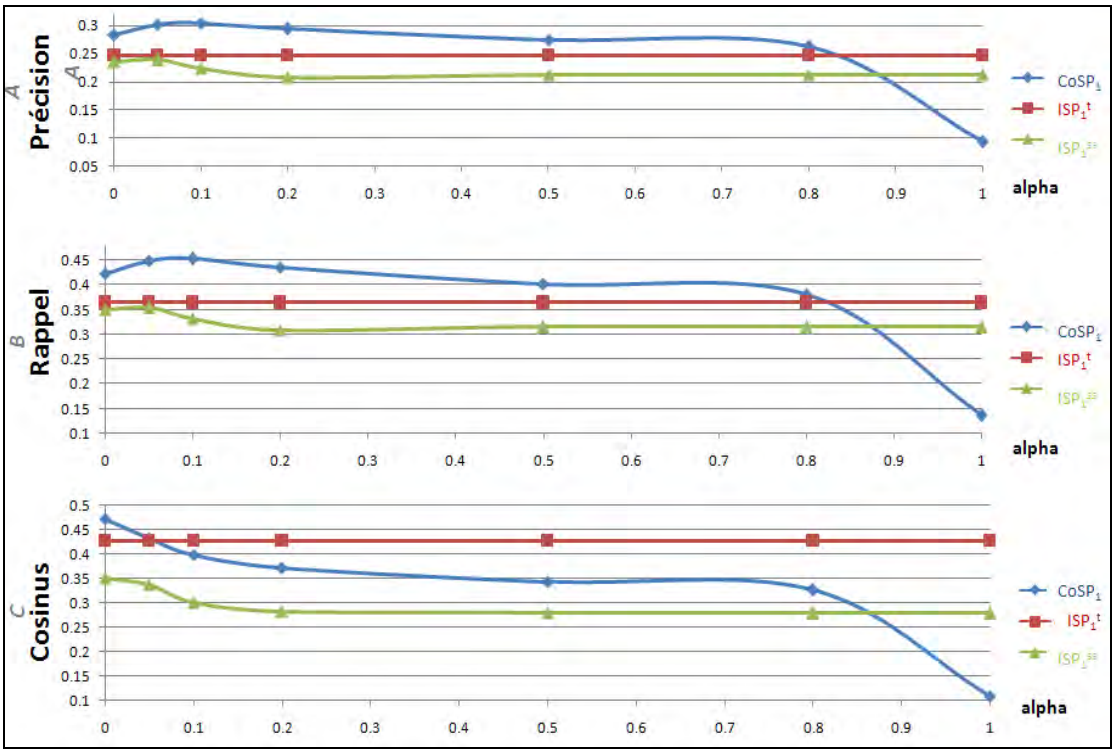



Figure 5. 17 : Comparatif de la pertinence des dimensions sociales construites par les algorithmes étudiés $CoSP_1$, IBSP1, $ISP_1^t$, en fonction du paramètre α, pour les auteurs dont la densité du réseau égocentrique est supérieure à 20%.

La figure 5.17A montre le comparatif par rapport à la précision, la figure 5.17B montre le comparatif par rapport au rappel, et la figure 5.17C montre le comparatif par rapport au cosinus de similarité.

Par rapport aux résultats observés dans les cas précédents, on remarque surtout que l'algorithme basé sur les communautés ($CoSP_1$) est beaucoup plus pertinent (les écarts entre les courbes sont plus nettes). Les meilleurs résultats sont également observés pour des valeurs de $\alpha \in [0, 0.1]$ (notamment en terme de précision et de rappel). Même au-delà de $[0, 0.1]$, entre $[0, 0.8]$ l'algorithme basé sur les communautés demeure le plus pertinent en terme de précision et de rappel.

#### *5.3.4.1.4 Sur les auteurs de densité supérieure à 30% (plus de 30% de l'échantillon)*

Pour confirmer les résultats obtenus dans l'analyse précédente (auteurs dont la densité su réseau égocentrique est supérieure à 20%), nous avons réalisé le même comparatif que précédemment en considérant uniquement les auteurs dont la densité du réseau égocentrique est cette fois supérieure à 30% (près du tiers de l'échantillon de données). Les comparatifs sont présentés sur la figure 5.18.

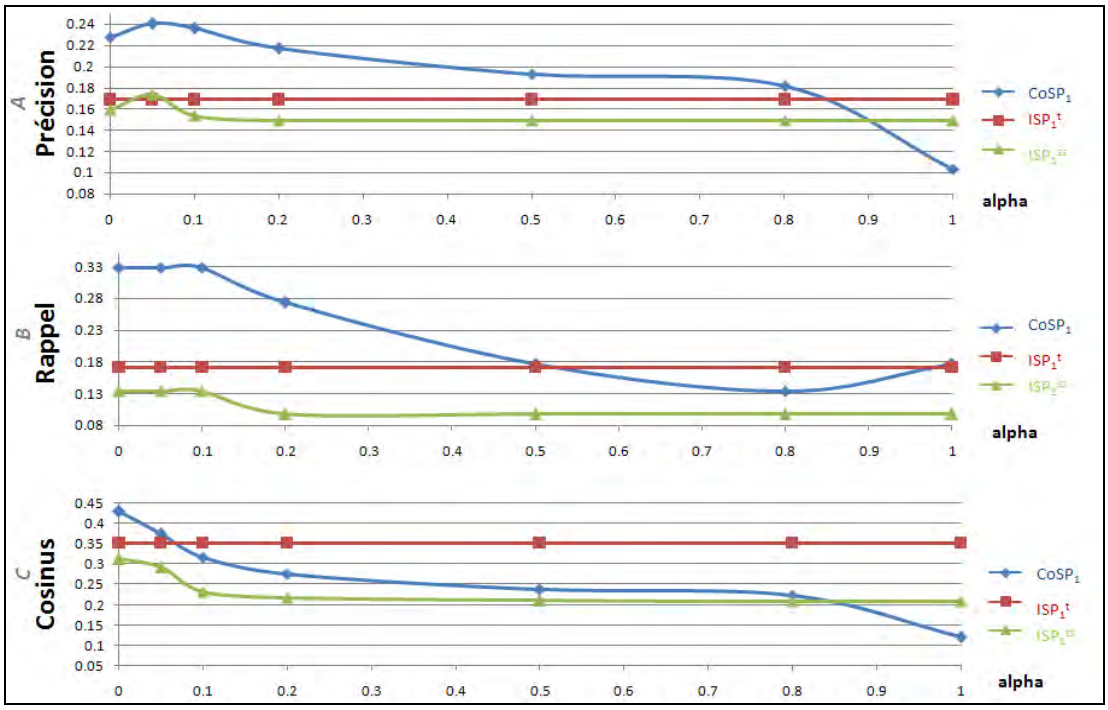



Figure 5. 18 : Comparatif de la pertinence des dimensions sociales construites par les algorithmes étudiés $CoSP_1$, $ISP_1^{ss}$, $ISP_1^{t}$, en fonction du paramètre α, pour les auteurs dont la densité du réseau égocentrique est supérieure à 30%.

La figure 5.18A montre le comparatif par rapport à la précision, la figure 5.18B montre le comparatif par rapport au rappel, et la figure 5.18C montre le comparatif par rapport au cosinus de similarité.

Par rapport aux résultats observés précédemment, on remarque surtout que l'écart entre la courbe bleue (algorithme basé sur les communautés) et les autres courbes (algorithmes basés sur les individus) est beaucoup plus importante. Les résultats les plus pertinents sont toujours observés pour des valeurs de $\alpha \in [0, 0.1]$ (notamment en terme de précision et de rappel).

Ces deux derniers comparatifs relatifs à la densité du réseau égocentriques nous permettent d'observer que plus la densité du réseau égocentrique est élevée, plus l'algorithme basé sur les communautés produit de meilleurs résultats comparativement aux algorithmes basés sur les individus. Ceci se justifie logiquement par le fait que plus le réseau égocentrique sera éparse (peu dense), moins les communautés extraites par l'algorithme de détection de communautés seront réellement significatives pour l'égo. Alors que plus ce réseau sera dense, l'algorithme de détection de communautés sera capable d'extraire des communautés plus significatives pour l'égo.

Nous avons réalisés des analyses plus spécifiques en fixant une valeur pour le paramètre α.

### *5.3.4.2 Comparaisons relatives à la densité et au nombre de co-auteurs*

Pour mieux évaluer l'impact de la densité sur les résultats des algorithmes étudiés, nous avons fixé le paramètre α à 0.08 (valeur quasi optimale selon les observations précédentes), et comparé les algorithmes suivant différentes valeurs de densité.

#### *5.3.4.2.1 Sur tous l'échantillon de données*

Le comparatif suivant la densité sur tout l'échantillon de données est présenté sur la figure 5.19.

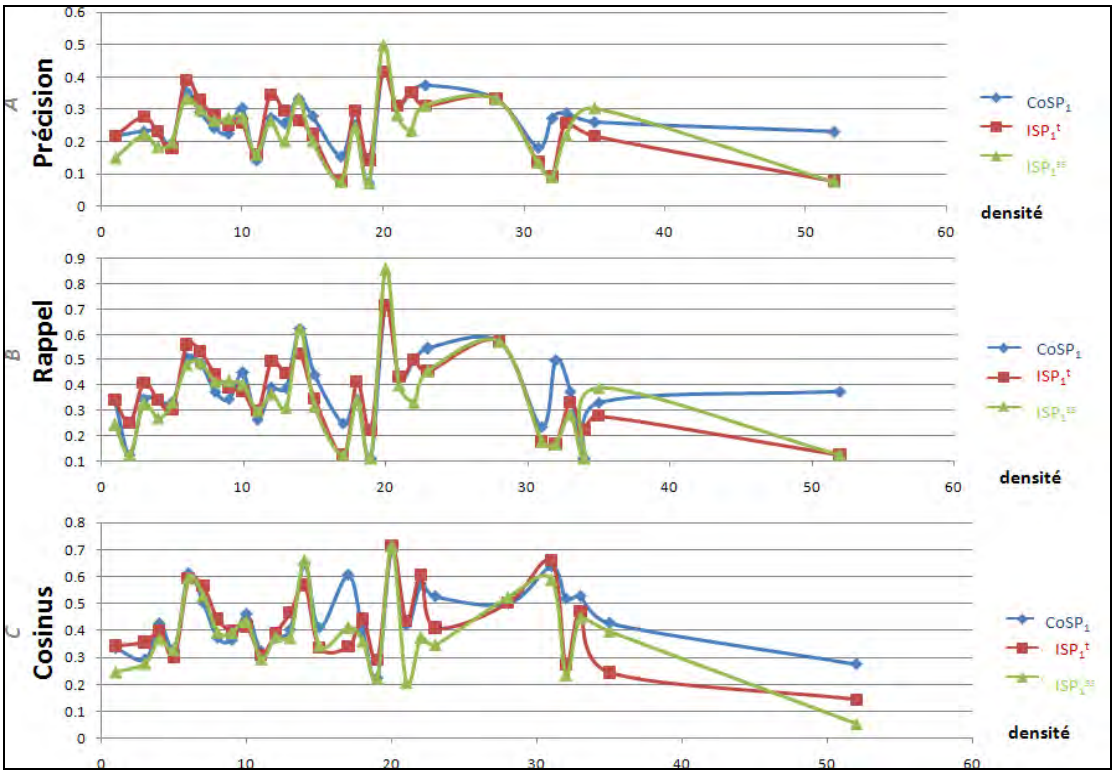



Figure 5. 19 : Comparatif de la pertinence des dimensions sociales construites par les algorithmes étudiés $CoSP_1$, $ISP_1^{ss}$, $ISP_1^{t}$, en fonction de la densité.

La figure 5.19A montre le comparatif par rapport à la précision, la figure 5.19B montre le comparatif par rapport au rappel, et la figure 5.19C montre le comparatif par rapport au cosinus de similarité.

Nous observons que pour les valeurs de densité < 30%, il est difficile de déterminer quel est le meilleur algorithme. Toutefois pour des valeurs de densité > 30%, l'algorithme basé sur les communautés fournit donne les meilleurs résultats quel que soit la mesure comparative (précision, rappel ou cosinus).

Ce résultat est un peu en contradiction avec les résultats (pour des densités < 30%) avec ceux observés lors des analyses en fonction du paramètre α (et pour des densités supérieures à 10% et 20%). Pour mieux comprendre cette contradiction, nous avons considéré un paramètre supplémentaire dans les analyses (le nombre d'individus dans le réseau k-égocentrique, nombre de co-auteurs d'un égo dans ce cas précis).

#### *5.3.4.2.2 Sur les auteurs ayant plus de 70 co-auteurs*

Nous avons réalisé le même comparatif que précédemment en considérant uniquement les auteurs ayant au minimum 70 co-auteurs (3/4 de l'échantillon de données, rappelons que la moyenne du nombre de co-auteur est de 98 dans l'échantillon de données). Le comparatif obtenu est présenté sur la figure 5.20.

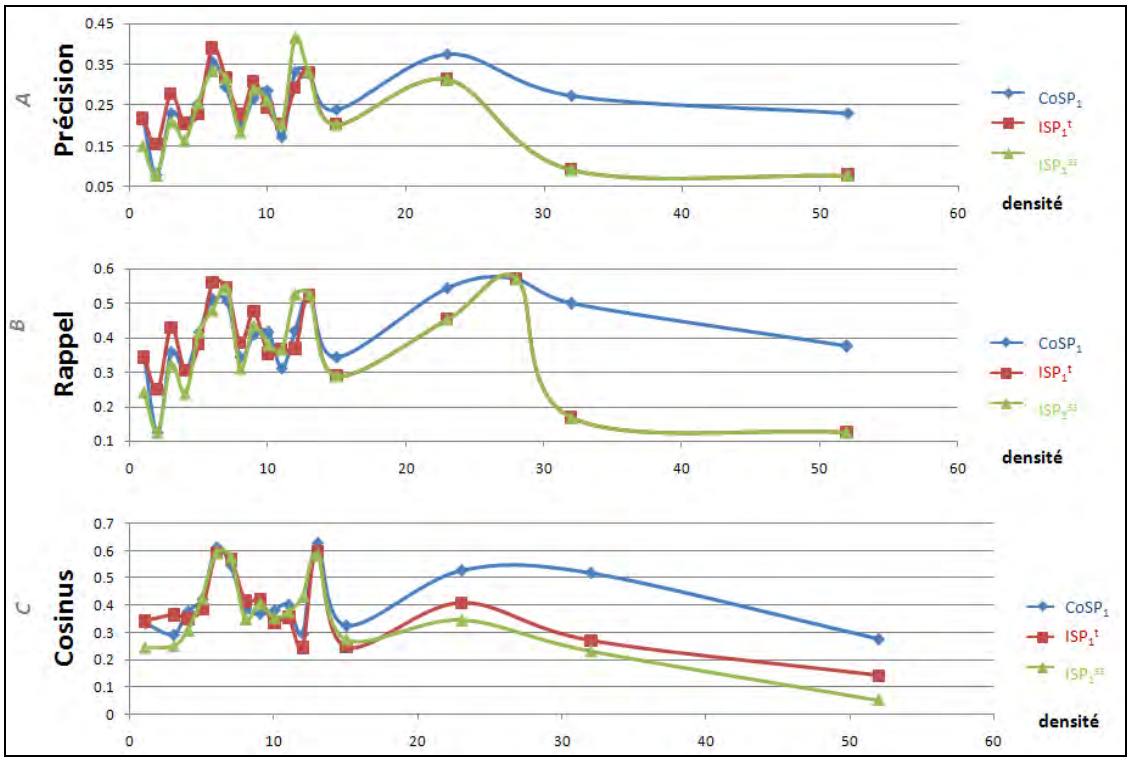



Figure 5. 20 : Comparatif de la pertinence des dimensions sociales construites par les algorithmes étudiés $CoSP_1$, $ISP_1^{ss}$, $ISP_1^{t}$, en fonction de la densité, pour les auteurs ayant plus de 70 co-auteurs.

La figure 5.20A montre le comparatif par rapport à la précision, la figure 5.20B montre le comparatif par rapport au rappel, et la figure 5.20C montre le comparatif par rapport au cosinus de similarité.

Nous observons sur ce comparatif que pour des densités supérieures à environ 12%, l'algorithme basé sur les communautés est nettement meilleur quel que soit la mesure comparative (précision, rappel ou cosinus). Ces résultats sont plus en adéquation avec ceux observés dans les analyses suivant le paramètre α. La densité qui produit des résultats optimaux se situe à entre [20, 25].

On remarque tout de même que même si l'algorithme basé sur les communautés reste meilleur pour des valeurs élevé de densité (entre [20, 25]), les résultats obtenus sont moins importants lorsque la densité devient supérieure à 30 (courbes décroissantes lorsque la densité >30). Donc, les réseaux égocentriques très denses fournissent des dimensions sociales moins pertinentes pour l'utilisateur. Ceci s'explique par le fait que si le réseau est très dense, il est probable que l'algorithme de détection de communautés détecte très peu de communautés (une seule grosse communauté par exemple) qui ne sera pas assez discriminante pour bien caractériser l'utilisateur par rapport aux cas où plusieurs communautés sont détectées.

Au final, on remarque que pour des densités très faibles (densité < 10%) et des densités très fortes (densité > 30%) les dimensions sociales construites sont moins pertinentes, du fait que les communautés extraites dans ces cas soient très peu discriminantes pour l'égo. Dans tous les cas l'algorithme basé sur les communautés produit de meilleurs résultats pour des réseaux égocentriques de densité supérieure à 10%. Pour des densités < 10% les trois algorithmes fournissent des résultats quasi identiques.

Tous les résultats présentés jusqu'ici exploitent la mesure de degré de centralité (de groupes ou d'individus) comme mesure de structure. Dans la section qui suit, nous présentons les résultats comparatifs impliquant d'autres mesures de structure.

### *5.3.4.3 Comparaisons de l'impact de différentes mesures de structure*

Les résultats présentés précédemment montrent la pertinence de l'algorithme basée sur les communautés par rapport aux algorithmes basés sur les individus. Dans cette section, nous nous intéressons uniquement à l'algorithme basé sur les communautés en recherchant la mesure de structure qui impacte le mieux la qualité des résultats obtenus (seule la centralité de degré a été utilisée dans les résultats précédents). Pour ce faire, nous avons utilisé à l'étape de caractérisation sémantico-structurelle (calcul de P'''), les mesures de centralité de degré (formule 14), de proximité (formule 15), et de cohésion (formule 27) et comparé chacune des dimensions sociales obtenues avec la dimension utilisateur des profils de tous les auteurs de notre échantillon.

Pour la mesure de proximité de groupe, nous avons utilisé trois mesures de distances ($d_f$ de la formule 15) : la distance moyenne entre les individus d'une communauté et ceux à l'extérieur de la communauté (proximité moyenne), la distance minimum entre les individus d'une communauté et ceux à l'extérieur de la communauté (proximité minimum), la distance maximum entre les individus d'une communauté et ceux à l'extérieur de la communauté (proximité maximum). Au total 5 mesures de structures sont donc utilisées : le degré, la cohésion, la proximité moyenne, la proximité minimale, la proximité maximum. Les résultats obtenus (précision, rappel, cosinus) sont présentés sur la figure 5. 21.

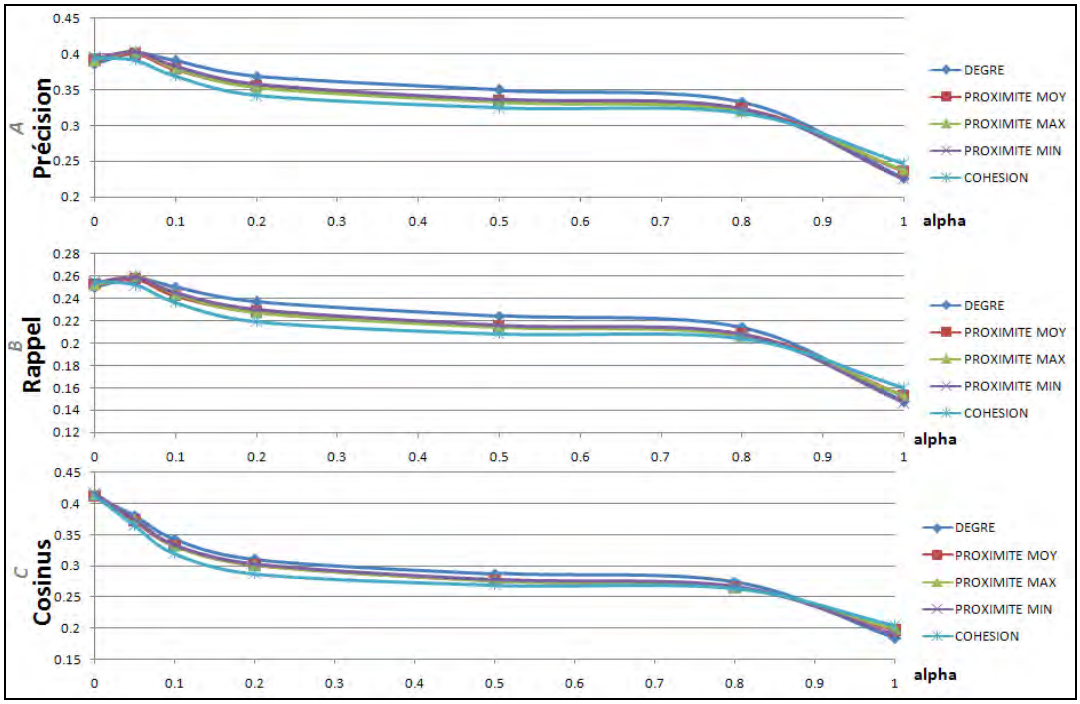



Figure 5. 21: Comparaison d'impact de différentes mesures de structure sur l'algorithme basé sur les communautés $CoSP_1$

Lorsque le paramètre α vaut 0, les résultats sont évidemment tous égaux car seule la mesure sémantique entre en jeu dans la pondération des centres d'intérêts. On constate de manière globale que les meilleurs résultats obtenus sont pour de très faibles valeurs du paramètre (intervalle [0, 0.1]) quelque soit la mesure de structure

(courbes de précision et rappel). On constate également que la centralité de degré (courbes en bleue foncé) fournit légèrement de meilleurs résultats que toutes les autres mesures. Les mesures de proximité (moyenne, minimale et maximale) fournissent des résultats quasi identiques (courbes rouge, vert, violet). Enfin, la mesure de cohésion des communautés fournit légèrement les moins bons résultats (courbes en bleu clair).

Les scores de structure impactent positivement (légère inflexion des courbes de précision et rappel) la qualité de résultats pour de très faibles valeurs du paramètre $\alpha$ (approximativement 0.08) par rapport aux cas où elles ne sont pas prises en compte (lorsque $\alpha$ vaut 0). Pour des valeurs du paramètre $\alpha$ supérieures à 0.1, l'impact de ces mesures est négatif sur la qualité des profils obtenus (courbes décroissantes à partir de la valeur 0.1). Au final, on retient alors que l'usage des scores de structure peut améliorer les résultats, à conditions qu'ils contribuent à hauteur de 8% environ (relativement aux scores sémantiques) dans le calcul de la pondération des centres d'intérêts dans la dimension sociale du profil de l'utilisateur.

# 5.4 Conclusion

Dans ce chapitre nous avons présenté les résultats des expérimentations et évaluations réalisées dans Facebook et DBLP afin d'évaluer la pertinence de notre approche de dérivation du profil social d'un utilisateur à partir de communautés de son réseau égocentrique.

L'expérimentation réalisée sur Facebook a été effectuée sur 15 réseaux égocentriques d'utilisateurs très actifs, pour un total de 3525 profils analysés. La dimension utilisateur de chaque profil a été construite à partir des activités des utilisateurs, et la dimension sociale a été construite à partir des activités des membres du réseau égocentrique de chaque utilisateur (avec chacun des trois algorithmes à évaluer). L'évaluation a alors consisté à déterminer l'algorithme de dérivation de la dimension sociale qui prédit le mieux les informations de la dimension utilisateur de chacun des profils avec la mesure du cosinus de similarité (suivant plusieurs niveaux de granularité dans la taxonomie de référence). Les résultats de cette évaluation nous a permis de tirer deux principales conclusions :

- L'algorithme proposé basé sur les communautés fournit de meilleurs résultats que les algorithmes basés sur les individus.
- Les scores de structure des communautés peuvent influencer positivement la qualité des dimensions sociales construites lorsque leur niveau d'importance est faible par rapport aux mesures sémantiques (notamment lorsque le paramètre de structure $\alpha$ est dans l'intervalle $]0, 0.1]$).

Toutefois, le nombre de réseaux égocentriques étudiés dans cette première évaluation n'étant pas très important (dû à la difficulté d'accès aux données dans Facebook), nous avons réalisé une seconde évaluation dans un environnement dans lequel l'accès aux données est complètement ouvert (DBLP).

L'expérimentation réalisée dans DBLP a été effectuée sur 105 réseaux égocentriques d'auteurs d'articles scientifiques. Afin d'éviter le biais qu'il peut y avoir à construire la dimension utilisateur du profil d'un égo à partir des mêmes articles présents dans les publications des membres de son réseau égocentrique, nous avons choisi d'utiliser comme dimension utilisateur, les profils renseignés explicitement par les égos eux-mêmes. Ainsi, nous avons exploité une seconde source de données (site Mendeley) dans lequel certains auteurs

indiquent de manière explicite la liste de leurs centres d'intérêts. L'expérimentation a alors consisté à construire les dimensions sociales des profils des auteurs (égos dans l'évaluation) à partir de leur réseau égocentrique dans DBLP (avec chacun des trois algorithmes à évaluer), et à rechercher celui qui prédit le mieux les centres d'intérêts exprimés explicitement par ces auteurs (égos) dans Mendeley. Le procédé de validation ici est alors assez similaire à la validation par confrontation à la perception humaine, même si les auteurs n'interviennent pas directement dans la validation. Les mesures d'évaluation considérées dans ce cas, sont la précision et le rappel en plus du cosinus de similarité.

Les résultats de ces évaluations nous ont permis de tirer quatre conclusions principales, parmi ces conclusions, deux ont été déduites de l'évaluation préliminaire sous Facebook, et confirmées dans l'évaluation à plus grande échelle sous DBLP/Mendeley :

- l'algorithme proposé basé sur les communautés fournit de meilleurs résultats que les algorithmes basés sur les individus,
- les scores de structure des communautés peuvent influencer positivement la qualité des dimensions sociales construites lorsque le paramètre $\alpha$ est faible (notamment dans $[0, 0.1]$).

Les deux autres conclusions sont tirées de l'évaluation sous DBLP/Mendeley qui disposait d'une taille d'échantillon plus importante (par rapport à l'évaluation sous Facebook) sur laquelle des analyses plus détaillées ont été réalisées :

- la qualité des dimensions sociales construites dépend de la densité des réseaux égocentriques des auteurs et du nombre de co-auteurs dans le réseau égocentrique. Les meilleurs résultats sont obtenus pour des auteurs ayant plus de 70 co-auteurs avec une densité du réseau égocentrique autour de 20%,
- les mesures de structure (degré, proximité, cohésion) produisent approximativement les mêmes résultats pour l'algorithme basé sur les communautés, toutefois la mesure de centralité de degré de communautés fournit légèrement de meilleurs résultats.

Au final, chacune de ces évaluations démontre que notre approche basée sur les communautés est plus pertinente que les approches basées sur les individus et confirme notre hypothèse de travail (pour des réseaux égocentriques, k=1). Les temps de calculs de chacune des approches sont sensiblement les mêmes, une fois que les communautés et leurs profils sont construits et sauvegardés pour chaque égo. Si notre approche produit de meilleurs profils sociaux d'utilisateurs, on peut donc logiquement penser qu'elle sera la meilleure approche à exploiter par tout mécanisme d'adaptation de l'information à l'utilisateur.

# 6 Conclusions générales et perspectives



## 6.1 Rappel du contexte et de la problématique

Le développement des profils utilisateurs (ou modélisation de l'utilisateur) dans un système d'information est un préalable et un enjeu majeur dans la mise en place de systèmes d'adaptation de l'information à l'utilisateur (personnalisation, recommandation, adaptation, etc.) et dans les systèmes d'analyses comportementales (détection des comportements à risques, détection de leaders, etc.). Les domaines d'application sont très nombreux : systèmes de recherche d'information tels que les moteurs de recherche (Google, Yahoo, etc.), systèmes de recommandation (Amazon, Ebay, etc.), hypermédias adaptatifs (environnements d'e-Learning, parcours personnalisés de sites Web, etc.), etc.

Le profil de l'utilisateur est généralement construit à partir de l'historique des activités de l'utilisateur, ce qui implique qu'il a besoin d'être constamment enrichi pour contenir le plus d'informations pertinentes pour les mécanismes (ou systèmes) qui s'y réfèrent. Ainsi, à tout moment, le profil d'un utilisateur peut ne pas contenir les informations nécessaires à un mécanisme, ce qui a pour effet de réduire l'efficacité de ce mécanisme dans la qualité de l'information qu'il proposera à l'utilisateur relativement aux besoins spécifiques de ce dernier. Ce problème est plus récurrent chez des nouveaux utilisateurs dans un système (profil vide) ou des utilisateurs très peu actifs (profil contenant très peu d'informations). Pour pallier ce problème d'enrichissement permanent du profil de l'utilisateur, outre l'activité propre de l'utilisateur dans un système donné, plusieurs travaux ont exploré d'autres sources de données potentiellement utiles pour dériver des informations pertinentes sur l'utilisateur. Il s'agit par exemple des informations en provenance des individus au comportement similaire à celui de l'utilisateur (systèmes de filtrage collaboratif) ou des informations produites par le même utilisateur dans d'autres systèmes d'information (systèmes de gestion des identités multiples ou profils utilisateurs multi-applications).

Ces dernières années, une nouvelle source de sources de données potentiellement très utile est exploitée dans les travaux nécessitant la modélisation de l'utilisateur : les réseaux sociaux. L'intérêt de plus en plus important pour l'usage des réseaux sociaux dans les systèmes nécessitant la modélisation de l'utilisateur a été principalement motivée par deux raisons : le fait que le comportement d'un individu dans la vie réelle est très souvent influencé par celui de son entourage, et la multiplication des réseaux sociaux en ligne (réseaux sociaux numériques) qui permettent d'avoir de plus en plus accès aux données du réseau social des utilisateurs. Les travaux utilisant les données du réseau social de l'utilisateur dans les systèmes nécessitant l'utilisation des profils utilisateurs entrent dans une nouvelle catégorie de systèmes d'adaptation de l'information à

l'utilisateur : les systèmes de filtrage social de l'information (systèmes de recommandation sociaux, systèmes de recherche d'information sociale, etc.). Toutefois, même si les systèmes de filtrage social actuels ont déjà démontré leur pertinence par rapport aux systèmes existants, nous avons noté qu'ils peuvent considérablement être améliorés (ou optimisés) par rapport à deux problématiques :

- ***L'amélioration (optimisation) de la qualité des profils construits :*** nous avons montré que les systèmes de filtrage social actuels exploitent d'une seule manière le réseau social de l'utilisateur : usage des individus ou d'une partie des individus directement liés à l'utilisateur, ainsi que de la force des relations entre ces individus et l'utilisateur. En se basant sur des travaux existants dans l'analyse des réseaux sociaux en sciences sociales, on peut envisager d'autres méthodes d'exploitation du réseau social de l'utilisateur pour dériver des éléments pertinents de son profil.

- ***La généricité de l'approche de modélisation sociale :*** nous remarquons que les travaux de la littérature dans le filtrage social ne séparent pas en général les phases de construction du profil de l'utilisateur en intégrant son réseau social, des phases d'usage des profils sociaux construits dans les mécanismes sous-jacents. Le réseau social de l'utilisateur est directement exploité dans les mécanismes de filtrage social. De plus, chaque proposition est fortement liée aux contraintes du domaine dans lequel elle s'applique (systèmes de recherche bibliographiques, systèmes de personnalisation ou de recommandation sur Internet, systèmes de personnalisation ou de recommandation dans des Intranets, etc.). On pourrait donc penser à proposer un modèle « social » de profil utilisateur qui soit réutilisable dans plusieurs domaines d'application et pour plusieurs types de mécanismes.

## 6.2 Résumé des contributions

Pour améliorer (optimiser) les systèmes d'adaptation basés sur le filtrage social de l'information selon les deux problématiques énoncées précédemment, nous situons notre contribution sur quatre plans.

Le premier plan est le positionnement de la problématique en elle-même. Jusqu'ici, tous les travaux de filtrage social de l'information se sont intéressés à démontrer l'intérêt de l'usage du réseau social de l'utilisateur pour l'amélioration des mécanismes d'adaptation de l'information à l'utilisateur. Chaque étude utilise à sa manière le réseau social de l'utilisateur. Aucune étude à notre connaissance ne s'est penchée sur la façon d'exploiter efficacement le réseau social de l'utilisateur dans les mécanismes de filtrage social de l'information. Cette problématique en elle-même est donc nouvelle.

Le deuxième plan est la proposition d'un modèle générique de profil social de l'utilisateur pour répondre à la problématique de généricité de profils utilisateurs exploités dans les mécanismes de filtrage social de l'information. Notre approche consiste, premièrement, à séparer dans les mécanismes de filtrage social, la phase de construction du profil social de l'utilisateur de la phase de son exploitation dans les mécanismes associés. Ainsi, un même modèle de profil social de l'utilisateur pourra être exploité par plusieurs mécanismes en fonction des besoins du mécanisme. Nous proposons un modèle générique de profil social d'un utilisateur

qu'on peut construire dans tout réseau social au sens premier du terme[55] quelque soit le mécanisme qui va l'utiliser. Pour ce faire, le modèle proposé est constitué de deux dimensions : une dimension utilisateur (qui sera construite à partir des activités de l'utilisateur) et une dimension sociale (qui sera construite à partir des activités des membres du réseau social de l'utilisateur). Ces deux dimensions sont caractérisées par les mêmes types d'attributs et sont ainsi comparables (c'est-à-dire exploitables indifféremment dans un mécanisme de filtrage social). De plus, nous considérons uniquement une portion significative du réseau social de l'utilisateur qui intègre les individus ou communautés d'individus pertinents vis-à-vis de l'utilisateur : son réseau k-égocentrique (réseau constitué des individus situés à une distance maximale k de l'utilisateur ainsi que leurs relations dans le réseau social global). Ce choix se justifie par des travaux existants dans les sciences sociales qui ont démontré l'intérêt du réseau égocentrique d'un utilisateur (égo) pour lui faciliter l'accès à de nouvelles informations et d'accroître ainsi son capital social. Le réseau k-égocentrique nous permet non seulement de considérer les individus déjà exploités dans les approches existantes de filtrage social, mais aussi de considérer les communautés d'individus autour de l'utilisateur qui, d'après notre hypothèse de travail permettront de dériver une dimension sociale du profil de l'utilisateur plus pertinente que lorsque cette dimension est dérivée à partir des individus (tel que le font les travaux existants).

Le troisième plan de notre contribution est la proposition d'un algorithme de dérivation de la dimension sociale du profil d'un utilisateur à partir des communautés de son réseau k-égocentrique. Cet algorithme se décompose en quatre étapes :

- La détection de communautés dans le réseau k-égocentrique de l'utilisateur : nous avons présenté les critères de choix de l'algorithme de détection de communautés qui serait le plus efficace dans notre contexte (recouvrement de communautés, fixation taille minimale de communautés, types de données de validation, techniques de construction et évaluation des communautés, prise en compte de la dynamique du réseau).

- Le profilage des communautés : par agrégation des dimensions utilisateur des profils des membres de la communauté. Le profil d'une communauté est représenté de la même manière que le profil d'un utilisateur (suivant une taxonomie de concepts).

- La caractérisation sémantico-structurelle des communautés : le profil de chaque communauté est caractérisé sémantiquement par rapport aux autres communautés (centres d'intérêts importants dans la communauté et peu important dans les autres communautés), et structurellement par rapport à sa centralité (centralité de degré, de proximité, d'intermédiarité, etc.) ou sa cohésion (importance du nombre de liens entre les membres de la communauté).

- La dérivation de la dimension sociale des communautés : les profils caractérisés des communautés sont combinés de manière optimale pour dériver les centres d'intérêts de la dimension sociale du profil de l'utilisateur. Nous avons justifié le choix de la technique de combinaison linéaire des profils de communautés en faisant une analogie avec les techniques de fusion de résultats de plusieurs moteurs de recherche en recherche d'information (pour déterminer le résultat renvoyé par un méta-système de recherche d'information intégrant plusieurs moteurs de recherche).

[55] La relation entre deux utilisateurs implique que ces derniers se connaissent dans la vie réelle

Le quatrième plan de notre contribution se décompose en deux parties. Premièrement, la proposition de différentes techniques d'évaluation des approches de dérivation de la dimension sociale du profil de l'utilisateur : évaluation directe dans les mécanismes de filtrage social, évaluation automatique par comparaison des dimensions sociales et de la dimension utilisateur pour des utilisateurs très actifs, évaluation par confrontation à la perception humaine dans laquelle les utilisateurs peuvent eux-mêmes valider la pertinence des profils sociaux construits. Ensuite, nous avons réalisé les expérimentations et évaluations proprement dites de notre approche par rapport aux principes des approches existantes dans deux contextes différents : réseaux sociaux numériques (cas de Facebook) et réseaux de co-auteurs d'articles scientifiques (cas de DBLP et Mendeley). Ces évaluations nous ont permis de démontrer de manière empirique la pertinence de notre approche par rapport aux approches existantes, et de montrer l'influence de paramètres tels que la mesure de centralité de communautés utilisée, la densité des réseaux égocentriques, ou le nombre d'utilisateurs dans le réseau égocentrique, sur la qualité de résultats.

## 6.3 Perspectives de recherche

La problématique abordée dans cette thèse étant relativement nouvelle dans le contexte des systèmes d'adaptation de l'information à l'utilisateur, plusieurs perspectives sont envisageables relativement :

- au modèle proposé,
- à l'algorithme proposé,
- aux expérimentations et évaluations,
- à l'ultime phase d'usage des dimensions utilisateur et sociales des profils construits dans les systèmes d'adaptation de l'information à l'utilisateur ou dans les systèmes d'analyses comportementales,
- à la sécurité des informations sur les utilisateurs dans les réseaux sociaux.

Le modèle proposé ici s'appuie sur une représentation sous forme de taxonomie de chacune des dimensions du profil utilisateur. Pour être encore plus générique, il pourra être judicieux d'opter pour une représentation sous forme d'ontologie afin de prendre en compte divers types de relations existantes dans les concepts (ou domaines d'intérêts) du profil d'un utilisateur.

L'algorithme basé sur les communautés du réseau k-égocentrique proposé s'appuie sur la structure du graphe induit par le réseau k-égocentrique à la différence des algorithmes existants qui eux s'appuient sur les individus du réseau égocentrique et éventuellement sur la force des liens entre ces individus et l'utilisateur. On peut envisager optimiser l'algorithme proposé dans le cas des réseaux k-égocentriques étendus (si on dispose de la force des liens entre l'égo et les membres de son réseau égocentrique) en incluant la force des liens entre les communautés et l'égo dans la dérivation de la dimension sociale. L'algorithme résultant serait alors un algorithme « mixte » qui exploiterait à la fois les avantages des communautés autour de l'utilisateur et la force des liens entre ces communautés et l'égo.

Les évaluations réalisées dans cette thèse n'ont considéré que le cas k=1 (réseau égocentrique). Il serait important de réaliser les mêmes évaluations avec des valeurs de k>1 (notamment k=2 par exemple) et

comparer les résultats obtenus avec ceux déjà observés. Bien évidemment les valeurs de k>1 augmenteraient considérablement la complexité de calcul compte tenu de la taille des graphes à analyser dans ce cas, mais si les résultats observés sont plus pertinents, on peut envisager des techniques de réduction de complexité en fonction des domaines d'application en sauvegardant par exemple les communautés et leurs profils pour chaque égo, et en différant périodiquement les mises à jour de la dimension sociale. De plus, d'autres mesures de structure non exploitées dans les expérimentations actuelles peuvent également être évaluées (centralité d'intermédiarité par exemple). D'autres mesures de caractérisation sémantique des communautés peuvent également être évaluées pour améliorer les résultats. La mesure tf.idf considérée par exemple dans ce document suit la logique de notre approche, mais serait inappropriée dans des réseaux k-égocentriques avec beaucoup de recouvrement, car un même centre d'intérêt aura de forte chance d'être présent dans toutes les communautés, et par conséquent sa valeur tf.idf vaudra 0 (cf. section 4.3.1.3.1). Pour pallier ce type de problème, des techniques telles que *tf.idf reduction* et LSA (*Latent Semantic Analysis*) peuvent également être exploitées et évaluées, de même que les techniques basées sur les machines à vecteurs de support (SVM). Enfin, les évaluations peuvent être réalisées dans plusieurs autres domaines : réseaux sociaux d'abonnés téléphoniques chez un opérateur de télécommunications, réseaux sociaux de clients de banques, réseaux d'échanges de mails dans une entreprise, etc. On peut également envisager dans le contexte du Web sémantique une évaluation dans laquelle plusieurs types de réseaux sociaux sont exploités pour extraire différents types de communautés autour d'un utilisateur (*LinkedIn* pour les communautés professionnelles, *Facebook* pour les communautés d'intérêts, etc.). Au-delà des réseaux sociaux, l'approche proposée peut également être étudiée pour une généralisation dans la prédiction de propriétés d'un nœud à partir de nœuds voisins dans tout type de graphe. Supposons par exemple la problématique de génération automatique de métadonnées caractéristiques d'un article scientifique. Ceci pourrait se faire de manière plus efficace suivant l'approche de réseaux k-égocentriques en exploitant les « communautés » d'articles cités par cet articles dans le réseau de co-citations d'articles.

Les évaluations réalisées dans cette thèse n'ont concerné que la phase de construction des profils afin de s'assurer dans un premier temps de la pertinence des profils construits, avant leurs usages par des mécanismes d'adaptation de l'information à l'utilisateur qui s'appuient sur son réseau social (recherche d'information sociale, systèmes de recommandations sociaux, etc.). Nous proposons ainsi un profil utilisateur à deux dimensions d'informations (dimension sociale et dimension utilisateur), et nous évaluons la pertinence de la dimension sociale construite. Pour aller plus loin dans les mécanismes d'adaptation de l'information, on peut s'imaginer plusieurs manières d'utiliser les deux dimensions proposés : par exemple, n'exploiter la dimension sociale que lorsque des informations ne sont pas présentes dans la dimension utilisateur, ou encore d'intégrer systématiquement les informations des deux dimensions dans le mécanisme de filtrage social. Ainsi, pour aller jusqu'au bout des mécanismes, ce serait intéressant de proposer différentes techniques d'usage des deux dimensions du profil en fonction des besoins dans les mécanismes de filtrage social de l'information, et d'évaluer les résultats renvoyés par ces mécanismes en fonction des techniques proposées.

Enfin, comme le démontrent nos résultats, il est possible de déduire des informations pertinentes des utilisateurs à partir de leur réseau social. Vue sous un angle de sécurisation des informations utilisateurs, ceci soulève une problématique de sécurisation non seulement des données utilisateurs, mais aussi des données du réseau social de l'utilisateur. Ceci est particulièrement d'actualité dans le cas des réseaux sociaux numériques (exemple de Facebook) dans lesquels l'utilisateur peut définir des paramètres de sécurité sur ses données, mais pas toujours sur la liste de ses contacts et sur les données en provenance de ses contacts.

# Liste des tableaux

# Liste des figures

# Bibliographie

## A

**ABBAR S., BOUZEGHOUB M., KOSTADINOV D., LOPES S.,** *A contextualization service for a Personalized Access Model.* 9èmes Journées Francophones Extraction et Gestion des Connaissances, EGC 2009: 265-270, 2009.

**ABBAR S., BOUZEGHOUB M., KOSTADINOV D., LOPES S., AGHASARYAN A., BETGÉ-BREZETZ S.,** *A personalized access model: concepts and services for content delivery platforms.* Information Integration and Web-based Applications & Services: iiWAS 2008, pp. 41—47, 2008.

**ADAMS P.,** *Real Life Social Networks vs Online*, Google, 2010.

**ADOMAVICIUS G., SANKARANARAYANAN R., SEN S., TUZHILIN A.,** *Incorporating contextual information in recommender systems using a multidimensional approach.* ACM Transactions on Information Systems (TOIS), 23, 103– 145, 2005.

**ADOMAVICIUS G., TUZHILIN A.** *Multidimensional recommender systems: a data warehousing approach.* Lecture Notes in Computer Science: Proceedings of the Second International Workshop on Electronic Commerce. 2232, 2001.

**ALIM S., ABDUL-RAHMAN R., NEAGU D., MICK R.,** *Data Retrieval from Online Social Network Engineering Applications,* In Technology and Secured Transactions, ICITST 2009, p. 1-5, 2009.

**AMATO G., UMBERTO S.,** *User Profile Modeling and Applications to Digital Libraries* . In ECDL 1999, European Conference on Research and Advanced Technology for Digital Libraries: pp184—197 (1999).

**AMATRIAN X.,** *Systèmes de recommandation,* Actes de l'Ecole d'été Web Intelligence « le Web centré sur l'utilisateur » du 5 au 9 juillet 2010, Saint-Germain-Au-Mont-d'Or, France, 2010.

**AMITAY E., CARMEL D., HAR'EL N., OFEK-KOIFMAN S., SOFFER A., YOGEV S., GOLBANDI N.,** *Social search and discovery using a unified approach.* In WWW 2009, World Wide Web Conference: pp. 1211--1212 (2009).

**ANCONA, D.,** *Outward bound: Strategies for team survival in the organization.* Academy of Management Journal 33: 334-365, 1990.

**ANDERSON J., SCHOOLER L.,** *Reflections of the environment in memory.* Psychological Science 2(6), 396–408 (1991).

**ANGAL R., KALER C., VAN GONG H. L., MALER E., MEDVINSKY A., SHEWCHUK J.,** *Web Single Sign On Metadata Exchange Protocol,* Microsoft, Sun Microsystems, URL= http://xml.coverpages.org/WebSSO200505.pdf, April 2005.

## B

**BAATARJAV, E. A., DANTU R., YAN TANG, CANGUSSU J.,** *BBN-Based Privacy Management Sytem for Facebook*, Intelligence and Security Informatics, 2009. ISI '09. IEEE International Conference Page(s):194 – 196, 2009.

**BALABANOVIĆ M., SHOHAM Y.,** Fab: *content-based, collaborative recommendation,* Communications of the ACM, 3 (40), 66-72, 1997.

**BEN JABEUR L., TAMINE L., BOUGHANEM M.,** *A social model for Literature Access: Towards a weighted social network of authors*. In: RIAO'10: Proceedings of the 9th international conference on Information Retrieval and its Applications, CDROM, 2010.

**BENDER M., CRECELIUS T., KACIMI, MICHEL M. 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.

**BERNERS-LEE T., HENDLER J., LASSILA O.,** *The semantic web*. Scientific American, 284, 28–37, 2001.

**BILLSUS D., PAZZANI M. J.,** *A hybrid user model for news story classification*. Proceedings of the seventh international conference on User modeling table of contents, 99–108, 1999.

**BLOEDORN E., MANI I., MACMILLAN T. R.,** *Machine Learning of User Profiles: Representational Issues*, In AAAI'96 Proceedings of the thirteenth national conference on Artificial intelligence - Volume 1, pages Pages 433-438, 1996.

**BLONDEL V. D., GUILLAUME J., LAMBIOTTE R., LEFEBVRE E.,** *Fast unfolding of communities in large networks*, Journal of Statistical Mechanics : Theory and Experiment, vol.10, 2008.

**BORGATTI S. P., CANDACE J., EVERETT M. G.,** *Network Measures of Social Capital*, INSNA, Volume 21, Issue 2, 27-36, 1998.

**BONHARD P., SASSE M. A.,** *Knowing Me, Knowing You – Using Profiles and Social Networking to Improve Recommender Systems*, in BT Technology Journal, Vol. 25 - No. 3, July 2006.

**BONNEAU J., ANDERSON J., DANEZIS G.,** *Prying data out of a Social Network*, In Advances in Social Network Analysis and Mining, ASONAM 2009, p. 249-254.

**BOTTRAUD J. C., BISSON G., BRUANDET M. F.,** *Expansion de requêtes par apprentissage automatique dans un assistant pour la recherche d'information.* COnférence en Recherche d'Information et Applications, CORIA 2004: 89-108, 2004.

**BOUADJENEK MOHAMED 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.

**BOUCHRA S.,** *Vers un système d'information Web fournissant des services Web sensibles au contexte*. Thèse de doctorat, Université Paul Sabatier, avril 2010.

**BOUGHANEM M.,** Recherche d'information contextuelle*, Actes de l'Ecole d'été Web Intelligence 2010 « le Web centré sur l'utilisateur » du 5 au 9 juillet 2010, Saint-Germain-Au-Mont-d'Or, France.*

**BOURDIEU P.,** *The forms of social capital*. in J.G. Richardson (ed.) Handbook of theory and research for the sociology of education. New York: Greenwood Press, Pp. 241-258, 1986.

**BOUZEGHOUB M., KOSTADINOV D.,** *Personnalisation de l'information: aperçu de l'état de l'art et definition d'un modèle flexible de profils*. COnférence en Recherche d'Infomations et Applications, CORIA 2005, 201—218 (2005).

**BOYD D., ELLISON N.,** *Social Network Sites: Definition, History, and Scholarship*, Journal of Computer-Mediated Communication, vol. 13, 2007.

**BRASS D.,** *Power in organizations: A social network perspective.* In G. Moore and J.A. White (Eds.) *Research in politics and society*. Pp. 295-323. Greenwich: JAI Press, 1992.

**BRESLIN J., DECKER S.,** *The Future of Social Networks on the Internet,* Published by the IEEE Computer Society, December 2007.

**BRIN S., PAGE L.,** *The anatomy of a large-scale hypertextual Web search engine*. Computer Networks and ISDN Systems v.30, p.107, 1998.

**BRINI A., BOUGHANEM M., DUBOIS D.,** *A Model for Information Retrieval Based on Possibilistic Networks.* String Processing and Information Retrieval (SPIRE 2005), Buenos Aires, ARGENTINE, LNCS, Springer Verlag, p. 271-282, janvier 2005.

**BRUSILOVSKY P.,** *Methods and techniques of adaptive hypermedia.* User Modeling and User Adapted Interaction, 6(2-3) :87–129, 1996.

**BRUSILOVSKY P.,** Adaptive *hypermedia*. User Modeling and User Adapted Interaction, 11(1-2):87–110, 2001.

**BRUT M., SEDES F.,** *Modélisation basée sur ontologies pour développer des recommandations personnalisées dans les systèmes hypermédia adaptatives*. 28ème congrès INFORSID (INFormatique des ORganisations et Systèmes d'Information et de Décision) INFORSID 2010: 61-76, 2010.

**BUCHHOLZ S., HAMANN T., HÜBSCH G.,** *Comprehensive Structured Context Profiles (CSCP): Design and Experiences,* in Proc. of the Second IEEE Annual Conference on Pervasive Computing and Communications Workshops, Orlando, Florida, 2004.

**BURT, R.S.,** *Range,* Beverly Hills: Sage Publications, in R.S. Burt and M.J. Minor (Eds.), *Applied Network Analysis,* Pp. 176-194, 1983.

**BURT, R.S.,** *Structural holes.* Cambridge: Cambridge University Press, 1992.

## C

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

**CABANAC G., CHEVALIER M., CHRISMENT C., JULIEN C.,** *Collective Annotation: Perspectives for Information Retrieval Improvement.* In : Large-Scale Semantic Access to Content (Text, Image, Video and Sound) (RIAO 2007), Pittsburgh, USA, 30/05/2007-01/06/2007, Centre de hautes études internationales d'Informatique Documentaire (C.I.D.), p. 529-548, mai 2007.

**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.,** *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, 2009.

**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.

**CHAFFEE J., GAUCH S.,** *Personal ontologies for web navigation,* Proceedings of the ninth international conference on Information and knowledge management, CIKM '00, Pages 227 – 234, 2000.

**CHEDRAWY Z., ABIDI S. S. R.,** *Case based reasoning for information personalization: using a context-sensitive compositional case adaptation approach*. 2006 IEEE International Conference on Engineering of Intelligent Systems, (2006).

**CHEN C. M., DUH L. J.,** *Personalized web-based tutoring system based on fuzzy item response theory*. Expert Systems With Applications, 34, 2298–2315, (2008).

**CHEN C. C., CHEN M. C., SUN Y.,** *PVA: A Self-Adaptive Personal View Agent,* Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, *KDD '01, Pages 257 – 262, 2001.*

**CHEN L., SYCARA K.,** WebMate: *A Personal Agent for Browsing and Searching,* In Proceedings of the Second International Conference on Autonomous Agents, *132—139, ACM press, 1998.*

**CHO Y. H., KIM J. K., KIM, S. H.,** *A personalized recommender system based on web usage mining and decision tree induction*. Expert Systems With Applications, 23, 329–342, (2002).

**CLAYPOOL M., LE P., WASEDA M., BROWN D.,** *Implicit interest indicators*. In International Conference on Intelligent User Interfaces, IUI 2001, pp. 33--40 (2001).

**CLAYPOOL M., GOKHALE, A., MIRANDA, T., MURNIKOV, P., NETES, D., SARTIN, M.,** *Combining content-based and collaborative filters in an online newspaper*. ACM SIGIR Workshop on Recommender Systems, (1999).

**CONLAN O., WADE V., BRUEN C., GARGAN M.,** *Multi-model, metadata driven approach to adaptive hypermedia services for personalized elearning*. Proceedings of the Second International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, 100– 111, (2002).

**CONKLIN J., BEGEMAN M. L.,** *GIBIS A hypertext tool for team design deliberation*. In Hypertext, pages 247–251. ACM, 1987.

**COUTANT A., STENGER T.,** *Processus identitaire et ordre de l'interaction sur les réseaux socionumériques*, Les Enjeux de l'Information et de la Communication, En cours de publication, août 2010 [en ligne] http://w3.u-grenoble3.fr/les_enjeux/2010/Coutant-Stenger/index.html

## D

**DAOUD M., TAMINE L., BOUGHANEM M.,** *Towards a graph based user profile modeling for a session-based personalized search.* Knowledge and Information Systems, Springer, Vol. 21 N. 3, p. 365-398, juillet 2009.

**DAS A. S., DATAR M., GARG A., RAJARAM, S.,** *Google news personalization: scalable online collaborative filtering*. Proceedings of the 16th international conference on World Wide Web, 271–280, (2007).

**DE LUCA E. W., PLUMBAUM T., KUNEGIS J., ALBAYRAK S.,** *Multilingual Ontology-based User Profile Enrichment*, In WWW'2010, International Conference on World Wide Web, pp. 41--42 (2010).

**DEY A. K.,** *Providing Architectural Support for Building Context-Aware Applications*, PhD Thesis, Georgia Institute of Technology (2000).

**DIMATRACOUPOULOU A., BRUILLARD E.,** *Enrichir les interfaces de forums par la visualisation d'analyses automatiques des interactions et du contenu*, Article de Recherche, Revue Sticef, Volume 13, 2006.

**DOUSSET B.,** *Tétralogie: a platform for scientific and technological survey*. International workschop on Webometrics, Infometrics and Scientometrics & seventh COLLNET Meeting, LORIA, Nancy France, 10/05/2006-12/05/2006.

**DROMZEE C., LABORIE S., ROOSE P.,** *Profil générique sémantique pour l'adaptation de documents multimédias*. INFormatique des ORganisations et Systèmes d'Information et de Décision (INFORSID 2012), 191-206, 2012.

**DWYER C., HILTZ STARR R. WIDMEYER G.,** *Understanding Development and Usage of Social Networking Sites: The Social Software Performance Model*, Proceedings of the 41st Hawaii International Conference on System Sciences – 2008.

## E

**EATON A.,** *RVW module for syndicating and aggregating reviews*. http://www.pmbrowser.info/rvw/0.2, 2004.

**ENCELLE B.,** *Accessibilité aux documents électroniques: personnalisation de la présentation et de l'interaction avec l'information.* Thèse de doctorat, Université Paul Sabatier, décembre 2005.

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

## F

**FAYYAD U. M., PIATETSKY-SHAPIRO G., SMYTH P.,** The *KDD Process for Extracting Useful Knowledge from Volumes of Data*. Commun. ACM 39(11): 27-34 (1996).

**FELT A., EVANS D.,** *"Privacy Protection for Social Networking APIs"*, 2008.

**FORTUNATO S.,** *Community detection in graphs*, Physics Reports, vol. 486, Feb. 2010, p. 75-174.

**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.

**FREEMAN, L.C.,** *Centrality in social networks: I. Conceptual clarification*. Social Networks 1: 215-239, 1979.

**FREITAG D., ARMSTRONG R., JOACHIMS T., MITCHELL T.,** *WebWatcher: A Learning Apprentice for the World Wide Web*, in Proceedings of the 1995 AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments, March 1995.

**FRIAS-MARTINEZ E., MAGOULAS G., CHEN S., MACREDIE R.,** *Automated user modeling for personalized digital*, International Journal of Information Management, 26, 234–248.

**FRIGGERI A., CHELIUS G., FLEURY E.,** *Triangles to Capture Social Cohesion*. The Third IEEE International Conference on Social Computing, SocialCom/PASSAT 2011: 258-265.

**FUKUYAMA F.,** *Trust: The social virtues and the creation of prosperity*. New York: Free Press, 1995.

G

**GAO M., LIU K., WU Z.,** *Personalisation in web computing and informatics: Theories, techniques, applications, and future research*. Information Systems Frontiers 12(5): pp. 607--629 (2010).

**GAUCH S., CHAFFEE J., PRETSCHNER A.,** *Ontology*-based personalized search and browsing, *Web Intelligence and Agent Systems, Vol. 1, 1—3, 2003.*

**GAUME B.,** *Balades aléatoires dans les petits mondes lexicaux*, I3 Information Interaction Intelligence, vol. 4, n° 2, 2004.

**GENTILI G., MICARELLI A., SCIARRONE F.,** *Infoweb: An Adaptive Information Filtering System for the Cultural Heritage Domain*, Applied Artificial Intelligence, vol. 17, no. 8-9, 2003 , Pages: 715-744 , DOI: 10.1080/713827256

**GERVASONI J.,** *compte rendu du stage sur les réseaux sociaux*, IUFM d'Aix-Marseille, Janvier 2012, URL=http://www.ac-aix-marseille.fr/pedagogie//jcms/c_111539/fr/c/r-stage-reseaux-sociaux

**GIACOMO D., MAHONEY M., BOLLEN D., MONROY-HERNANDEZ J. A., MERAZ R. C. M., MYLIBRARY,** *A personalization service for digital library environments*. Proceedings of the Second DELOS Network of Excellence Workshop on Personalisation and Recommender Systems in Digital Libraries (ERCIM Workshop Proceedings 01/W03). Dublin, (2001).

**GIRVAN M., NEWMAN M. E. J.,** *Community structure in social and biological net works*, Proceedings of the National Academy of Sciences of the United States of America, vol. 99, n 12, 2002, p. 7821–7826.

**GODOY D.,** *Learning user interests for user profiling in personal information agents*: Thesis, AI Communications, Volume 19 Issue 4, December 2006, Pages 391 – 394, IOS Press Amsterdam.

**GOECKS J., SHAVLIK J.,** *Learning User's Interests by Unobtrusively Observing their Normal Behavior*, Proceedings of the 2000 International Conference on Intelligent User Interfaces, New-Orleans, USA, Jan 9-12, 2000, ACM, pp129-132, 2000.

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

**GOLBECK J., ROTHSTEIN M.,** *Linking Social Networks on the Web with FOAF: A Semantic Web Case Study*. In Proceedings of the Twenty third National Conference on Artificial Intelligence, AAAI 2008: 1138-1143.

**GOLBECK J., PARSIA B., HENDLER J. A.,** *Trust Networks on the Semantic Web*. In WWW'2003, International Conference on World Wide Web (Posters).

**GOLBECK J., HENDLER J. A.,** *Inferring binary trust relationships in Web-based social networks*. ACM Trans. Internet Techn. 6(4): 497-529 (2006).

**GOLEMATI M., KATIFORI A., VASSILAKIS C., LEPOURAS G., HALATSIS C.,** *Creating an Ontology for the User Profile: Method and Applicat*ions. In RCIS 2007, Research Challenges in Information Science: pp. 407-- 412 (2007).

**GOMEZ C., JUAN D., BOTHOREL C., POULET F.,** *Entropy based community detection in augmented social networks,* International Conference on Computational Aspects of Social Networks, Salamanca : 19-21 october 2011, Salamanca, Spain, 2011, pp. 163-168.

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

**GUARINO N., VETERE G., MASOLO C**., *OntoSeek: Content-Based Access to the Web,* IEEE Intelligent System, 14 (3), 70-80, 1999.

**GUHA S., TANG K., FRANCIS P**., *NOYB: Privacy in Online Social Networks,* WOSN'08, Seattle, Washington, USA, August 18, 2008.

**GUY I., JACOVI M., SHAHAR E., MESHULAM N., SOROKA V., FARRELL S.,** *Harvesting with SONAR: the value of aggregating social network information*. In CHI 2008, Computer Human Interaction: pp. 1017-1026 (2008).

**GUY I., JACOVI M., MESHULAM N., RONEN I., SHAHAR E.,** *Public vs. private: comparing public social network information with email.* 2008 ACM International Conference on Computer Supported Cooperative Work, CSCW 2008: 393-402, 2008.

**GUY I., RONEN I., WILCOX E.,** *Do you know? recommending people to invite into your social network*. 2009 International Conference on *Intelligent User Interfaces,* IUI 2009: 77-86, 2009.

**GUY I., ZWERDLING N., RONEN I., CARMEL D., UZIEL E**., *Social media recommendation based on people and tags.* 2010 Special Interests in Information Retrieval SIGIR 2010: 194-201, 2010.

**GUY I., ZWERDLING N., CARMEL D., RONEN I., UZIEL E., YOGEV S., OFEK-KOIFMAN S.,** *Personalized recommendation of social software items based on social relations.* In *Proceedings of the 2009 ACM conference on Recommender systems,* RecSys 2009: 53-60, 2009.

## H

**HADDADI E. A., DOUSSET B., BERRADA L., KASSOU I.,** *Construction d'une ontologie de domaine fondée sur le "text mining"*. Colloque Veille Stratégique Scientifique et Technologique (VSST 2009), Nancy, 30/03/2009-31/03/2009, Bernard Dousset (Eds.), Université Paul Sabatier - Toulouse, (support électronique), mars 2009.

**HARARY F.,** *Graph theory*. New York: Addison-Wesley, 1969.

**HECKMANN D., KRUEGER** A., *A User Modeling Markup Language (UserML) for Ubiquitous Computing,* Lecture Notes in Computer Science, Volume 2702/2003, 148, 2003.

**HERLOCKER J. L., KONSTAN J.A., BORCHERS A., RIEDL J.,** *An Algorithmic Framework for Performing Collaborative Filtering*. 1999 Special Interests Group on Information Retrieval, SIGIR 1999: 230-237.

**HOLLAND S., WERNER K.,** *Situated Preferences and Preference Repositories for Personnalized Database Applications*, 23rd International Conference on Conceptual Modeling, ER 2004, Shanghai, China, (2004).

**HOFMANN T.,** *Latent semantic models for collaborative filtering*. ACM Transactions on Information Systems (TOIS), 22, 89– 115.

**HOFMANN T.,** *Probabilistic latent semantic analysis*. matrix, 50, 2, 1999.

**HORVITZ E., BREESE J., HECKERMAN D., HOVEL D., ROMMELSE K.,** *The Lumiere project: Bayesian user modeling for inferring the goals and needs of software users*. Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, 256–265, (1998).

**HUBERT G., LOISEAU Y., 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, 2007.

**HUHNS M. N., STEVENS L. M**., *Personal Ontologies*, IEEE Internet Computing, Volume 3, Issue 5, 1999, pages 85-87.

## I

**ISOZAKI H., KAZAWA H.,** *Efficient support vector classifiers for named entity recognition*. Proceedings of the 19th international conference on Computational linguistics-Volume 1, 1–7, (2002).

## J

**JIN X., ZHOU Y., MOBASHER, B.,** *Web usage mining based on probabilistic latent semantic analysis*. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 197–205, 2004.

**JIN Y., MATSUO Y., ISHIZUKA M.,** *Extracting a Social Network among Entities by Web mining*. In European Semantic Web Conference (ESWC 2007), Innsbruck, Austria, 2007.

**JOACHIMS T.,** *A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization*. Proceedings of the Fourteenth International Conference on Machine Learning, 143–151, (1997).

**JUNG K.,** *Modeling Web user interest with implicit indicators*, master thesis, Florida Institute of Technology, 2001.

**JUNG S. Y., HONG J. H., KIM, T. S.,** *A statistical model for user preference*. IEEE Transactions on Knowledge and Data Engineering, 834–843, (2005).

## K

**KAUTZ H., SELMAN B., SHAH M.,** *Referral Web: combining social networks and collaborative filtering*. *Commun.* ACM *40, 3 (March 1997).*

**KELLY, D.,** *Understanding implicit feedback and document preference: A naturalistic user study*. Ph.D. Dissertation, Rutgers University, (2004).

**KELLY D., TEEVAN J.,** *Implicit Feedback for inferring user preferences: a bibliography*, SIGIR Forum, Number 37, Vol. 2, 18—28, 2003.

**KLYNE G., REYNOLDS F., WOODROW C., HIDETAKA O., HJELM J., BUTLER M. H., TRAN L.,** *Composite Capability/Preference Profiles (CC/PP): Structure and Vocabularies 1.0.* URL=http://www.w3.org/Mobile/CCPP/, 2004.

**KIEßLING W.,** *Foundations of preferences in database systems*, VLDB '02 Proceedings of the 28th international conference on Very Large Data Bases, pp. 61--82 (2002).

**KIM H., CHAN P.,** *Implicit Indicators for Interesting Web Pages*. WEBIST 2005, Proceedings of the First International Conference on Web Information Systems and Technologies, Miami, USA, May 26-28, 2005.

**KIM H., CHAN P.,** *Learning implicit user interest hierarchy for context in personalization*, IUI '03 Proceedings of the 8th international conference on Intelligent user interfaces, 101-108, 2003.

**KIM J. K., CHO Y. H., KIM W. J., KIM J. R., SUH J. H.,** *A personalized recommendation procedure for Internet shopping support*. Electronic Commerce Research and Applications, 1, 301–313. (2002).

**KIM J. W.,** *Application of decision-tree induction techniques to personalized advertisements on internet storefronts*. International Journal of Electronic Commerce, 5, 45–62, (2001).

**KITTS B., FREED D., VRIEZE M.,** *Cross-sell: a fast promotion tunable customer-item recommendation method based on conditionally independent probabilities*. Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, 437–446, (2000).

**KNIGHT K., LUK S. K.,** *Building a large-scale knowledge base for machine translation*, AAAI '94 Proceedings of the twelfth national conference on Artificial intelligence (vol. 1), Pages 773-778, 1994.

**KOBSA A.,** *Generic User Modeling Systems*, User Modeling and User Adapted International Journal, Vol. 11, 49-63, 2001.

**KONSTAN J. A., MILLER B. N., MALTZ D., HERLOCKER J. L., GORDON L. R., RIEDL J.**, *GroupLens: applying collaborative filtering to Usenet news*. Communications of the ACM, 40, 77–87. (1997).

**KONSTAS I., STATHOPOULOS V., JOSE J. M.,** *On social networks and collaborative recommendation*. 2009 Special Interests Group on Information Retrieval, SIGIR 2009: 195-202.

**KOSTADINOV D.,** *La personnalisation de l'information, une approche de gestion des profils et de reformulation de requêtes*, Thèse de Doctorat, Université de Versailles Saint-Quentin-En-Yveline, décembre 2007.

## L

**LABORIE S., MANZAT A-M., FLORENCE SÈDES.,** *Managing and querying efficiently distributed semantic multimedia metadata collections.* IEEE MultiMedia, IEEE Computer Society, Los Alamitos - USA, Numéro spécial Special issue on Multimedia-Metadata and Semantic Management, Vol. 16 N. 4, p. 12-20, décembre 2009.

**LABROU Y., FININ T.,** *Yahoo! as an ontology: using Yahoo! categories to describe documents*, Proceedings of the eighth international conference on Information and knowledge management, CIKM '99, Pages 180 – 187, 1999.

**LAI H. J., LIANG T. P., KU Y. C.,** *Customized Internet news services based on customer profiles.* Proceedings of the 5th international conference on Electronic commerce, 225–229, (2003).

**LE P., WASEDA M.,** *A curious browser: implicit rating,* Technical reports, 2000.

**LEMIEUX *V.*, OUIMET *M.*,** *L'analyse structurale des réseaux sociaux,* Methodes En Sciences Humaines, De Boeck, mai 2004.

**LEMLOUNA T., LAYAIDA N.,** *Content Adaptation and Generation Principles for Heteregeneous Clients*, W3C Workshop on Device Independent Authoring Techniques, Gernany, 2002.

**LEY M.,** *DBLP - Some Lessons Learned*. PVLDB 2(2): 1493-1500 (2009).

**LEY M.,** *DBLP XML Requests*, Appendix to the paper DBLP — Some Lessons Learned (June 17, 2009).

**LI Y., LU L., XUEFENG L.,** *A hybrid collaborative filtering method for multiple-interests and multiple-content recommendation in E-Commerce*. Expert Systems With Appli cations, 28, 67–77, (2005).

**LIANG T.-P., LAI H.-J.,** *Discovering user interests from Web browsing behavior: an application to Internet news services*. System Sciences, 2002. HICSS. Proceedings of the 35th Annual Hawaii International Conference on System Sciences, (2002).

**LIEBERMAN H.,** *Letizia : an agent that assists web browsing,* IJCAI'95 Proceedings of the 14th international joint conference on Artificial intelligence - Volume 1, Pages 924-929, 1995.

**LIN C., XUE G-R., ZENG H-J., YU Y.,** *Using Probabilistic Latent Semantic Analysis for Personalized Web Search*, Proceedings of APWeb. 2005, 707-717, Springer-Verlag.

**LIN N.,** *Conceptualizing social support*. In N. Lin, A. Dean, and W. Ensel (Eds.) Social support, life events and depression. New York: Academic Press, 1986.

**LINDEN, G., SMITH B., YORK J.,** *Amazon.com recommendations: item-to-item collaborative filtering*. IEEE Internet computing, 7, 76–80, (2003).

M

**MALEK M.,** *Introduction à l'analyse des réseaux sociaux*, Rapport EISTI-LARIS, 2009, http://mma.perso.eisti.fr/HTML-IAD/RapportR1.pdf.

**MARSDEN P.V.,** *Homogeneity in confiding relations*. Social Networks 10: 57-76, 1988.

**MASSA P., AVESANI P.,** *Trust-Aware Collaborative Filtering for Recommender Systems*. CoopIS/DOA/ODBASE (1) 492-508, 2004.

**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, 2007.

**MASSA P., AVESANI P.,** *Trust Metrics in Recommender Systems*. Computing with Social Trust, Springer, 259-285, 2009.

**MATSUO Y., HAMASAKI M., TAKEDA H., NISHIMURA T., HASIDA K., ISHIZUKA M.,** *POLYPHONET: An advanced social network extraction system*. In proceedings Word Wide Web Conference, 2006.

**MCNEE S.M., ALBERT I., COSLEY D., GOPALKRISHNAN P., LAM S.K., RASHID A.M., KONSTAN J.A., RIEDL J.,** *On the recommending of citations for research papers*. In: CSCW'02: Proceedings of the 2002 ACM conference on Computer supported cooperative work, ACM, New York, NY, USA, pp 116–125, (2002).

**MELVILLE P., MOONEY R. J., NAGARAJAN R.,** *Content-Boosted collaborative filtering for improved recommendations*. Proceedings of the Eighteenth National Conference on Artificial Intelligence, 187– 192, (2002).

**MICARELLI A., SCIARRONE F.,** *Anatomy and Empirical Evaluation of an Adaptive Web-Based Information Filtering System,* User Modeling and User Adapted interaction 14 (2-3): 159-200, 2004.

**MIKA P.,** *Flink: Semantic Web Technology for the Extraction and Analysis of Social Networks*. Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 3, No. 2-3., 2005, p. 211-223.

**MOBASHER B.,** *Data mining for web personalization*. The Adaptive Web: Methods and Strategies of Web Personalization, Brusilovsky, P., Berlin Heidelberg New York: Springer Verlag, (2007).

**MOCK K. J., VEMURI V. R.,** *Information filtering via hill climbing, wordnet, and index patterns*. Information Processing & Management, Vol 33, N° 5, 633-644 (1997).

**MONTANER M., LOPEZ B., DELA ROSA J. L.,** *A taxonomy of recommender agents on the Internet*. Artificial Intelligence Review, 19, 285–330, (2003).

**MONTEBELLO M., GRAY W. A., HURLEY S.,** *A Personal Evolvable Advisor for WWW Knowledge-Based Systems. Proceedings of Workshop on Reuse of Web information at the Seventh International World Wide Web Conference*, Brisbane, Australia, Available online at http://www.mel.dit.csiro.au/~vercous/REUSE/pos7/index.html, 1997.

**MOONEY R. J., ROY L.,** *Content-based book recommending using learning for text categorization*. Proceedings of the fifth ACM conference on Digital libraries, 195–204, (2000).

N

**NANOPOULOS A., KATSAROS D., MANOLOPOULOS Y.,** *Effective prediction of web-user accesses: A data mining approach*. WEBKDD, 1, 2001.

**NAVARRO E., CAZABET R.,** *Détection de communautés, étude comparative sur graphes réels*, In 1er Journées Modèles et l'Analyse des Réseaux : Approches Mathématiques et Informatique (MARAMI 2010), Toulouse, 2010.

**NICHOLS D.,** *Implicit Rating and Filtering*. In: Proceedings of the Fifth DELOS Workshop on Filtering and Collaborative Filtering. ERCIM, Budapest, pp. 31-36. (1998).

## O

**OARD D. W., KIM J.,** *Modeling Information Content Using Observable Behavior*, In proceedings of the 64th annual meeting of the American Society for Information Science and Technology, 38—45, 2001.

**O'DONOVAN J., SMYTH B., EVRIM V., MCLEOD D.,** *Extracting and Visualizing Trust Relationships from Online Auction Feedback Comments*. In 2007 International Joint Conference on Artificial Intelligence IJCAI 2007: 2826-2831.

**O'MAHONY M. P., HURLEY NEIL J., GUENOLE C. M. S.,** *Collaborative Filtering - Safe and Sound?* ISMIS 2003: 506-510

**O'MAHONY M. P., HURLEY NEIL J., GUENOLE C. M. S.,** *Recommender systems: Attack types and strategies*. In Proceedings of the 20th National Conference on Artificial Intelligence (AAAI-05), Pittsburgh, Pennsylvania, USA, 9–13, Jul 2005. AAAI Press.

## P

**PALLA G., DERENYI I., FARKAS I., VICSEK T.,** *Uncovering the overlapping community structure of complex networks in nature and society*, Nature, vol. 435, Juin. 2005, p. 814-818.

**PATSAKIS C., ASTHENIDIS A., CHATZIDIMITRIOU A.,** *Social networks as an attack platform: Facebook case study*. Networks, 2009. ICN, p. 245-247.

**PAZZANI M., MURAMATSU J., BILLSUS D.,** *Syskill Webert: Identifying interesting web sites,* In Proceedings of the Thirteenth National Conference on Artificial Intelligence, 1996, 54—61, AAAI Press.

**PITKOW J. E., SCHÜTZE H., CASS TODD A., COOLEY R., TURNBULL D., EDMONDS A., ADAR E., BREUEL T. M.,** *Personalized search*. Commun. ACM 45(9): 50-55 (2002).

**PONS P.,** *Détection de communautés dans les grands graphes de terrain,* Thèse de doctorat, Université Paris 7 - Denis Diderot, 2007.

**PORCEL C., LÓPEZ-HERRERA A.G., HERRERA-VIEDMA E.,** *A recommender system for research resources based on fuzzy linguistic modeling*. Expert Syst Appl 36(3):5173–5183, (2009).

**PRESTCHNER A.,** *Ontology Based Personalized search*, Master's thesis, University of Kansas, June 1999.

**PUTNAM R.,** *Bowling alone: America's declining social capital*. Journal of Democracy 6(1):65-78, 1995.

## Q

**QUIROGA L. M., MOSTAFA J.,** *Empirical evaluation of explicit versus implicit acquisition of user profiles in information filtering systems, DL '99* Proceedings of the fourth ACM conference on Digital libraries, Pages 238 – 239, 1999.

## R

**RAZMERITA L.,** *Modèle Utilisateur et Modélisation Utilisateur dans les systèmes de Gestion des connaissances : une approche fondée sur les ontologies.* Thèse de doctorat, Université Paul Sabatier, décembre 2003.

**REFFAY C., LANCIERI L.,** *Quand l'analyse quantitative fait parler les forums de discussion,* Article de Recherche, Revue Sticef, Volume 13, 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.

**ROBERTSON S., SPARCK J. K.,** *Relevance weigthing for search terms*. Journal Of the American Society for Information Science, 27 (3), 129-146, 1976.

**ROSVALL M., BERGSTROM C.T.,** *An information-theoretic framework for resolving community structure in complex networks,* PNAS, vol. 104, p. 7327-7331, Mai. 2007.

**RUTHVEN I., LALMAS M.,** *A survey on the use of relevance feedback for information access systems,* Journal The Knowledge Engineering Review Volume 18 Issue 2, Pages 95 – 145, June 2003.

**RUVINI J. D.,** *Adapting to the user's internet search strategy.* Proceedings of the 9th International Conference on User Modeling (UM2003), Pittsburgh, 55–64, (2003).

## S

**SAAD MISSEN M. M., BOUGHANEM M., CABANAC G.,** *Opinion Detection in Blogs: What is still Missing* ASONAM'10: Proceedings of the 2nd International Conference on Advances in Social Networks Analysis and Mining 270-275 IEEE Computer Society, (2010).

**SAID A., DE LUCA E.W, ALBAYRAK S.,** *How social relationships affect user similarities***.** IUI 2010 workshop on Social Recommender Systems (SRS), Hong Kong, China, 2010.

**SALTON G., YANG C.S.,** *On the specification of terms values in automatic indexing,* Journal of Documentation, Vol. 29 Iss: 4, pp.351 – 372, (1973).

**SALTON G.,** *The SMART retrieval system: experiments in automatic document processing,* Englewood Clis, NJ: Prentice-Hall, 1971.

**SAKAGAMI H., KAMBA T.,** *Learning personal preferences on online newspaper articles from user behaviors,* In proceedings of the sixth International World Wide Web Conference (WWW'06), 7—11, 1997.

**SARUKKAI R. R.,** *Link prediction and path analysis using Markov chains.* Computer Networks, 33, 377–386, (2000).

**SCHAEFFER S. E.,** *Graph clustering,* Computer Science Review, vol. 1, n 1, p. 27–64, 2007.

**SCHUBERT P., KOCH M.,** *The power of personalization: customer collaboration and virtual communities.* Proceedings of the Eighth Americas Conference on Information Systems (AMCIS), 1953– 1965, (2002).

**SEN S., VIG J., RIEDL J.,** *Tagommenders: connecting users to items through tags,* WWW '09 Proceedings of the 18th international conference on World wide web, Pages 671-680, 2009.

**SHAVLIK J., CALCARI S., ELIASSI-RAD T., SOLOCK J.,** *An instructable, adaptive interface for discovering and monitoring information on the world-wide web,* In Proceedings of the 1999 international conference on Intelligent user interfaces, 157—160, 1999.

**SHAVLIK J., ELASSI-RAD T.,** *Intelligent Agents for Web-based Tasks: An Advice-Taking Approach,* In AAAI/ICML Workshop on Learning for Text Categorization, 63—70, AAAI Press, 1998.

**SIDIR M., LUCAS N., GIGUET E.,** De *l'analyse des discours à l'analyse structurale des réseaux sociaux : une étude diachronique d'un forum éducatif,* Article de Recherche, Revue Sticef, Volume 13, 2006.

**SIERSDORFER S., SIZOV S.,** *Social recommender systems for web 2.0 folksonomies.* 20th ACM Conference on Hypertext and Hypermedia, Hypertext 2009: 261-270.

**SINHA R. R., SWEARINGEN K.,** *Comparing Recommendations Made by Online Systems and Friends,* In DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries (2001).

**SLUIJS V. D., HOUBEN K. G.J.,** *Towards a Generic User Model Component,* Workshop on Personalization on the Semantic Web (PerSWeb05), workshop at the 10th International Conference on User Modeling (UM2005), pp. 47-57, Edinburgh, Scotland (2005).

**SOMLO G. L., HOWE A. E.,** *Using web helper agent profiles in query generation,* In Proceedings of the second international joint conference on Autonomous agents and multiagent systems *(AAMAS'03), 812-818, ACM Press, 2003.*

**SPERETTA M., GAUCH S.,** *Personalized Search Based on User Search Histories.* Web Intelligence 622-628, 2005.

**STEFANI A., STRAPPARAVA C.,** *Personalizing Access to Web Sites: The SiteIF Project,* Proceedings of the 2nd Workshop on Adaptive Hypertext and Hypermedia HYPERTEXT'98, Pittsburgh, USA, June 20-24, 1998.

**STENGER T., COUTANT A.,** *How teenagers deal with their Privacy on social network sites? Results from a national survey in France,* actes de l'Intelligent Information Privacy Management Symposium, Stanford (USA), pp. 169-174, mars 2010.

**STORY H., HARBULOT B., JACOBI I., JONES M.,** *FOAF+SSL: RESTful Authentication for Social Web,* URL= http://dig.csail.mit.edu/2009/Papers/SPOT/foaf-ssl-spot2009.pdf, 2009.

**STUTZMAN F.,** *An evaluation of identity-sharing behavior in social networks communities,* iDMAa Journal, vol.3,no.1, 2006. http://www.ibiblio.org/fred/pubs/stutzman pub4.pdf

## T

**TANUDJAJA F., MUI L.,** Persona: *A Contextualized and Personalized Web Search,* Proceedings of the 35th Hawaii International Conference on System Sciences, page 67, 2002.

**TCHUENTE D., CANUT M-F., JESSEL N., PÉNINOU A., SÈDES F.,** *Visualizing the relevance of social ties in user profile modeling.* Web Intelligence and Agent Systems, IOS Press, Vol. 10 N. 2, p. 261-274, mai 2012.

**TCHUENTE D., CANUT M-F., JESSEL N., PÉNINOU A., 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, 26/08/12-29/08/12, IEEE Computer Society, 2012 (BIS).

**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, 09/08/10-11/08/10, IEEE Computer Society, p. 370-374, août 2010.

**TCHUENTE D., CANUT M-F., JESSEL N., PÉNINOU A., SÈDES F.,** *Dérivation de profils utilisateurs à partir de des réseaux sociaux : une approche par communautés de réseaux égocentriques.* Ingénierie des Systèmes d'Informations, numéro spécial Social, Localisation et Mobilité, de nouveaux enjeux pour la gestion des données, 2013 (à paraître).

**TCHUENTE D., CANUT M-F., JESSEL N., PÉNINOU A., SÈDES F.,** *Comment dériver efficacement le profil de l'utilisateur à partir de son graphe social ? Expérimentation dans DBLP et Mendeley.* Ingénierie des Systèmes d'Informations, numéro spécial Extension INFORSID, 2013 (BIS) (à paraître).

**TCHUENTE D., CANUT M-F., JESSEL N., COUTANT A., STENGER T., RAMPNOUX O.,** *Pour une approche interdisciplinaire des TIC : le cas des réseaux socionumériques.* Document numérique, Hermès, Numéro spécial Collaboration interdisciplinaire au service de la complexité, Vol. 14, N. 1, p. 31-57, avril 2011.

**TCHUENTE D., JESSEL N., CANUT M-F.,** *Accès à l'information dans les réseaux socionumériques.* Hermès, CNRS EDITIONS, Numéro spécial Ces réseaux numériques dits sociaux, Vol. 59, p. 59-64, avril 2011.

**TCHUENTE D., CANUT M-F., JESSEL N.,** *Détection des profils à court-terme et à long-terme dans les réseaux sociaux.* Revue des Nouvelles Technologies de l'Information, Cépaduès Editions, Numéro spécial Actes de la Conférence Internationale Francophone Extraction et Gestion des Connaissances - EGC 2011, Vol. Hors-série, p. 365-371, 2011.

**TCHUENTE D., CANUT M-F., JESSEL N., PÉNINOU A.,** *Quelle modélisation des profils utilisateurs des réseaux sociaux numériques?* Colloque International EUTIC - Usages et Enjeux des TIC (EUTIC 2009), Bordeaux, 18/11/09-20/11/09, Maison des Sciences de l'Homme, p. 163-175, 2010.

**TCHUENTE D., CANUT M-F., JESSEL N., PÉNINOU A., SÈDES F.,** *Modèle et techniques de dérivation de profils utilisateurs à partir de réseaux sociaux égocentrés.* INFormatique des Organisations et Systemes d'Information et de Décision (INFORSID 2012), Montpellier, 29/05/12-31/05/12, Association INFORSID, p. 207-222, mai 2012.

**TCHUENTE D., PÉNINOU A., JESSEL N., CANUT M-F., SÈDES F.,** *Modélisation générique du processus de développement des profils utilisateurs dans les systèmes d'information.* Colloque Veille Stratégique Scientifique et Technologique (VSST 2012), Ajaccio, 24/05/12-25/05/12, Université Paul Sabatier - Toulouse, (support électronique), mai 2012.

**TCHUENTE D., JESSEL N., CANUT M-F.,** *Conception de profils visuels d'utilisateurs à partir de réseaux égocentriques : Cas de Facebook.* Colloque Veille Stratégique Scientifique et Technologique (VSST 2010), Toulouse, France, 25/10/10-29/10/10, 2010.

**TEBRI H., BOUGHANEM M., CHRISMENT C., TMAR M.,** *Incremental profile learning based on a reinforcement method.* SAC'2005- 20th ACM Symposium on Applied Computing, Santa Fe, New Mexico, USA, 13/03/2005-17/03/2005, ACM, p. 1096-1101, mars 2005.

**TEEVAN J., DUMAIS S. T., HORVITZ E.,** *Personalizing search via automated analysis of interests and activities,* SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval New York, NY, USA: ACM Press, p. 449—456, (2005).

**TIANJUN F., HSINCHUN C.,** *Analysis of cyberactivism: A case study of online free Tibet activities, Intelligence and Security Informatics*, Ingenierie des Systèmes d'information, ISI 2008, P. 1-6.

**TRAJKOVA J., GAUCH S.,** *Improving Ontology-Based User Profiles*, Dans Recherche d'information assistée par ordinateur RIAO 2004: 380-390.

U

**ULDIS B., PASSANT A., CYGANIAK R., BRESLIN J.,** *Weaving SIOC into the Web of Linked Data,* LDOW2008, Beijing, China.

V

**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.

W

**WAP FORUM.,** *User Agent Profile WAG UAProf Wireless Application Protocol,* 2001. URL= http://www.wapforum.org.

**WEI H., ARTHIR-NICOLAE M., CRISTINA M.,** *Sensing learner interest through eye tracking.* Ninth IT & T Conference, Dublin Institute of Technology, Dublin, Ireland, 22nd.-23rd. October (2009).

**WEN J-R., LAO N., MA W-Y.,** *Probabilistic Model for Contextual Retrieval,* SIGIR '04 Proceedings of the 27th annual international ACM SIGIR conference on Research, and development in information retrieval, Pages 57 – 63, 2004.

**WHITE R. W., JOSE J. M., RUTHVEN I.,** Comp*aring explicit and implicit feedback techniques for Web retrieval:* TREC-10 interactive track report , Proceedings of the Tenth Text REtrieval Conference (TREC 2001), 2001.

**WIDYANTORO D. H., YIN J., EL NASR M. S., YANG L., ZACCHI A., YEN J.,** *Alipes: A Swift Messenger in Cyberspace,* In Spring Symposium on Intelligent Agents in Cyberspace Palo Alto: (March 1999) , p. 62--67.

Z

**ZAYANI C.,** *Contribution à la définition et à la mise en oeuvre de mécanismes d'adaptation de documents semi-structurés.* Thèse de doctorat, Université Paul Sabatier, mai 2008.

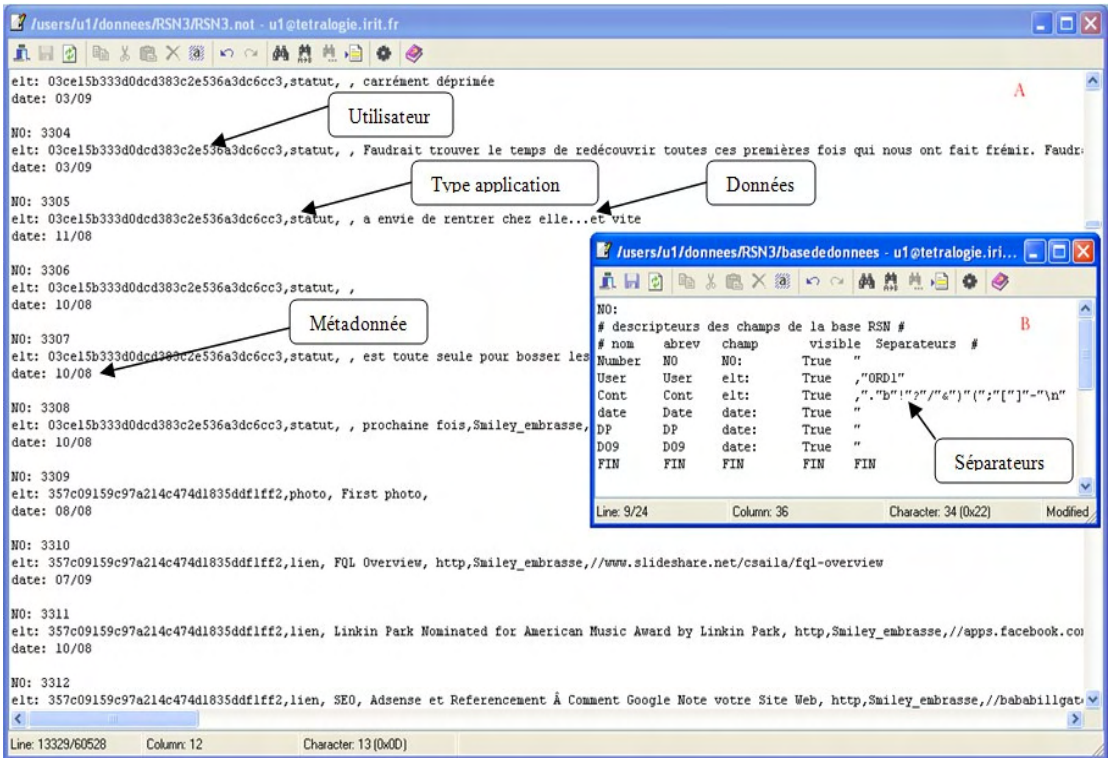
**ZEMIRLI W. N.,** *Modèle d'accès personnalisé à l'information basé sur les diagrammes d'influence intégrant un profil multidimensionnel.* Thèse de doctorat, Université Paul Sabatier, juin 2008.

**ZENG Y., YAO Y.Y., ZHONG N.,** *Dblp-sse: A dblp search support engine*. In: Proceedings of the 2009 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 626–630 (2009).

**ZENG Y., WANG Y., HUANG Z., DAMLJANOVIC D., ZHONG N., WANG** C., *User Interests: Definition, Vocabulary, and Utilization in Unifying Search and Reasoning, 2010* International *Conferences* on Active Media Technology, AMT 2010: 98-107, 2010.

# Annexe A : Exemple de documents, filtres et graphe réalisés dans Tétralogie

## 1. *Exemple de représentation de données, métadonnées et descripteurs sous le format de l'environnement Tétralogie (Tchuente et al., 10) (Tchuente et al., 11)(Dousset, 06).*



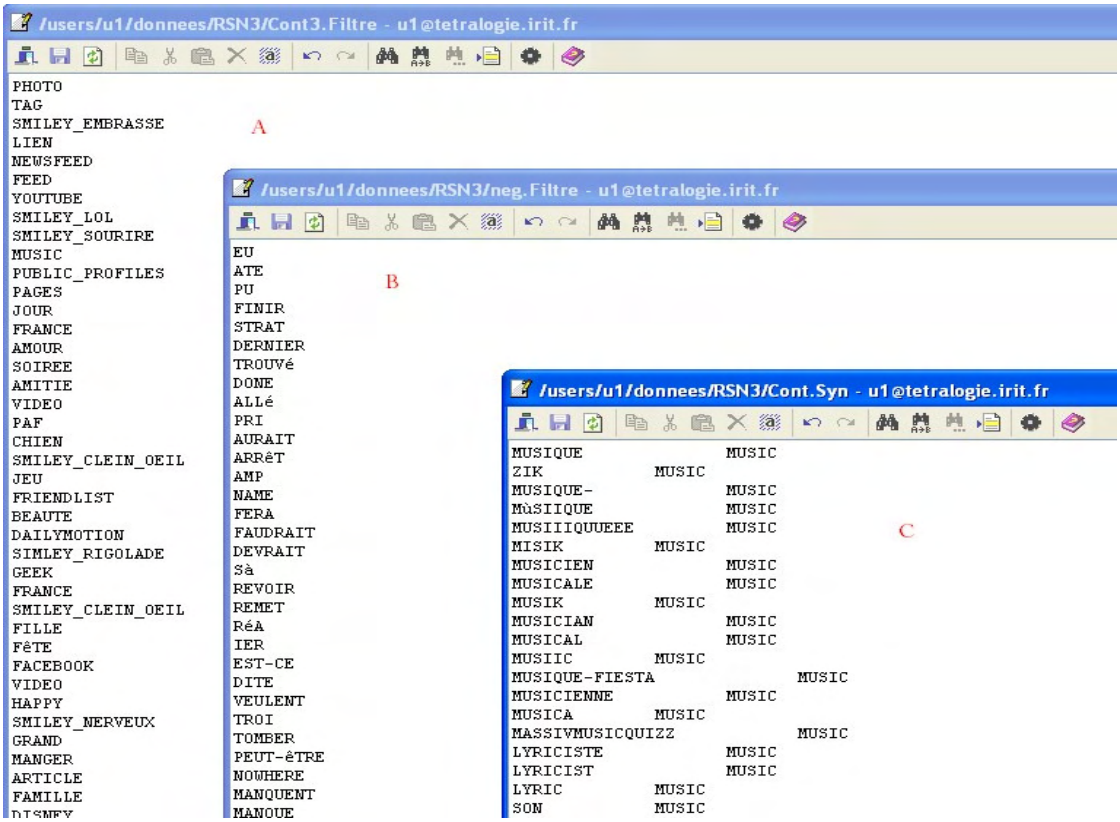

Le logiciel Tétralogie permet de structurer un texte suivant des champs (*NO*, *elt*, et *date* dans ce cas), figure 1A. Le champ *NO* ici représente le numéro d'une activité d'un utilisateur dans Facebook (commentaire, statut, lien, page, groupe, événement, etc.). Le champ *elt* représente le contenu textuel associé à l'activité. Il peut s'agir par exemple du texte d'un statut publié par l'utilisateur, de la description d'un lien publié, ou encore de la description d'une page (groupe ou événement) à laquelle est connecté l'utilisateur, etc. Le champ *date* ici représente une métadonnée (exemple donnée de contexte dans le modèle de profil proposé) qui pourra ensuite être exploitée pour réalisée des analyses évolutives.

Pour extraire les termes significatifs des champs de données, un fichier descripteur de données (figure 1B) définit les séparateurs qui vont permettre le découpage de chaque champ en termes distincts. Le fichier de description de données peut également définir en même temps des champs supplémentaires à partir du contenu des champs existants. Le champ *User* (utilisateur) contiendra par exemple l'ensemble des utilisateurs. Le séparateur (,« ORD1 ») utilisé pour définir le champ *User* à partir du champ *elt* indique que le nouveau champ *User* aura pour contenu les premiers termes avant la première virgule dans le champ *elt* (ce qui correspond bien à un identifiant crypté d'un utilisateur sur le document de la figure 1A). Le champ *Cont* (pour contenu) est également un nouveau champ qui contiendra tous les termes distincts extraits à partir du champ

*elt* en appliquant la liste de séparateurs indiqués. Dans cette liste, « b » représente par exemple un blanc (espace).

Les deux documents présentés sur cette figure vont ainsi permettre de structurer un texte et d'en extraire des champs significatifs qui contiendront chacun un ensemble de termes. Ces termes pourront ensuite être filtrés suivant des dictionnaires, filtres positifs ou filtres négatifs (Annexe A.2 qui suit).

### 2. *Exemple de dictionnaires et filtres Tétralogie (Tchuente et al., 10) (Tchuente et al., 11)(Dousset, 06)*




Une fois les champs définis par un descripteur de documents, ils contiendront des termes qu'il est possible de filtrer principalement de trois manières : filtres positifs, filtres négatifs et dictionnaires.

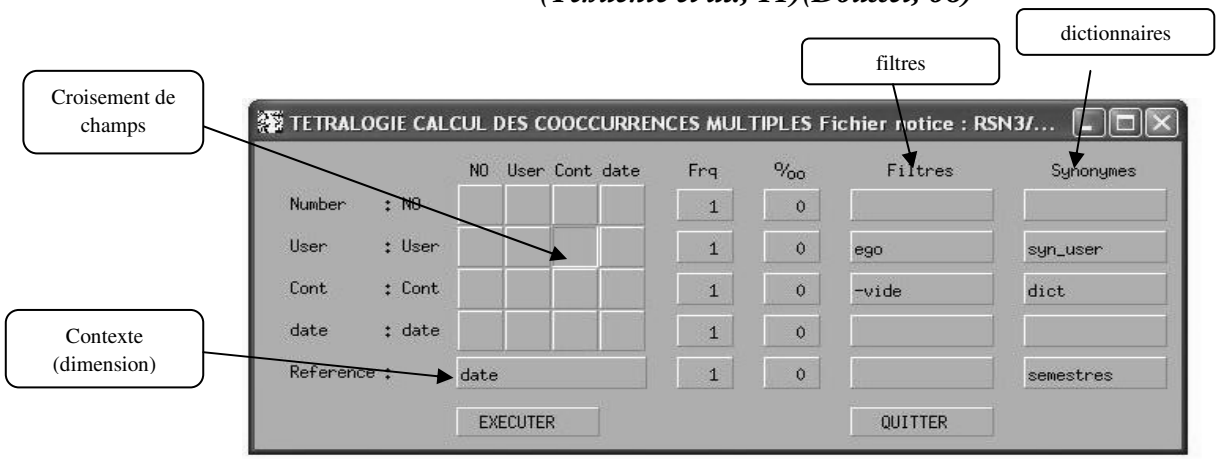
Les filtres positifs permettent d'indiquer explicitement les termes qui doivent retenus dans les analyses (figure 2A). Ceci peut être réalisé par un expert du domaine, ou en s'appuyant sur une ontologie du domaine (s'il en existe). Sinon, dans un contexte très ouvert comme les réseaux sociaux numériques où la terminologie peut être très diversifiée, nous réalisons les filtres positifs en utilisant tous les termes d'un champ, qu'on raffine en appliquant les filtres négatifs et les dictionnaires.

Les filtres négatifs sont les termes n'ayant pas de sens pris tout seul (figure 2B). Ce sont par exemple les articles, certains verbes, etc. Pour la langue française et anglaise, le logiciel Tétralogie dispose d'une liste prédéfinie de termes directement exploitable comme filtre négatif.

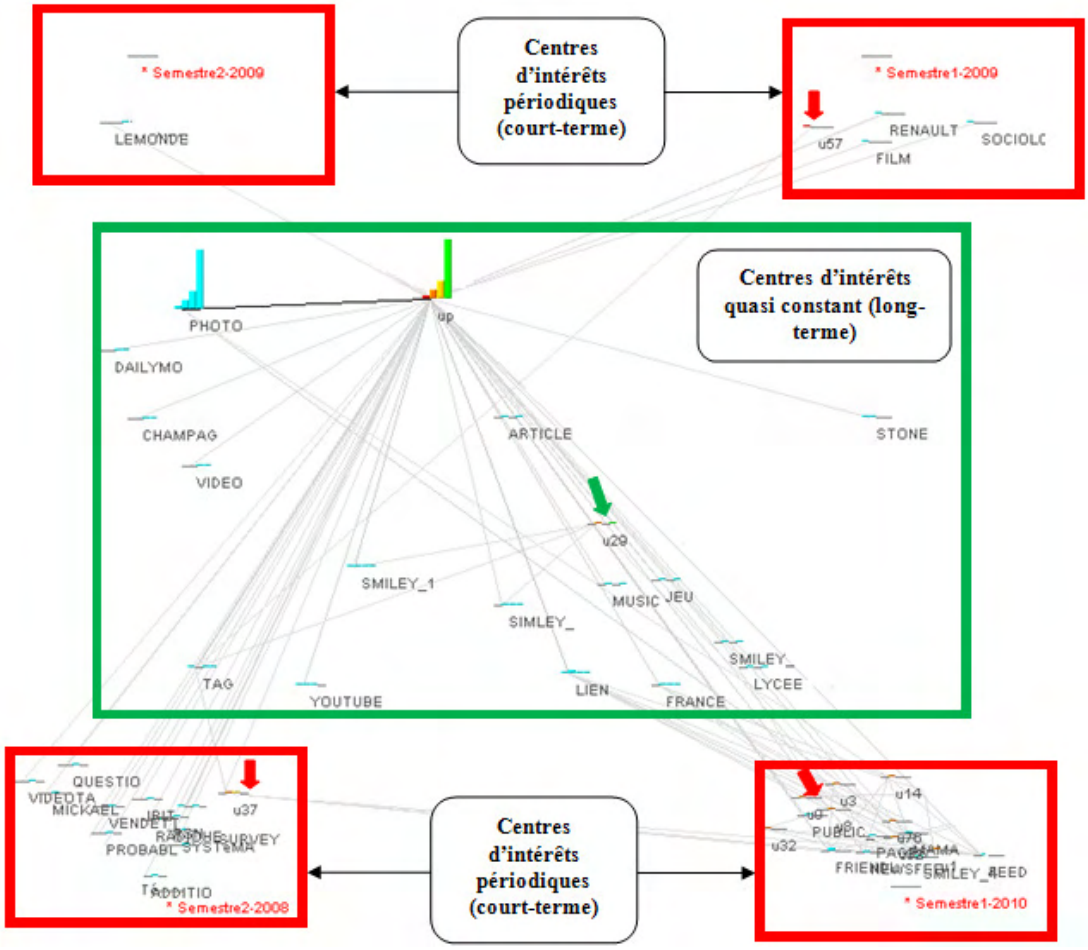
Les dictionnaires (dictionnaires de synonymes en particulier) permettent de regrouper plusieurs termes en un seul (figure 2.3). Le logiciel Tétralogie dispose des algorithmes de linguistiques qui peuvent aider à la

réalisation de pareils dictionnaires à partir d'un corpus. Sinon, l'analyste a également la possibilité de fournir son propre dictionnaire (créer à partir d'une ontologie de domaine par exemple) lors des analyses.

### 3. *Exemple de graphe biparti généré à partir de Tétralogie (Tchuente et al., 10) (Tchuente et al., 11)(Dousset, 06)*




A- Exemple de croisement entre champs, filtres et dictionnaires




B- Exemple de graphe biparti temporel (utilisateurs-centres d'intérêts dans le temps)

Une fois que les champs, les filtres et les dictionnaires sont construits, le logiciel Tétralogie offre de nombreuses possibilités de visualisation de graphe que nous avons exploité dans notre cas pour visualiser les profils construits (graphes bipartis utilisateurs-centres d'intérêts).

Pour visualiser les graphes, il faut déterminer les champs dont le contenu (termes significatifs) déterminera les nœuds du graphe. Dans notre cas, nous réalisons un croisement entre le champ contenu *Cont* (qui contient les termes significatifs considérés comme centres d'intérêts des utilisateurs) et le champ *User* (qui contient les utilisateurs) (figure 3A). Pour extraire les termes pertinents de chaque champ, les dictionnaires et filtres sont exploités. Pour le champ *User*, nous utilisons un filtre positif appelé *ego* (qui contient l'égo et tous les membres de son réseau égocentrique) et un dictionnaire *syn_user* (qui renomme simplement les identifiants utilisateurs sous la forme U1, U2, U3, etc.). Pour le champ *Cont,* nous utilisons un filtre négatif *(-vide*) pour extraire les termes n'ayant pas de sens pour les analyses (« mots vides »), et un dictionnaire de synonyme (*dict*) pour rassembler les termes ayant un même sens. L'usage de deux champs distincts dans le croisement va permettre de générer un graphe biparti. Dans le cas de cet exemple, nous souhaitons réaliser une analyse contextuelle pour suivre l'évolution du profil de l'utilisateur (et de celui de son réseau égocentrique). Pour ce faire, nous rajoutons au croisement un champ de référence (*date*) qui permettra d'avoir un graphe biparti temporel dans lequel on pourra visualiser l'évolution des profils sur le temps (figure 3B).

Sur le graphe biparti temporel présenté sur la figure 3B, chaque nœud (un histogramme) représente l'utilisateur (barres à plusieurs couleurs) ou ses centres d'intérêts (barres vertes). Chaque barre représente la fréquence de l'activité de l'utilisateur ou la fréquence d'un centre d'intérêt sur une période de temps. Pour chaque nœud, les barres se succèdent dans le sens des aiguilles d'une montre à partir de la première période de temps (semestre 1 de l'année 2009 dans ce cas). Par exemple, la barre rouge représente la fréquence de son activité au semestre 1, année 2009, la barre orange représente la fréquence de son activité pour le semestre 1, année 2010 (qui ne prend actuellement en compte que les mois de janvier et février 2010), la barre jaune représente la fréquence de son activité au semestre 2, année 2008, la barre verte représente la fréquence de son activité au semestre 2, année 2009). Ce graphe permet d'identifier les centres d'intérêts périodiques (profil à court-terme) de l'utilisateur qui sont les histogrammes proches des périodes (cadres en rouge). Par exemple, on constate que l'utilisateur « $u_p$ » s'est intéressé aux films et à la marque Renault au semestre 1, année 2009. Les centres d'intérêts quasi constants (profil à long-terme) de l'utilisateur (histogrammes dont les barres sont significatives dans toutes les périodes de temps) se rapprochent du centre du graphe (cadre vert). Dans ce cas, on peut considérer que les centres d'intérêts du profil à long-terme de l'utilisateur sont les photos, les vidéos, les articles, le jeu, la musique, le champagne, etc.

# Annexe B : détail des calculs de l'exemple illustratif sur la dérivation de la dimension sociale à partir de communautés (cf. figure 4.6)

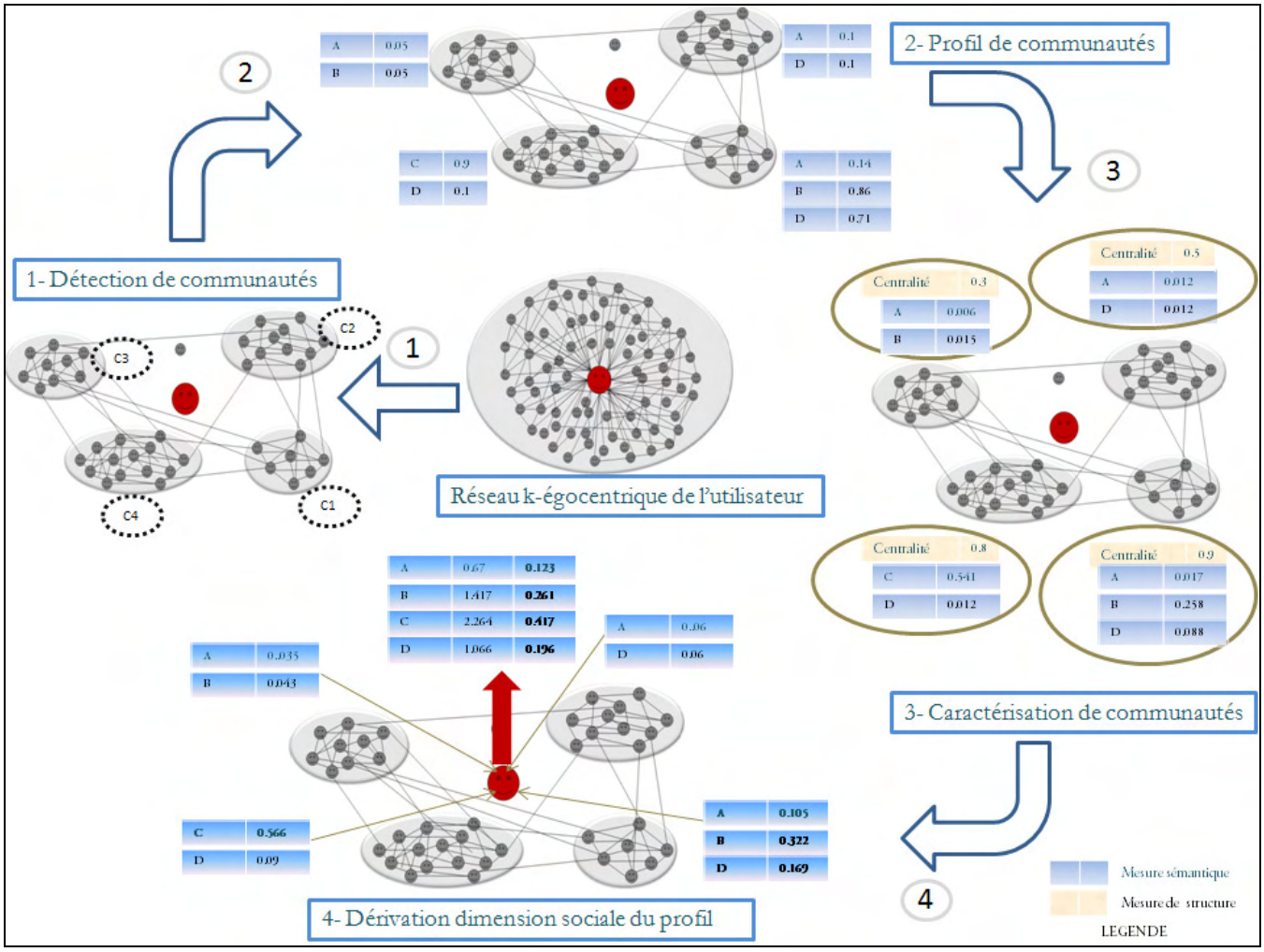



Figure 4. 8: Illustration du processus de dérivation de la dimension sociale à partir de communautés

Les détails de calcul des poids pour chacune des étapes de cette figure selon l'algorithme ($CoBSP_k$) sont présentés dans les tableaux ci-dessous.

| $P'$ | 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** |

*1. Poids P' des 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** |

*2. Poids P'' caractérisé sémantiquement (via tf.idf) du profil de 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** |

*3. Poids P''' caractérisés sémantiquement et structurellement (cf. formule 28, les mesures de centralités utilisées : 0.9, 0.5, 0.3, 0.8 sont arbitraires, cf. figure 4.6), la valeur du paramètre a est de 0.1 dans cet exemple.*

| $P^{social}$ | A | B | C | D |
|---|---|---|---|---|
| **Dimension sociale** | 0.105*4+0.06*+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** |

*4. Dérivation des poids $P^{social}$ la dimension sociale par la combinaison linéaire Lin_CombMNZ (cf. formule 31), les poids sont ensuite normalisés en divisant par la somme de tous les poids (0.67+1.417+2.264+1.066=5.417).*