Temporalité et réseaux sociaux : prise en compte de l'évolution dans la construction du profil utilisateur

Sirinya On-At

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *29/05/2017* **par :**

Sirinya ON-AT

Temporalité et réseaux sociaux : prise en compte de l'évolution dans la construction du profil utilisateur

---

JURY

Christine LARGERON	Professeur, Université Jean Monnet Saint-Etienne	Rapporteur
Josep LLADOS	Professeur, Universitat Autonoma de Barcelona	Rapporteur
Ladjel BELLATRECHE	Professeur, ENSMA Poitiers	Président du Jury
Hubert DUBOIS	Ingénieur, CEA Tech Toulouse	Examinateur
Florence SEDES	Professeur, Université de Toulouse	Directrice
André PENINOU	Maître de conférence, Université de Toulouse	Co-Encadrant
Marie-Françoise CANUT	Maître de conférence, Université de Toulouse	Co-Encadrant, Invitée
Nadine JESSEL	Maître de conférence (HDR),Université de Toulouse	Invitée

---

École doctorale et spécialité :
MITT : Image, Information, Hypermédia
Unité de Recherche :
Institut de Recherche en Informatique de Toulouse (UMR 5505)
Directeur(s) de Thèse :
Florence SEDES
Rapporteurs :
Christine LARGERON et *Josep LLADOS*

RESUME

Pour pouvoir restituer des informations qui correspondent aux besoins de l'utilisateur, les mécanismes d'adaptation doivent disposer de métadonnées sur celui-ci telles que ses caractéristiques personnelles, ses préférences générales, ses centres d'intérêt. De ce fait, le profil utilisateur construit à partir de celles-ci devient central dans tout système basé sur la personnalisation.

Dans cette thèse, nous nous focalisons sur l'approche qui consiste à s'appuyer sur le réseau social de l'utilisateur pour enrichir le profil de cet utilisateur, les métadonnées explicites étant complétées par les informations issues de notre processus d'analyse. Nous appelons les techniques ou processus associés à cette approche « profilage social ». Le terme « profil social » désigne un profil construit à l'aide du réseau social de l'utilisateur. Un profil social contient les métadonnées traduisant les intérêts de l'utilisateur extraits à partir des informations partagées par les individus de son réseau social.

Les intérêts de l'utilisateur évoluant au fil du temps dans la vie réelle, il en est de même pour ceux extraits depuis son réseau social : pertinents à un moment donné, ils peuvent ne plus être significatifs ultérieurement. Partant de ce constat, les principales informations que nous souhaitons étudier pour détecter un changement de centres d'intérêt ne sont pas ciblées sur l'utilisateur lui-même mais sur les éléments de son réseau social (liens entre les membres, informations qui circulent entre eux) : l'évolution du profil social de l'utilisateur est donc liée à l'évolution de son réseau social.

Nous proposons une démarche générique de profilage social efficace permettant de construire un profil social représentatif de l'utilisateur prenant en compte différents types de réseaux ainsi que leurs caractéristiques évolutives. Pour prendre en compte l'évolution des intérêts dans le profil social, nous avons proposé d'améliorer l'efficacité des processus de construction du profil social existants en intégrant la prise en compte de l'évolution du réseau social de l'utilisateur. Nous proposons d'intégrer un facteur temporel dans ces processus (approche basée sur des individus et approche basée sur les communautés). La solution permet de privilégier les intérêts provenant d'informations significatives et à jour. Il s'agit donc d'intégrer une mesure temporelle dans l'étape d'extraction et pondération des intérêts. Cette mesure est calculée d'une part, à partir de la pertinence temporelle des informations utilisées pour extraire cet intérêt et d'autre part, à partir de la pertinence temporelle de l'individu qui partage ces informations.

Nous mettons en œuvre la méthode proposée au travers d'expérimentations dans deux réseaux sociaux différents : *DBLP*, un réseau de publications scientifiques et *Twitter*, un réseau de micro-blogs. Les résultats de ces expérimentations nous ont permis de montrer l'efficacité de la méthode temporelle proposée par rapport aux processus de construction du profil social qui ne prennent pas en compte des critères temporels. En étudiant les résultats en fonction des techniques de pondération des intérêts ou fonctions temporelles utilisées, nous constatons que la fonction temporelle et la technique utilisées donnant les meilleurs résultats varient selon l'approche de construction du profil social choisie, selon la taille et la densité du réseau étudié mais aussi selon sur le type de réseau.

La problématique abordée dans cette thèse est relativement nouvelle dans le contexte des systèmes de personnalisation de l'information et ouvre de nombreuses perspectives : évaluation du profil social dans un système de recommandation par exemple, application de la méthode proposée dans d'autres types de réseaux sociaux, application de techniques de mise à jour du profil, conception d'une plateforme de construction du profil social selon les caractéristiques du réseau.

Mots-clés : *Système de personnalisation, Profil utilisateur, Profil social, Réseau égocentrique, Evolution.*

ABSTRACT

User profiling is essential for personalization systems (e.g. personalized information retrieval systems, recommendation systems) to identify user information (preference, interests...), in order to propose relevant content based on his/her specific needs and requirements.

Many works have shown that user's social neighbors can be a meaningful source to infer his/her interests. Besides, sociology works have shown that the user is better described by people around him/her, especially the people that are directly connected to him/her (his egocentric network). In this work, the term "social profiling" is considered as the interest extraction approach that consists in extracting user interests from information of his/her social neighbors. The user's profile built within this approach is called "social profile".

As user behaviors evolve over time, it is necessary to take into consideration the evolution of user interests in user profiling process. In the case of social profile, user interests are extracted from the information shared by his/her social neighbors. Hence, the evolution of extracted interests is related to the evolution of information shared on user social network and to the evolution of relationships between the user and his/her social neighbors. This issue becomes particularly important in the Online Social Networks (OSNs) context where user behavior changes quickly. For a user, the relationships and information in his/her social network can evolve and become obsolete for him/her overtime. Two users creating a relationship are not required to know each other in real life. Thus, the relationship persistence is not always maintained in this case. Social events or viral marketing (buzz) are also factors that enhance online social content sharing.

In this work, we propose a generic approach that considers the evolution in user's social network in the social profiling process and can be applied in different types of social network. To handle this, we propose to apply a time-aware method into existing social profile building process (individual based and community based approaches). This strategy aims at weighting user's interests in the social profile based on their temporal score. The temporal score of an interest is computed by combining the temporal score of information used to extract the interests (computed by considering their freshness) with the temporal of individuals who share the information in the network (computed by considering the freshness of the interaction with the user). The technique and temporal function used to compute the temporal score are customizable. Thus, we can find out the most appropriate technique or temporal function depending on the types or characteristics of the adopted social network.

The experiments conducted on DBLP and Twitter showed that the so-called time-aware social profiling process applying our proposed time-aware method outperforms the existing time-agnostic social profiling process. We also found that the most appropriate technique, temporal function and social profiling approach vary depending on the network characteristics (size, density) and to the social network type.

Our approach opens many opportunities for future studies in social information filtering and many application domains as well as on the Web (e.g. evolution of social profile in personalization of search engines, recommender systems in e-commerce,). Our long-term perspective consists in the proposal of a generic platform that extracts the information and builds the user social profile based on the type and the specific characteristics of the underlying social network. Such a platform would be parameterized by the characteristics of the targeted social network using a machine learning approach.

Keywords: *Personalization system, User Profile, Social Profile, User's interests, social networks, egocentric network, social network evolution, temporal method*

REMERCIEMENTS

Cette thèse est le fruit de la participation d'un ensemble de personnes qui ont permis de près ou de loin que ces travaux de recherche aboutissent. Je souhaiterais adresser tous mes remerciements à toutes celles avec qui j'ai pu échanger et qui m'ont aidée pour la réalisation de ce mémoire.

Je commencerai par remercier et exprimer ma profonde gratitude envers ma directrice de thèse, Mme Florence Sèdes pour la confiance qu'elle m'a accordée en acceptant de diriger mes recherches depuis mon stage de master, stage qui s'est concrétisé par la réalisation de cette thèse. Sa pédagogie, ses conseils avisés m'ont permis d'avoir une vision plus large de mes travaux. Je n'oublierai pas sa sympathie et ses soutiens moraux qui m'ont beaucoup aidé dans l'avancement dans mes travaux.

Je tiens à exprimer ma profonde gratitude à mes co-encadrants, Marie-Françoise Canut et André Péninou pour leur encadrement, leurs conseils, leur grande disponibilité et le temps qu'ils ont consacré aux vérifications et à l'avancement de mon travail, et ce, depuis mon stage de master. Ce fut un véritable plaisir de travailler avec eux. Leur grande patience, leurs encouragements et leur sympathie ont été pour moi une grande motivation. Ils m'ont beaucoup aidée à avancer dans les moments les plus difficiles. Je ne les remercierai jamais assez et je leur en serai toujours reconnaissante.

Je remercie vivement Mme Christine Largeron, Professeure à l'université Jean Monnet de Saint - Etienne, M. Josep Llados, Professeur à l'université autonome de Barcelone, M. Ladjel Bellatrech, Professeur à l'école nationale supérieure de mécanique et d'aéronautique de Poitiers, M. Hubert DUBOIS, Ingénieur chercheur au CEA Tech Toulouse pour avoir montré un intérêt pour ces travaux en acceptant d'être rapporteur ou examinateur de cette thèse.

Je remercie Nadine Baptiste-Jessel, pour les multiples heures de discussions qu'elle a consacrées sur mon travail et pour sa disponibilité, ses conseils et surtout pour ses critiques constructives qui m'ont permis de consolider mes idées et d'améliorer mes recherches.

Je remercie Arnaud Quirin, ancien post-doctorant de l'équipe, pour ses compétences techniques, ses discussions rigoureuses et sa sympathie qui m'ont beaucoup aidée dans la réalisation ce travail. Ce fut un grand plaisir d'avoir travaillé avec lui.

Je remercie la direction du laboratoire de l'IRIT et toute l'équipe SIG pour m'avoir accueillie chaleureusement. Je remercie Josiane Mothe et Olivier Teste, tous deux responsables de l'équipe SIG pour tous les efforts qu'ils consacrent au bon déroulement des travaux des doctorants.

Je remercie Guillaume Cabanac, Thierry Millan, Gilles Hubert et Pierre Règnier, pour m'avoir accordé leur confiance lors des enseignements réalisés à l'IUT et à l'université Paul Sabatier.

Je n'oublie pas Chantal Morand, Françoise Agar, Jean-Pierre Baritaud, Jean-Philippe Cornille, les responsables de la plateforme OSIRIM : Jacques Thomazau et Guillaume Dubreule, et tout le personnel de l'IRIT et de l'école doctorale pour leur aide durant ces années au laboratoire.

Je remercie tous les anciens et actuels doctorants de l'équipe SIG pour leur complicité et les bons moments partagés. Je pense en particulier à mes collègues du bureau Manel, Hamdi et Franck, sans oublier les anciens : Imen, Jérémy et Liana. Ce fut un véritable plaisir de partager le bureau avec vous. Je pense également à ma camarade de classe Chiraz, qui aura débuté et terminé sa thèse le même jour que moi ! Mais aussi à Mohammed, Hamid, Baptiste, Mahdi, Franck, Wafa, Mahmoud, Arpit, Ghada, Jiefu, ... sans oublier les anciens : Dieudonné, Anna-Maria, Dana, Andra, Ahmed.

J'adresse ma sincère reconnaissance à l'ensemble de mes amis de Thaïlande et de France (P'Bow, Guillaume, P'Tuk, P'Toy, Stéphane, François, David, P'Ple, Aom, Mery, ...) pour leur aide, leur soutien et les moments partagés ensemble durant mon séjour en France.

Je remercie les familles Fontayne et Moreau pour leur accueil chaleureux et les bons moments partagés ensemble. Leur gentillesse et leur sympathie m'ont fait chaud au cœur même lorsque j'étais loin de ma famille.

Un énorme merci à mon co-capitaine, Rémy pour sa présence sans faille à mes côtés et pour m'avoir fait garder le cap dans les hauts et les bas durant tout ce parcours. Merci pour ta patience, tes conseils, tes encouragements. C'est une très grande chance de te connaître et de partager des moments inoubliables avec toi.

Merci à ma sœur pour nos échanges et pour son soutien dans les moments difficiles. Je remercie également mon frère, mes oncles, mes tantes et mes cousins pour leur soutien, sans oublier ma grande mère qui aurait été si fière ...

Enfin, mes derniers remerciements, vont à l'égard des deux êtres qui me sont le plus chers : mes parents qui auront consacré leur vie à mon bonheur, et à mon avenir. Merci pour leur confiance et leur soutien sans faille. Je tiens à leur dédier cette thèse.

Merci à tous, merci, merci !

TABLE DES MATIERES

1. INTRODUCTION GENERALE

1.1. Contexte

Pour pouvoir restituer des informations qui correspondent aux besoins de l'utilisateur, les mécanismes d'adaptation doivent disposer de métadonnées sur celui-ci telles que ses caractéristiques personnelles, ses préférences générales, ses centres d'intérêt. De ce fait, le profil utilisateur construit à partir de celles-ci devient central dans tout système basé sur la personnalisation.

Généralement, dans les approches classiques, le profil utilisateur est construit à partir des activités propres de l'utilisateur, soit de façon explicite par lui-même, soit de façon implicite par l'extraction de ses intérêts lors de ses interactions avec le système (Gauch et al., 2007).

Le foisonnement du Web 2.0 plonge l'utilisateur dans un univers d'informations générées par lui-même et par les autres utilisateurs du système : c'est cet univers que nous nommons écosystème et qui offre une source très riche d'informations pour extraire des connaissances. Les divers contenus analysés viennent ainsi enrichir notre vision de son profil ou des processus dans lesquels il est impliqué, par exemple, les informations disponibles dans son(ses) réseau(x) social(aux).

En effet, les relations dans le réseau social de l'utilisateur sont basées sur le fait que tous partagent des intérêts ou des caractéristiques communes. L'utilisation des informations issues dudit réseau social permet d'une part, de compléter la représentation que le système se fait de l'utilisateur, et d'autre part, de restreindre l'espace d'investigation dans l'écosystème puisqu'il n'est plus nécessaire d'accéder à l'espace entier mais uniquement à ce sous-ensemble relatif à l'utilisateur.

Dans cette thèse, nous nous focalisons sur l'approche qui consiste à s'appuyer sur le réseau social pour enrichir le profil de l'utilisateur, les métadonnées explicites étant complétées par les informations issues de notre processus d'analyse. Nous appelons les techniques ou processus associés à cette approche, « profilage social ».

La construction de profil avec cette approche permet d'une part de compléter le profil d'un nouvel utilisateur généré a priori donc parfois vide (i.e. *cold start problem*) ou celui d'un utilisateur peu actif, cas pour lesquels le profil individuel de l'utilisateur s'avère insuffisant pour les mécanismes de personnalisation d'informations, et d'autre part, d'enrichir un profil utilisateur existant.

Le terme « *profil social* » désigne un profil construit à l'aide du réseau social de l'utilisateur. Un profil social contient donc les métadonnées traduisant les intérêts de l'utilisateur extraits à partir des informations partagées par les individus de son réseau social.

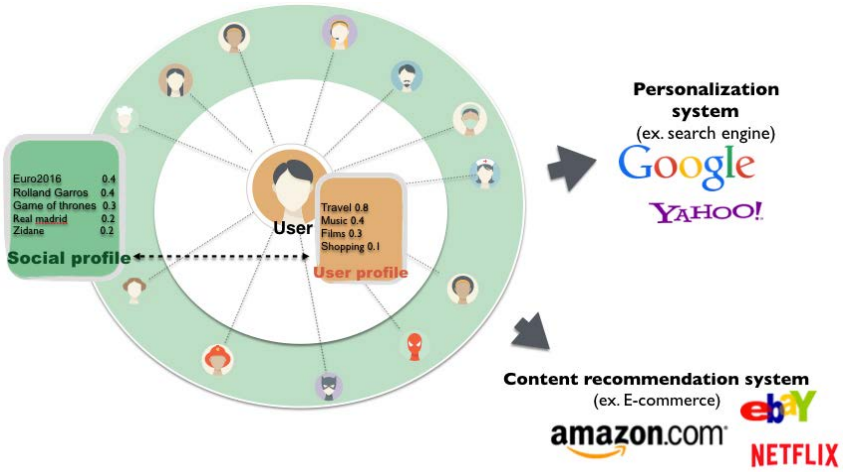


Figure 1.1 Profil utilisateur et profil social dans les systèmes de personnalisation d'informations

Les intérêts de l'utilisateur évoluant au fil du temps, il en est de même pour ceux extraits depuis son réseau social : pertinents à un moment donné, ils peuvent ne plus être significatifs ultérieurement.

L'évolution des intérêts de l'utilisateur est un problème soulevé et traité dans plusieurs travaux relatifs à la construction du profil utilisateur. A partir des travaux de la littérature, nous pouvons distinguer deux approches de gestion de l'évolution des intérêts dans le profil utilisateur : la première consiste à prendre en compte cette évolution pendant l'étape d'extraction des intérêts, la deuxième consiste à gérer cette évolution des intérêts après les avoir extraits et correspond dès lors à un processus de mise à jour du profil utilisateur.

Dans le cas du profil social, les principales informations que nous souhaitons étudier pour détecter un changement des intérêts ne sont pas ciblées sur l'utilisateur lui-même mais sur les éléments de son réseau social (liens entre les membres, informations qui circulent entre eux). L'évolution du profil social de l'utilisateur est donc liée à l'évolution de son réseau social.

L'arrivée des médias sociaux ces dernières années a permis d'offrir aux utilisateurs de nouvelles fonctionnalités ainsi qu'une meilleure accessibilité aux réseaux sociaux numériques (RSNs). Dans les médias sociaux, l'information partagée évolue sans cesse du fait des interactions sociales en ligne (partage, échange d'informations, signal social (« like », …)) qui génèrent (plus facilement) un volume important d'informations.

Construire le profil social sans prendre en compte les caractéristiques évolutives de ce type de réseau peut amener à construire un profil social contenant des intérêts non pertinents à terme ou à l'alourdir par des ajouts successifs et cumulatifs.

De plus, avec la diversité des RSNs, on peut trouver différents types de réseaux sociaux selon les objectifs d'utilisation, le type de relations mis en place, le type d'informations échangées ou même selon le comportement évolutif du réseau. Plus précisément :

- En ce qui concerne les objectifs d'utilisation, certains réseaux sociaux sont utilisés dans un contexte général, les informations partagées sont aussi générales (micro-blogs, sites de réseautage social, …) alors que dans d'autres réseaux sociaux, les informations sont plus ciblées et restreintes à un domaine (réseau de publications scientifiques, forum de discussions, réseau spécialisé, réseau professionnel, …). Les utilisateurs sont parfois aussi caractérisés par des comportements différents.

- Les types de relations varient également selon le réseau social. Sur les sites de réseautage social, les relations sont plus restreintes et sont généralement créées en se basant sur la connaissance qu'a chaque utilisateur de l'autre, alors que sur un réseau de partage d'information, nous pouvons trouver les relations de « suiveur d'informations » : un utilisateur établit une relation afin de suivre les informations que l'autre partage, sans forcément le connaître et sans que cela ne soit forcément réciproque.
- En ce qui concerne le type des informations échangées, chaque RSN offre la possibilité aux utilisateurs de partager ou d'annoter différents types d'informations (image, vidéo, texte, …). Par exemple, dans Facebook, les utilisateurs peuvent partager du texte, des images, des vidéos. Ils peuvent également attribuer la mention « aimer » (*like*) ou laisser des commentaires dans les contenus partagés par d'autres personnes. Depuis les dernières mises à jour, les utilisateurs peuvent même exprimer leur émotion sur un *post* à l'aide d'émojis (aimer, adorer, mécontent, triste). Dans Twitter, les utilisateurs peuvent partager des textes courts (*tweets*), images, vidéos, annoter les *posts* par des *hashtags* ou adresser le *post* à un autre utilisateur (@...).
- En ce qui concerne le comportement évolutif du réseau, l'évolution des relations et des informations dans les RSNs est différente selon le type de réseau social. Sur les sites de partage d'informations, les créations des liens ont lieu quand un utilisateur commence à suivre (*follows*) un autre utilisateur. Dans ce type de médias, les créations de liens peuvent augmenter rapidement suite à des événements importants. Les informations sont partagées sous la forme de flux d'informations. Les utilisateurs actifs partagent fréquemment leurs billets (*tweets*). Ce peut être toutes les heures, ou bien toutes les minutes. A l'inverse, dans les blogs, des informations sont partagées périodiquement : quotidiennement, hebdomadairement, ou mensuellement.

De par l'hétérogénéité des réseaux sociaux, nous pouvons penser que le type et les caractéristiques du réseau social peuvent avoir un impact sur le processus de construction du profil social. Or, ce constat, bien que partagé par toute une communauté autour de l'analyse de réseaux sociaux (SNA), n'est pas forcément pris en compte dans les travaux de recherche sur les systèmes de personnalisation de l'information.

Nos travaux tentent de pallier cette lacune dans l'objectif de proposer un profil social à la fois pertinent et à jour.

1.2. Problématique

L'évolution et l'hétérogénéité des RSNs comme vus précédemment peuvent avoir un impact sur la construction du profil social de l'utilisateur. Cela nous permet d'énoncer les problématiques suivantes :

- ***Comment construire un profil social à la fois pertinent et à jour dans le contexte des RSNs ?***

L'une des étapes du processus de construction du profil social de l'utilisateur est celle de sélection des informations pertinentes. Cette tâche devient cruciale dans les RSNs où l'évolution des informations est rapide et où l'accès à celles-ci est largement facilité. Dans les RSNs, les utilisateurs peuvent se connecter facilement en ligne avec d'autres personnes sans forcément les connaître, et sans pouvoir forcément filtrer le flux transmis par celles-ci. Ils peuvent ainsi être destinataires d'informations partagées qui ne correspondent pas forcément à

leurs intérêts. De plus, avec du recul sur leur usage, on observe que les relations créées peuvent ne plus être « entretenues », être supprimées, voire sur certains médias, être non persistantes (volatilité).

De la même manière, une information partagée par un ami dans le réseau social à un instant donné peut rapidement ne plus être significative pour l'utilisateur. Cela peut provenir du changement d'intérêt de l'utilisateur lui-même ou d'un changement d'intérêt de ses amis. On ne peut donc pas prendre en compte ou donner la même importance à toutes les informations existant dans les réseaux sociaux pour refléter les intérêts d'un utilisateur à un moment donné.

- ***Comment étudier l'influence des caractéristiques des RSNs sur la pertinence du profil social construit ?***

Compte tenu de la grande diversité des RSNs, il est difficile de trouver un processus de construction du profil social qui peut être appliqué à tous les types de réseau social. Une technique qui marche sur un réseau social pourrait ne pas marcher pour d'autres. Il s'avère donc important d'étudier l'impact du type et de la caractéristique du réseau social sur la pertinence du profil social construit par une technique donnée.

C'est sur la base de ces interrogations que nous avons construit les contributions de ce mémoire, à partir d'un état de l'art des usages en cours, des travaux académiques sur la gestion du profil et de la prise en compte de son évolution.

1.3. Contribution

En nous basant sur les travaux de (Tchuente et al., 2013) réalisés dans notre équipe, notre point de point de départ a consisté à considérer l'utilisation des informations extraites du réseau égocentrique de l'utilisateur : celui-ci peut être considéré comme un réseau social à part entière et est largement utilisé en sociologie. Il s'agit d'un graphe composé des relations entre les individus situés à distance d=1 (i.e. directement reliés) de l'utilisateur (appelé égo), l'égo étant bien entendu exclu de ce graphe. L'utilisation du réseau égocentrique de l'utilisateur est motivée par le fait que nous considérons que, dans la vie réelle, les intérêts d'un utilisateur sont principalement liés aux personnes qu'il côtoie dans la société (sphères privée et publique).

Dans nos travaux, le processus de construction du profil social est composé de 3 principales étapes qui permettent d'extraire les métadonnées qui constitueront le profil :

- Sélection des informations partagées par les membres du réseau égocentrique de l'utilisateur,
- Extraction des mots-clés de ces informations et pondération de ces mots-clés selon des métriques choisies,
- Agrégation des mots-clés pour obtenir les intérêts finaux pondérés avant de les dériver pour construire le profil social de l'utilisateur.

Pour aborder la prise en compte de l'évolution des RSNs mais aussi leurs différents types, nous proposons une démarche générique de profilage social efficace permettant de retourner un profil social représentatif de l'utilisateur prenant en compte différents types de réseaux ainsi que leurs caractéristiques évolutives. Notre contribution dans le cadre de cette recherche se décline selon deux axes :

- ***Prise en compte de critères temporels dans les processus de construction du profil social***

A partir du constat du caractère évolutif des RSNs, nous considérons deux types d'évolution dans le réseau social de l'utilisateur : l'évolution des relations entre l'utilisateur et ses amis et l'évolution des informations partagées par les amis dans son réseau social.

Il s'avère donc nécessaire d'étudier comment intégrer ces deux types d'évolution dans le processus de construction du profil social de l'utilisateur. Pour cela, il est nécessaire de trouver les modèles et méthodes qui permettent de prendre en compte ces deux types d'évolution afin d'extraire les intérêts les plus pertinents de l'utilisateur.

Il s'agit donc de sélectionner les sources d'informations (individus) les plus pertinentes pour l'utilisateur, puis, à partir des informations partagées par ces individus sélectionnés, de ne sélectionner que les informations les plus pertinentes pour l'utilisateur (qui correspondent à ses intérêts les plus à jour).

Dans ce travail, nous envisageons de prendre en compte des critères temporels sur les relations entre l'utilisateur et les individus dans son réseau égocentrique (prise en compte de l'évolution des relations) ainsi que des critères temporels sur les informations échangées dans son réseau égocentrique (prise en compte de l'évolution des informations).

Nous proposons d'intégrer un critère temporel dans les processus de construction du profil social existants (approche basée sur des individus et approche basée sur les communautés). La solution proposée doit permettre de privilégier les intérêts provenant des informations significatives et à jour.

Pour cela, nous utilisons une mesure temporelle dans l'étape d'extraction et de pondération des intérêts. Cette mesure est calculée d'une part, à partir de la pertinence temporelle des informations utilisées pour extraire cet intérêt et d'autre part, à partir de la pertinence temporelle de l'individu qui partage ces informations. La pertinence temporelle de l'information représente son importance, au moment t, vis-à-vis de l'utilisateur. De la même manière, la pertinence temporelle d'un individu représente l'importance de sa relation avec l'utilisateur, au moment t.

Cette pondération des intérêts selon ces critères temporels fait intervenir différents paramètres. Le poids dit « temporel » est une combinaison entre des poids « temporels » de l'individu et des informations. En effet, il existe plusieurs façons de faire la combinaison entre ces poids (somme, moyenne etc.). Comme décrit précédemment, dans le cas du réseau social, l'importance de chaque poids pourrait être différente selon le type du réseau social étudié. Il est donc judicieux de combiner les deux poids en prenant en compte l'influence de chacun sur la pertinence du poids « temporel » final. Concernant les facteurs temporels, nous avons observé que selon le type de réseau social, la pertinence temporelle des informations et/ou des individus peut varier. Dans un réseau où les informations évoluent très vite, la pertinence de ces informations pourrait changer très vite et vice-versa. Pour cela, il faut introduire et étudier un paramètre qui permet de moduler le taux d'évolution des informations en fonction de la caractéristique évolutive du réseau social concerné.

De ce fait, nous effectuons dans nos travaux, une étude paramétrique qui permet de trouver la meilleure combinaison des paramètres étudiés afin d'obtenir le poids temporel le plus significatif. Ceci permet également d'étudier l'impact de l'évolution des relations et des informations sur l'efficacité de la méthode proposée sur le réseau social concerné.

Les techniques et fonctions temporelles utilisées pour calculer le poids temporel sont paramétrables. Cela nous permet également de mettre en évidence la technique et la fonction temporelle appropriées selon le type et la caractéristique du réseau social. La méthode proposée s'avère donc suffisamment générique pour pouvoir être appliquée sur différents types du réseau.

- ***Etudes de cas sur différents types de réseaux sociaux***

La méthode proposée est mise en œuvre sur deux réseaux sociaux différents : DBLP qui est un réseau de publications scientifiques, Twitter qui est un réseau de micro-blogs. Ces deux réseaux sociaux possèdent des caractéristiques différentes en termes d'objectif d'utilisation, de type d'informations partagées, de type de relations et interactions entre les individus dans le réseau et enfin par rapport à leurs caractéristiques évolutives. Ces expérimentations permettent d'étudier l'impact des caractéristiques du réseau social sur la pertinence du processus de construction du profil social.

1.4. Organisation du mémoire

La suite du mémoire se décline en 5 chapitres :

Le chapitre 2 présente tout d'abord la notion de profil utilisateur ainsi que son utilisation dans le contexte de systèmes de personnalisation d'information, sujet central de cette thèse. Les problèmes existants dans ce contexte sont ensuite déclinés nous permettant de positionner nos travaux pour nous amener vers les études qui concernent la problématique visée. L'état de l'art sur les différentes phases et techniques de développement du profil utilisateur est ensuite présenté. La prise en compte de l'évolution du profil utilisateur est évoquée à la fin de ce chapitre.

Le chapitre 3 présente les travaux associés à la construction du profil de l'utilisateur qui s'appuie sur les informations de son réseau social, que nous appelons dans cette thèse « **profilage social** ». L'état de l'art sur cette partie porte sur deux axes d'étude : la construction du profil, qui repose sur l'extraction de connaissances comme vue dans la section précédente, et l'analyse des réseaux sociaux. Nous présentons tout d'abord le contexte de notre travail portant sur le profilage social. Ensuite, nous présentons les éléments fondamentaux d'un réseau social ainsi que les études d'analyse de réseaux sociaux associées à notre travail. Puis nous présentons les travaux existants dans le domaine du profilage social avant d'en faire la synthèse pour amener vers notre positionnement et notre contribution par rapport à ces travaux existants.

Le chapitre 4 présente notre contribution : nous présentons dans un premier temps, notre positionnement vis-à-vis des travaux existants, ensuite nous présentons les notions et concepts du profil social sur lesquels s'appuie notre travail. Enfin, nous présentons notre méthode de prise en compte de l'évolution du réseau social dans la construction du profil social ainsi que l'intégration de cette méthode dans les processus existants.

Le chapitre 5 présente l'évaluation de notre méthode. Nous commençons par donner la synthèse sur les stratégies d'évaluation qui pourraient être adaptées dans notre contexte. Puis nous présentons le protocole d'évaluation choisi. Ensuite, nous présentons les expérimentations menées pour chaque réseau social (DBLP, Twitter) ainsi que les résultats obtenus.

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

2. PROFIL UTILISATEUR

Dans ce premier chapitre, nous étudions de façon générale l profil utilisateur. Nous présenterons d'abord la notion de profil utilisateur ainsi que son utilisation dans les systèmes de personnalisation de l'information, contexte dans lequel se situent nos travaux. Nous présenterons ensuite, l'état de l'art sur les différentes phases et techniques de construction du profil utilisateur. La prise en compte de l'évolution du profil utilisateur sera également abordée dans la dernière section de ce chapitre.

2.1. Notion de profil utilisateur

La notion de profil utilisateur est largement abordée dans la modélisation utilisateur (*User Modeling*) qui peut être considérée comme le processus d'extraction de connaissances dans le but d'identifier les informations et les caractéristiques de l'utilisateur ou d'un groupe d'utilisateurs (Kobsa, 2007). La modélisation de l'utilisateur est une discipline de recherche datant des années 70, évoquée en premier lieu dans les travaux de (Allen, 1979 ; Cohen et Perrault, 1979), dans le domaine de l'interaction homme machine (*IHM*). L'utilisation du modèle de l'utilisateur dans ce domaine permet d'améliorer la qualité des *IHMs* : la déduction des préférences et contextes de l'utilisateur à partir des activités observées sert à déterminer dans un premier lieu, le type de dialogue que le système va avoir avec l'utilisateur, dans un deuxième temps, les métaphores graphiques les plus appropriées et enfin les modalités d'affichage des résultats. Le terme « modèle de l'utilisateur » est largement utilisé dans ce domaine pour assigner la notion de profil qui contient les informations de l'utilisateur (Zemirli, 2008).

Par la suite, les travaux liés à la modélisation de l'utilisateur, se retrouvent dans de nombreux domaines : système d'aide à l'apprentissage(Sehaba, 2012), systèmes hypermédia adaptatifs (Bra et al., 2002)et les systèmes de personnalisation d'informations (Anand et Mobasher, 2005 ; Fink et Kobsa, 2000).

Dans les systèmes de personnalisation d'informations, domaine dans lequel se situent nos travaux, le profil utilisateur est utilisé par des mécanismes tels que les systèmes de recommandation et les systèmes de recherche d'information personnalisée. Ces deux systèmes permettent d'adapter et de proposer des contenus appropriés qui correspondent aux besoins spécifiques de l'utilisateur (Gauch et al., 2007).

Ces différents systèmes de personnalisation d'informations, utilisent des mécanismes qui leur sont propres. Tous ces systèmes ont un point commun : l'utilisation d'un profil utilisateur, point central de nos travaux. La section suivante donne une définition du profil utilisateur à partir des travaux de la littérature.

2.1.1. Définition du profil utilisateur

Le profil utilisateur dans le contexte des systèmes de personnalisation d'informations, peut être défini comme une structure qui permet de modéliser et stocker des informations relatives à l'utilisateur. Le profil utilisateur peut contenir (Brusilovsky, 1996) :

- **ses données personnelles** telles que, son identité (nom, prénom, etc.), ses données démographiques (âge, genre, adresse, situation familiale, etc.), ses données professionnelles.
- **son historique/ feedbacks** qui regroupe l'ensemble des informations collectées sur son comportement, de façon explicite ou implicite (par exemple, le nombre de clics qu'il a effectués sur le lien d'une page ou le nombre des requêtes qu'il a émises, etc.).
- **les annotations** associées par l'utilisateur aux documents, peuvent être sous différentes formes (par exemple, les annotations textuelles, les signets qui mémorisent les liens vers d'autres documents, les tags, qui sont les références sous forme d'un ensemble de mots-clés choisis librement par l'utilisateur pour identifier le document visité…).
- **ses préférences** qui désignent les caractéristiques de l'utilisateur, en termes de présentations ou d'interactions avec les informations (par exemple, des couleurs et/ou les styles de présentation de page web préférés, etc.).
- **ses intérêts** qui expriment ses domaines d'expertise ou son périmètre d'exploration. Ils sont généralement définis par un ensemble de mots clés ou concepts, le plus souvent pondérés.

Les données personnelles sont relativement stables dans le temps et ne demandent pas a priori de mise à jour automatique, alors que les préférences et les intérêts tendent à changer au fil du temps.

Un profil utilisateur a pour objectif de permettre à un système de s'adapter à l'utilisateur. La section suivante explique cet aspect, dans le contexte des systèmes de personnalisation d'informations.

2.1.2. Utilisation du profil utilisateur dans le contexte de la personnalisation d'informations

On a pu constater ces dernières années, une augmentation exponentielle des données dans les systèmes d'informations et dans le web en général. Ce grand volume de données provient principalement de l'évolution des technologies du web 2.0 qui permet l'interactivité entre les utilisateurs. Il devient donc de plus en plus difficile pour les utilisateurs, de retrouver les informations qui correspondent précisément à leurs besoins dans cette masse de données. Ce problème est reconnu comme le problème de la surcharge cognitive de l'utilisateur. En outre, l'utilisateur peut se retrouver face au problème de la désorientation (il ne sait plus quel chemin suivre lors de la navigation via une interface utilisateur) pour trouver des informations qui correspondent vraiment à ses besoins. Les systèmes de personnalisation d'informations, ont pour but de pallier ces problèmes de surcharge cognitive et de désorientation de l'utilisateur. Pour cela, ils utilisent un profil de l'utilisateur (cf. Figure 2.1).

Un système de personnalisation d'informations, est un système qui intègre les informations de l'utilisateur pendant l'accès aux informations, afin de délivrer des contenus pertinents en fonction de caractéristiques et besoins spécifiques de l'utilisateur.

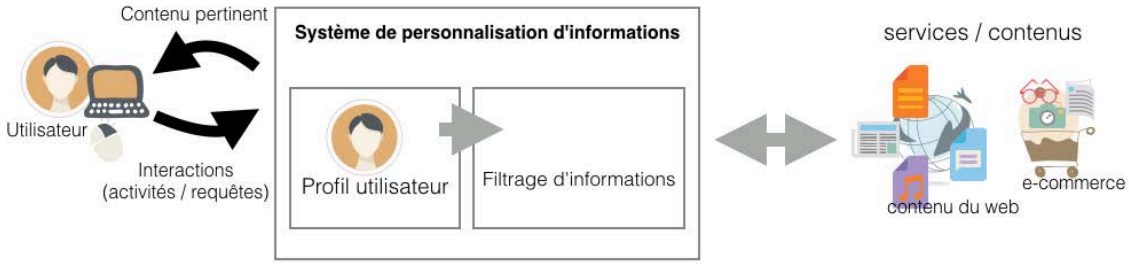


Figure 2.1 L'utilisation du profil utilisateur dans un système de personnalisation d'informations

L'efficacité du système de personnalisation, peut être montrée grâce à la pertinence des contenus qu'il propose, selon les besoins de l'utilisateur. Nous citons dans ce cas deux types de pertinence : **la pertinence utilisateur** qui désigne l'ensemble des jugements de l'utilisateur et **la pertinence système** qui désigne la vision que le système a de la pertinence du contenu proposé » (Denos, 1997).

Pour avoir un système efficace de personnalisation d'informations, la pertinence système devrait se rapprocher de la pertinence utilisateur. Pour ce faire, il s'avère important que le système intègre les informations individuelles de l'utilisateur (ses intérêts, ses préférences, …) et ses besoins vis-à-vis de l'information. Pour y arriver, deux stratégies de personnalisation sont envisageables (Chevalier, 2011):

- **le paramétrage** ou « *customization* » qui permet à l'utilisateur, de configurer les systèmes relatifs à des options ou préférences proposées initialement par le système (par exemple pour Google : le thème de l'interface, la langue de l'interface, le nombre de résultats par page...). Cette stratégie nécessite une démarche active de l'utilisateur puisqu'il est à l'origine de cette configuration. Cette dernière est cependant limitée aux seules options proposées par le système,
- **le profilage** (Cho, Kim et Kim, 2002) qui permet au système de construire une connaissance de l'utilisateur (intérêts, préférences, …) qu'il acquiert principalement au travers de l'interaction avec celui-ci.

Dans ce mémoire, nous abordons uniquement la stratégie de profilage qui offre le plus de possibilités d'acquérir les caractéristiques de l'utilisateur. Nous assimilons le terme profilage avec la construction du profil utilisateur, qui comprend les différentes étapes d'acquisitions d'informations et d'extraction des connaissances de l'utilisateur.

Avant d'aborder la construction du profil utilisateur dans la section qui suit, nous présentons d'abord l'utilisation du profil utilisateur dans les mécanismes de personnalisation d'informations : la recommandation et la recherche d'information personnalisée.

2.1.2.1. Utilisation du profil utilisateur dans un système de recommandation

Un système de recommandation consiste à fournir à un utilisateur des ressources (contenu ou items) pertinentes en fonction de ses intérêts ou besoins spécifiques. Il permet donc de diminuer le temps de recherche de ressources, d'aider les utilisateurs indécis lors de la sélection de ressources, mais également, de suggérer de nouvelles ressources qui pourraient les intéresser. Ces systèmes sont largement utilisés dans les sites de e-commerce comme Amazon[1] qui recommandent les produits appropriés à l'utilisateur, parmi la très importante variété de produits disponibles. Il existe aussi des systèmes de recommandations de « news », qui consistent à sélectionner pour un utilisateur, des nouvelles intéressantes parmi les nombreuses publiées sur le web. Il existe également, les systèmes de recommandation de sites de références d'articles comme *CiteSeerX*[2], les sites de recommandation de films comme *MovieLens*[3].

Les systèmes de recommandations s'appuient fondamentalement sur le filtrage d'informations (*Information filtering*), qui a pour objectif d'identifier dans un flux documentaire, les documents correspondants aux intérêts d'un utilisateur (profil utilisateur) (Belkin et Croft, 1992 ; Hanani, Shapira et Shoval, 2001).

On peut distinguer trois types de systèmes de recommandation : ceux basés sur le filtrage par contenu, ceux basés sur le filtrage collaboratif et ceux basés sur le filtrage hybride.

a. Système de recommandation basé sur le filtrage par contenu

Le filtrage basé sur le contenu consiste à comparer le profil des documents/items avec le profil de l'utilisateur, afin de filtrer les documents/items pertinents selon leur similarité avec le profil de l'utilisateur (Lops, Gemmis et Semeraro, 2011 ; Pazzani et Billsus, 2007). Différentes fonctions de similarité peuvent être appliquées. La fonction la plus utilisée est le cosinus de similarité qui mesure le cosinus de l'angle entre le vecteur représentant le profil de l'utilisateur et le vecteur des documents/items (Adomavicius et al., 2005).

Les systèmes de recommandation basés sur le contenu, consistent à proposer à l'utilisateur des nouveaux contenus ou items, en s'appuyant sur les évaluations positives ou négatives de l'utilisateur, effectuées auparavant sur un ensemble de documents/items (cf. Figure 2.2).

[1] www.amazon.com
[2] citeseerx.ist.psu.edu
[3] movielens.org

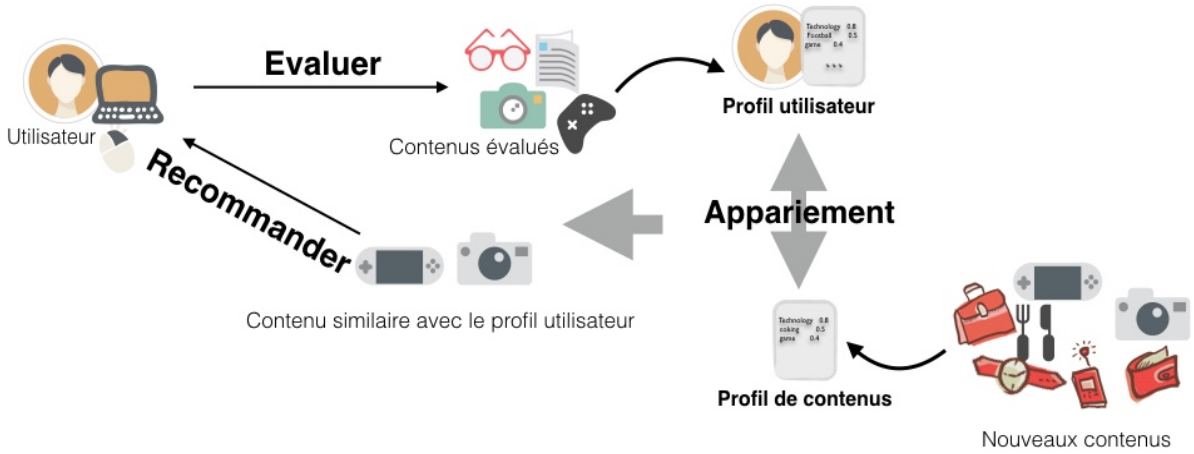


Figure 2.2 Recommandation par contenu

b. Système de recommandation basé sur le filtrage collaboratif

Le filtrage collaboratif consiste à utiliser les évaluations des autres utilisateurs dans le système, pour proposer des ressources appropriées pour l'utilisateur. Cette technique est introduite dans le contexte de la recommandation collaborative (Ekstrand, Riedl et Konstan, 2011). Avec cette technique, seuls les items bien évalués par les utilisateurs peuvent être recommandés.

La littérature distingue deux approches différentes de filtrage collaboratif : le filtrage collaboratif centré sur l'utilisateur et le filtrage collaboratifs centré sur les items.

Le filtrage collaboratif centré sur l'utilisateur (Resnick et al., 1994) est la technique la plus utilisée dans la littérature. Il s'agit de construire une matrice Utilisateurs*Items contenant les poids attribués par chaque utilisateur U sur chaque item I du système. Une difficulté rencontrée par cette technique, est que la matrice construite est généralement éparse car le ratio entre le nombre d'items notés par un utilisateur et le nombre total des items peut être très faible. Une autre difficulté, importante cette fois, est la complexité de calcul lors du passage à l'échelle, car le nombre de calculs sur cette matrice augmente linéairement en fonction du nombre d'items et du nombre d'utilisateur du système. Pour pallier ces problèmes, l'indexation des similarités entre utilisateurs par matrices Utilisateur*Utilisateurs est proposée. Il s'agit de rechercher des individus ayant un profil similaire au profil de l'utilisateur et ainsi, lui recommander uniquement les items qui sont bien évalués par les individus similaires (cf. Figure 2.3 a). Toutefois, le problème pour ce type de recommandation, est la construction du profil des utilisateurs. Elle nécessite de connaitre le profil de l'utilisateur et celui des autres individus pour pouvoir calculer leur similarité.

Le filtrage collaboratif centré sur les items (Sarwar et al., 2001) consiste à construire une matrice Items*Items. Au lieu d'exploiter la similarité entre les utilisateurs, cette technique utilise la similarité entre les items, en comparant l'évaluation de ces items. L'évaluation des items, peut se mesurer par exemple, par des votes (« j'aime », je n'aime pas »), des poids (notes données). On peut simplifier le principe de cette technique par « les utilisateurs qui aiment l'item **x** comme l'utilisateur **u** aiment également l'item **y** donc on recommande l'item **y** à cet utilisateur **u** » (cf. Figure 2.3 b).

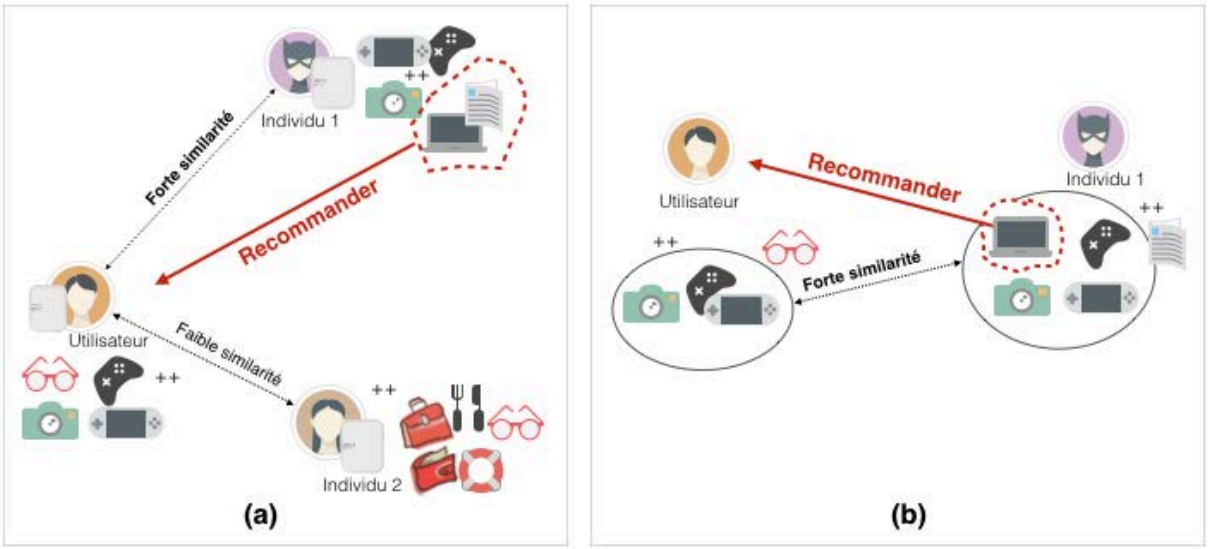


Figure 2.3 (a) Filtrage collaboratif basé sur les utilisateurs. (b) Filtrage collaboratif basé sur le contenu.

Après le filtrage par contenu et le filtrage collaboratif, qui mettent le focus soit sur les contenus, soit sur les utilisateurs, la section suivante étudie le filtrage hybride dont l'objectif est de combiner ces deux filtrages.

c. Système de recommandation basé sur le filtrage hybride (hybrid filtering)

Pour pallier les limites du filtrage basé sur le contenu et celles du filtrage collaboratif, le filtrage hybride (*hybrid filtering*) combine ces deux derniers en gardant les avantages de chacun (Burke, 2002 ; Godoy et Amandi, 2008 ; Melville, Mooney et Nagarajan, 2002). Selon (Adomavicius et al., 2005 ; Burke, 2002), le filtrage hybride peut se faire de différentes manières, par exemple :

- Implémenter chacune des méthodes séparément et combiner leurs résultats en une seule liste,
- Présenter les résultats des différentes méthodes sans forcément les combiner dans une seule liste,
- Choisir l'algorithme qui donne de meilleurs résultats selon le contexte particulier,
- Rajouter certaines caractéristiques du filtrage collaboratif dans le filtrage par contenus ou vice-versa,
- Utiliser la sortie d'un algorithme comme entrée d'un autre algorithme
- Construire un modèle unifié incorporant les caractéristiques de chacune des méthodes.

Dans la section qui suit, nous abordons l'utilisation du profil utilisateur dans le contexte de la recherche d'information personnalisée.

2.1.2.2. Utilisation du profil utilisateur dans un système de recherche d'information personnalisée

a. Recherche d'information

Un système de recherche d'information est *un système qui permet de retourner à partir d'un ensemble de documents, ceux dont le contenu correspond le mieux à un besoin en informations d'un utilisateur, exprimé à l'aide d'une requête (Daoud, 2009)*. La recherche d'information (RI) regroupe des techniques qui permettent l'acquisition, l'organisation, le stockage, la recherche et la sélection d'informations (Rijsbergen, 1979 ; Salton et McGill, 1986). Les principales phrases de la recherche d'information présentées dans la Figure 2.4 sont : l'indexation des documents, l'indexation des requêtes et l'appariement entre la requête et les documents afin de sélectionner et d'ordonner (*ranking*) les documents à présenter à l'utilisateur. L'indexation consiste à déterminer et à extraire les termes représentatifs du contenu d'un document ou d'une requête qui couvrent au mieux leur contenu sémantique. Cela permet de retrouver rapidement les documents contenant les mots clés de la requête. L'appariement consiste à calculer la pertinence de chaque document vis-à-vis de la requête donnée par l'utilisateur selon une mesure de correspondance du modèle de RI, avant de retourner la liste des résultats à l'utilisateur.

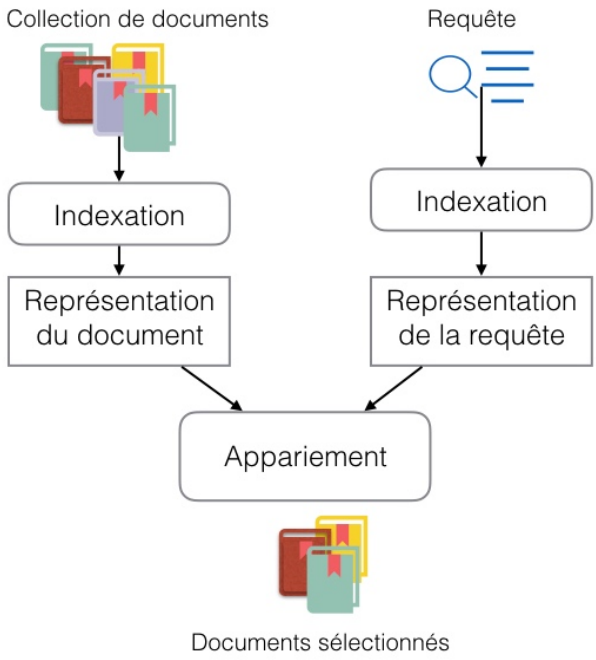


Figure 2.4 Processus de RI

b. Recherche d'information et personnalisation

Avec la masse de données produites dans le web 2.0, il devient difficile pour les systèmes de recherche d'information (SRI) traditionnels, de satisfaire les besoins spécifiques des utilisateurs.

Premièrement, face au grand volume d'information, les SRI classiques peuvent retourner une liste importante de documents, ayant tous des estimations de pertinence élevées par rapport à la requête donnée. Par conséquent l'utilisateur se trouve face à une surcharge informationnelle qui le désoriente et dans laquelle il doit trouver seul ce qui est pertinent de ce qui ne l'est pas.

Deuxièmement, il peut exister des ambigüités dans les requêtes données par l'utilisateur. Par exemple, pour la même requête « *Orange* » dans un moteur de recherche, deux utilisateurs peuvent s'attendre à des résultats de différents domaines. L'un peut avoir l'intention de chercher les informations liées au fruit et l'autre peut vouloir chercher des informations liées à la marque du même nom.

La RI personnalisée a pour objectif de pallier ce problème en considérant le profil de l'utilisateur. Le but fondamental de la RI personnalisée est d'exploiter des informations concernant l'utilisateur, en plus de la requête donnée, pour sélectionner les contenus correspondant aux besoins spécifiques de l'utilisateur. Nous citons ici les techniques d'exploitation du profil utilisateur pour la reformulation de requête et l'exploitation du profil utilisateur, pour la sélection de l'information et l'ordonnancement des résultats de recherche (Audeh, 2014 ; Daoud, 2009 ; Liu, Yu et Meng, 2004 ; Xu et al., 2007 ; Zemirli, 2008) (cf. Figure 2.5) :

- **La sélection d'information personnalisée** consiste à intégrer les informations du profil utilisateur pendant l'étape de l'appariement entre la requête de l'utilisateur et chaque document indexé.
- **La reformulation de requête** consiste à introduire dans la requête de l'utilisateur, les termes ou une partie des termes provenant du profil utilisateur.
- **L'ordonnancement des résultats** consiste à intégrer les informations du profil utilisateur pour réordonnancer les résultats trouvés après l'étape d'appariement requête/documents.

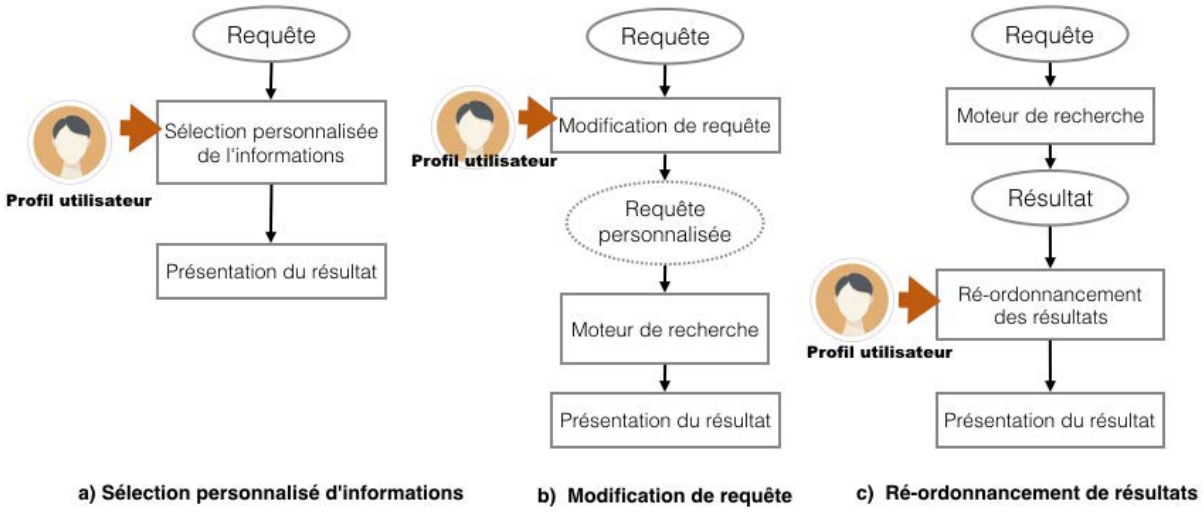


Figure 2.5 Intégration du profil utilisateur dans un système de RI personnalisée

Ces trois techniques d'intégration du profil, produisent de meilleurs résultats qu'un système de RI classique.

Avec l'exploitation d'un profil utilisateur, la RI personnalisée, constitue l'un des fondements du filtrage d'informations basé sur le contenu. En effet, ce dernier est souvent considéré comme une évolution du filtrage d'information issu du domaine de la recherche d'information (Berrut et Denos, 2003).

Le filtrage basé sur le contenu, se base uniquement sur les évaluations de l'utilisateur actif et non sur les autres utilisateurs similaires comme dans le cas du filtrage collaboratif. La

recommandation/personnalisation est donc plus explicite et ne nécessite pas des informations sur d'autres utilisateurs. On peut donc recommander des produits qui ne sont jamais évalués par d'autres utilisateurs, ce qui n'est pas le cas des systèmes utilisant le filtrage collaboratif (Lops, Gemmis et Semeraro, 2011).

Cependant, le filtrage par contenu a besoin de connaitre l'utilisateur ; c'est-à-dire de disposer des évaluations qu'il a émises sur des contenus dans le système. Plus l'utilisateur va utiliser le système, plus la pertinence des ressources qui lui seront proposées sera améliorée. Cependant, un utilisateur ne se verra jamais proposer d'items qui sont trop différents de ses précédentes évaluations d'items (Lops, Gemmis et Semeraro, 2011). Par exemple, un utilisateur ne s'intéressant qu'aux articles sur l'informatique, ne se verra jamais proposer des articles de sport. Avec cette technique, on rencontre également un problème pour les nouveaux utilisateurs arrivant dans le système. Un utilisateur qui n'a jamais interagi avec le système ne se verra pas proposer d'items pertinents car le système manque d'informations sur ses intérêts pour sélectionner des items. Ce problème est connu comme le problème de démarrage à froid (*cold start problem*) (Massa et Avesani, 2007).

Les approches de filtrage collaboratif offrent l'avantage de ne pas avoir besoin de disposer d'informations spécifiques sur les items, qu'elles recommandent par rapport à l'approche basée sur le contenu (profil d'item). En effet, il suffit de connaitre la similarité entre les utilisateurs ou entre les items pour sélectionner les items à recommander.

Cependant, quelle que soit la technique de filtrage collaboratif, le problème de démarrage à froid demeure. Pour les nouveaux utilisateurs ou ceux peu actifs, qui n'ont jamais attribué de score sur des items, le système ne dispose que d'un profil vide. Il n'est donc pas possible de trouver des individus qui leur sont similaires et donc impossible d'effectuer des recommandations. Dans le cas d'une faible activité, les recommandations peuvent être peu intéressantes pour l'utilisateur, car le profil peut ne pas être assez précis. Le même problème se pose dans le cas du filtrage collaboratif centré sur les items. Pour les items ayant très peu ou pas de scores attribués par les utilisateurs, il n'est pas possible de trouver les items qui leur sont similaires pour effectuer les recommandations.

Le filtrage hybride identifie plus efficacement la relation entre l'utilisateur et les items, afin de fournir des recommandations plus pertinentes. Cependant, cette approche peut s'avérer complexe et couteuse en temps de calcul (Su et Khoshgoftaar, 2009).

L'efficacité des approches de filtrage d'information présentées ci-dessus, dépend pour beaucoup, de la pertinence du profil utilisateur. Elle dépend aussi des mécanismes d'appariement utilisés (profil/information par exemple) mais ce point, hors du contexte de cette thèse, ne sera pas abordé.

Nous présentons dans la section qui suit, l'état de l'art sur la construction du profil utilisateur.

2.2. Méthodologie de construction du profil utilisateur

Indépendamment de son modèle de représentation, la construction du profil utilisateur repose sur deux phases principales : la phase de collecte des sources d'informations et la phase d'exploitation de ces sources d'informations pour construire et représenter le profil utilisateur avant son utilisation par des techniques de personnalisation d'information (Figure 2.6).

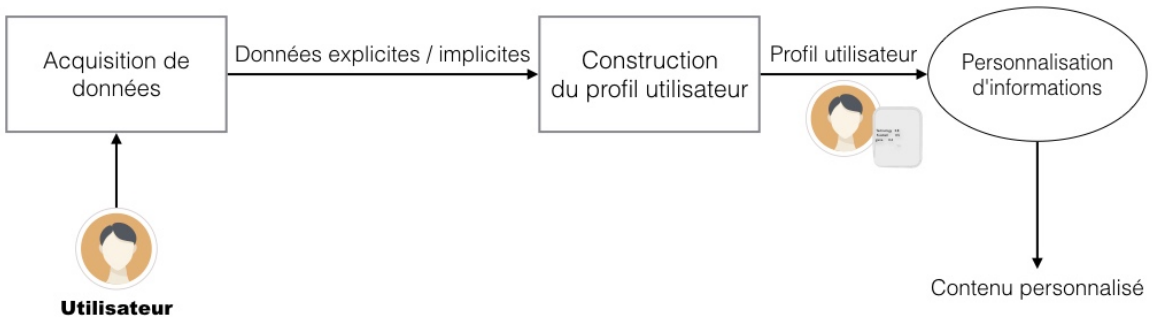


Figure 2.6 Les phases de construction du profil utilisateur

Nous présentons, dans un premier temps, les techniques d'acquisition des sources d'informations et le prétraitement des données. Ensuite, nous présentons les techniques de construction du profil selon différents modèles de représentation du profil utilisateur.

2.2.1. Acquisition des données

Nous pouvons distinguer deux façons d'acquérir les données : l'acquisition de données explicites et l'acquisition de donnée implicites (Gauch et al., 2007). Ces deux approches sont détaillées dans les deux sous-sections suivantes.

2.2.1.1. Acquisition des données explicites

La technique d'acquisition de données explicites est une technique simple, qui consiste à interroger l'utilisateur, pour lui demander des informations personnelles, démographiques et/ou ses intérêts (Gauch et al., 2007).

Cela peut se faire en demandant à l'utilisateur de remplir un formulaire d'informations personnelles pour construire son profil comme le demande par exemple MyYahoo![4]. On peut également considérer le feedback explicite de l'utilisateur, exprimé lors de son interaction avec le système, par exemple les notes de l'utilisateur sur les documents trouvés, les films regardés ou les produits achetés sur Internet, ou encore son choix de recommander tels articles ou tels produits à d'autres utilisateurs.

Les données explicites peuvent être directement intégrées dans un profil utilisateur et ainsi, être exploitées par des mécanismes de personnalisation. Cependant, cette technique d'acquisition de données peut entrainer une surcharge cognitive pour l'utilisateur, en particulier suite aux demandes de jugement répétitives et fréquentes. Cela peut entrainer également un désintéressement et un possible abandon de l'utilisateur, amenant ensuite la réduction de pertinence du système. Enfin, la pertinence du profil dépend du degré d'implication de l'utilisateur pour fournir des réponses exactes et complètes.

2.2.1.2. Acquisition des données implicites

Les limitations dans l'acquisition des données explicites, ont orienté les travaux vers des techniques d'acquisition des données implicites de l'utilisateur. Il s'agit dans ce cas, de ne plus demander à l'utilisateur de fournir explicitement ses données, mais de trouver des sources et des données tierces, permettant d'extraire des connaissances sur l'utilisateur et de construire son profil.

[4] https://my.yahoo.com

Deux questions principales se posent alors et sont explorées dans les sous-sections suivantes : quelles données peuvent être exploitées puisqu'elles ne sont pas explicitement demandées à l'utilisateur ? Et qui produit ces données ?

a. Données disponibles

Les données utilisées dans cette approche, sont les données collectées en observant le comportement et/ou en extrayant les informations des utilisateurs au travers de leurs activités. D'après (Kelly et Teevan, 2003), ces données sont aussi variées que : les documents propres à l'utilisateur, les requêtes et documents sélectionnés lors de l'utilisation d'un moteur de recherche, les pages Web consultées, les contenus publiés sur le web ou sur les réseaux sociaux (annotations, commentaires), les fichiers logs sur les applications telles que les applications de messagerie, les fichiers logs sur les consultations de bases de données, etc. Ces données sont généralement utilisées pour extraire les intérêts de l'utilisateur, dans la mesure où elles ont « un lien direct » avec lui puisqu'il les a consultées ou produites.

b. Producteurs de données

Dans la littérature, les données implicites peuvent être produites soit par l'utilisateur lui-même, soit par d'autres utilisateurs. Dans ce dernier cas, ce sont des individus considérés comme « proches » de l'utilisateur : les individus similaires (Ekstrand, Riedl et Konstan, 2011 ; Resnick et al., 1994) ou les individus dans son réseau social (Carmel et al., 2009 ; Guy et al., 2009 ; Wen et Lin, 2011 ; Zhang et al., 2010). Nous analysons ces trois cas dans les paragraphes qui suivent.

- **Données produites par l'utilisateur.** L'acquisition des données à partir des données produites par l'utilisateur est le cas le plus courant dans la littérature. Toutefois, ce type d'acquisition peut poser des problèmes lorsque le système dispose de très peu de données générées par l'utilisateur. C'est souvent le cas des nouveaux utilisateurs ou des utilisateurs peu actifs comme évoqué dans (Zhang et al., 2010).
- **Données produites par les individus similaires à l'utilisateur.** Dans ce cas, les données à acquérir sont extraites des données produites par les utilisateurs similaires à l'utilisateur courant. La similarité entre les utilisateurs est généralement calculée par croisement (cosinus de similarité par exemple) du profil de l'utilisateur avec ceux de tous les autres utilisateurs du système (Gao, Liu et Wu, 2010). Ensuite, les informations des individus similaires sont utilisées, pour calculer les informations ou les intérêts de l'utilisateur. Ce principe est à la base des différentes techniques de filtrage collaboratif (section 2.1.2.1.b). Cependant, cette technique nécessite beaucoup de temps de calcul pour des systèmes qui possèdent un nombre très élevé d'utilisateurs (temps de calcul de similarités entre les utilisateurs). De plus, elle ne peut pas être exploitée efficacement pour les utilisateurs ayant un profil vide ou très pauvre (nouvel utilisateur dans le système par exemple), car il devient impossible de retrouver des individus similaires à ce dernier.
- **Données produites par les individus dans le réseau social de l'utilisateur.** Dans ce cas, les données à acquérir sont extraites des données produites par les utilisateurs en lien avec l'utilisateur concerné dans un réseau social. Un réseau social est un graphe de relations entre individus. Les liens entre les individus dans le réseau social, représentent les relations entre eux et donc une certaine similarité. Cette approche se base sur les théories qui montrent qu'un utilisateur crée des relations avec ceux qui lui sont similaires (homophilie) car il partage avec eux des intérêts communs (Aral

et Walker, 2013 ; Carmel et al., 2009). Cette approche d'acquisition de données, restreint par construction le nombre d'individus similaires à l'utilisateur, en évitant d'explorer tous les utilisateurs du système. Ceci peut réduire drastiquement le temps de calcul de similarités entre l'utilisateur et les individus. De plus, cette approche peut être utilisée comme alternative, pour pallier les limites des méthodes précédentes, en cas de manque d'informations sur l'utilisateur (profil vide ou pauvre).

Le principal avantage de l'acquisition de données implicites est qu'elle ne nécessite aucune action explicite de la part de l'utilisateur. Cependant avec cette technique, on peut faire face au problème d'informations biaisées ou de manque d'informations. En effet, avec les données acquises sans vérification de la part de l'utilisateur, il se peut que ces dernières ne soient pas pertinentes pour lui (ex. les données que l'utilisateur produit par erreur ou qui contiennent des informations obsolètes). On peut également manquer d'informations importantes pour extraire des connaissances sur l'utilisateur par exemple, lorsque l'utilisateur n'interagit pas souvent avec le système, les données ne seront pas suffisantes pour extraire les préférences ou intérêts de cet utilisateur. Avec cette technique d'acquisition de données, il est donc nécessaire d'appliquer un prétraitement et un traitement de données qui s'avèrent plus complexes que la technique d'acquisition de données explicites. La problématique du prétraitement des données est abordée dans la section suivante.

2.2.1.3. Prétraitement des données

Les données issues de la sélection des données, comme expliqué précédemment, peuvent contenir de nombreuses inconsistances telles que : des données incomplètes (manque de valeurs ou attributs importants), des données biaisées (présence d'erreurs produites lors des saisies ou de la collection automatique de données), des incohérences (nommages ou codages différents dans les données, …). De plus, ces données brutes peuvent ne pas être conformes au modèle ou au format d'entrée de l'algorithme de construction du profil utilisateur. Après l'étape de sélection des données, il est nécessaire de mettre en œuvre une étape de prétraitement avant l'étape finale de construction du profil utilisateur.

On utilise plusieurs types de prétraitement selon les données (García, Luengo et Herrera, 2015 ; Liu, 2007).

- Pour les données incomplètes, biaisées ou incohérentes, on peut appliquer des techniques de nettoyage de données, qui consistent à ignorer les données manquantes ou à utiliser la valeur moyenne d'un attribut en remplacement ou encore à utiliser la valeur la plus probable (formule bayésienne ou arbre de décision) en remplacement, etc.
- La discrétisation des données peut être appliquée pour convertir des attributs continus vers des attributs ordinaux.
- La réduction des données peut être appliquée pour obtenir une représentation réduite du jeu de données, plus petite en volume, mais qui produit (ou presque) les mêmes résultats analytiques.
- Pour rendre les données conformes au modèle ou à l'algorithme utilisé, on peut appliquer des techniques de transformation de données qui permettent par exemple, de ne conserver qu'un résumé d'un texte à partir d'un texte entier, de traduire un texte dans une autre langue …

Après la collecte et le prétraitement des données, celles-ci sont utilisées en entrée de la phase de construction du profil utilisateur présentée dans la section suivante.

2.2.2. Construction du profil utilisateur

La construction d'un profil utilisateur nécessite d'effectuer des analyses sur les données en fonction de la modélisation utilisateur mise en œuvre. En termes de type de modélisation, on peut distinguer deux grands types : la modélisation du comportement et la modélisation des intérêts.

La modélisation de comportement consiste à analyser les comportements des utilisateurs via ses interactions avec le système. Elle est généralement utilisée dans les services Web (ex. historique de navigation, transaction avec le serveur Web). Ce type de modélisation a pour but de prédire ou de déterminer les préférences ou les feedbacks de l'utilisateur (ex. déterminer les parcours de navigation récurrents, valider la pertinence des campagnes marketing).

La modélisation des intérêts consiste, à partir d'une analyse des données, à construire une liste représentant du point de vue du système, ce que sont les intérêts de l'utilisateur. Pour extraire les intérêts de l'utilisateur, on peut le faire de façon directe à partir des données explicites de l'utilisateur ou de façon implicite à partir des données collectées (cf. section 2.2.1).

Dans cette thèse, nous nous intéressons particulièrement à la modélisation des intérêts de l'utilisateur. Plus précisément on s'intéresse aux techniques d'extraction des intérêts de l'utilisateur à partir de ses informations implicites.

La construction d'un profil utilisateur nécessite aussi de décider d'un modèle de représentation du résultat produit. Nous retiendrons trois catégories de représentation : représentation ensembliste, représentation par réseau sémantique et représentation conceptuelle (Gauch et al., 2007). Nous présentons dans les sections suivantes les techniques d'extraction des intérêts et de construction du profil utilisateur selon les trois modèles de représentation du profil retenus.

2.2.2.1. Construction d'un profil utilisateur ensembliste

La représentation ensembliste fait partie des premières représentations du profil utilisateur qui ont été proposées et reste largement utilisée. Le profil est représenté par un ensemble de termes ou mots-clés, éventuellement pondérés. Notons que dans ce mémoire, nous utilisons indifféremment « terme » ou « mot-clé ». La pondération permet de moduler l'importance de chaque intérêt par rapport à tous les autres intérêts de l'ensemble. Chaque terme peut simplement représenter un intérêt (ex. football, tennis) ou une catégorie d'intérêt qui regroupe des intérêts qui sont dans le même domaine (ex. sport, culture). Les intérêts peuvent être structurés soit en une unique liste ou vecteur de termes pondérés (ou non) où chaque terme correspond à un intérêt (Armstrong et al., 1995), soit par un ensemble de vecteurs de termes pondérés (ou non) indépendants où chaque vecteur correspond à un domaine d'intérêt et contient les termes correspondant à ce domaine (Pazzani et Billsus, 1997).

La construction d'un profil ensembliste se base sur des techniques d'extraction des termes à partir des données récoltées. Généralement, l'extraction des termes comprend les phases de traitements automatisés suivantes : extraction des termes (segmentation), élimination des mots vides, normalisation et pondération.

- **Extraction de termes** : cette phase consiste à extraire/segmenter les textes issus des données récoltées en termes. La segmentation (*tokenization*) du texte est une première étape importante dans ce processus. Les termes appelés les « *tokens* » sont extraits en utilisant des délimiteurs tels que les espaces, les traits d'union, les ponctuations. Les délimiteurs peuvent être définis différemment selon le contexte (type de contenu du texte, langue utilisée dans le texte)

- **Élimination des mots vides** : cette phase consiste à enlever les termes non significatifs appelés mots vides (ex. pronoms, prépositions, …). La suppression de ces termes peut réduire de manière considérable la taille du corpus de termes à traiter. Cependant il peut parfois être difficile de définir l'ensemble des mots vides. Il se peut que certains termes considérés comme des mots vides fassent partie de termes importants pour le traitement de termes. Par exemple, le terme « *a* » dans le mot « *vitamine a* ».
- **Lemmatisation** : cette phase permet de regrouper les variantes d'un mot. En effet, dans un texte, il peut y avoir différentes formes d'un terme désignant le même sens. Le but de ce processus est de les représenter par un seul terme qui porte sur un concept commun. Par exemple, remplacement de termes « Algorithmes », « Algorithme », « Algorithmique », « Algo », « Algos » par le terme « Algorithmique ». La lemmatisation peut se faire par analyse grammaticale en utilisant un dictionnaire, par utilisation de règles de transformation de type condition-action (ex. l'algorithme de Porter (Porter, 2006)), en tronquant des suffixes à *n* caractères ou encore par la méthode des *n-grammes* utilisée souvent pour la langue chinoise (Mayfield et McNamee, 2003).
- **Pondération des termes** : les termes extraits sont ensuite pondérés selon la technique de pondération choisie. La pondération des intérêts est souvent basée sur la fonction *TF*IDF*(Robertson et Sparck Jones, 1988). Cette fonction est largement utilisée dans le domaine de la recherche d'information pour trouver les termes d'un document qui représentent le mieux son contenu sémantique.

  La fréquence d'un terme, notée *TF (Term Frequency)*, est une mesure relative à la fréquence de ce terme dans le document. Cette pondération est donc une pondération au niveau local. L'idée de cette mesure est que, plus le terme est fréquent dans un document, plus il est important. On trouve plusieurs variantes de cette mesure. *TF* d'un terme *t* dans un document *d* peut être calculée par le nombre brut d'occurrences de *t* dans le document *d* (formule (2.1)).

$$TF(t,d) = occurence\ (t,d) \qquad (2.1)$$

  Pour éviter les biais liés à la longueur du document on peut calculer le *TF* de *t* par le nombre d'occurrences de *t* dans *d* normalisé par la somme totale des occurrences de tous les termes *k* dans *d* (formule (2.2)).

$$TF(t,d) = \frac{occurence\ (t,d)}{\sum_k occurence\ (k,d)} \qquad (2.2)$$

  On peut également utiliser la fonction logarithmique d'occurrence suivante (formule (2.3)) :

$$TF(t,d) = 1 + log\big(occurence(t,d)\big) \qquad (2.3)$$

  Cette formule peut être appliquée pour réduire l'écart entre les termes qui apparaissent le plus souvent et les termes qui apparaissent le moins souvent dans le

document : le terme t_1 qui apparait 10 fois dans le document est plus pertinent par rapport au terme t_2 qui n'apparait qu'une seule fois. Cependant le terme t_1 n'est pas forcement 10 fois plus pertinent que le terme t_2.

La fréquence inverse du document, notée *IDF* (*Inverse Document Frequency*), mesure l'importance d'un terme dans toute la collection, ce qui représente la pondération globale du terme dans la collection. L'idée de cette mesure est que les termes qui apparaissent dans peu de documents de la collection sont plus représentatifs que ceux qui apparaissent dans tous les documents de la collection. Cette mesure peut être calculée par la formule (2.4) suivantes :

$$IDF(t, D) = log\ (\frac{|D|}{n_t})$$

(2.4)

Ou bien par la formule (2.5) suivante :

$$IDF(t, D) = log\ (\frac{|D|}{n_t + 1})$$

(2.5)

D représente l'ensemble des documents de la collection et n_t représente le nombre de document de *D* dans lesquels le terme t apparait.

Finalement, la fonction *TF*IDF* (Robertson et Sparck Jones, 1988) est le produit des deux mesures *TF* et *IDF*. Cette fonction permet de palier à la limite de la mesure *TF* au niveau de la pertinence globale.

La fonction *TF*IDF* trouve ses limites dans le cas d'un document qui comporte beaucoup de termes. Dans ce cas, des solutions dérivées telles que *TF*IDF réduction* et LSA (*Latent Semantic Analysis*) (Landauer, Foltz et Laham, 1998) peuvent être utilisées. Les techniques d'apprentissage telles que les machines à vecteurs de support (SVM) peuvent également être exploitées (Isozaki et Kazawa, 2002). La pondération probabiliste qui consiste à pondérer chaque terme via sa probabilité de pertinence pour l'utilisateur peut être également appliquée (Joachims, 1997).

Pour terminer, soulignons que même si les modèles de représentation ensembliste permettent de traduire une multiplicité d'intérêts de l'utilisateur, cette représentation manque parfois de structuration, de cohérence, de niveaux de généralités/spécificités et de relations de corrélation entre les divers intérêts de l'utilisateur. Toutefois, la représentation ensembliste du profil utilisateur possède l'avantage d'être simple à mettre en œuvre. De ce fait, elle est souvent utilisée et appliquée sur de grandes collections de documents.

2.2.2.2. Construction d'un profil utilisateur basé sur les réseaux sémantiques

Ce type de représentation consiste à enregistrer les intérêts de l'utilisateur dans un réseau sémantique dont les nœuds représentent un terme traduisant un intérêt de l'utilisateur et les liens entre les nœuds représentent la proximité sémantique entre les nœuds.

Cette représentation permet de résoudre le problème de la polysémie des termes (que l'on peut rencontrer dans la représentation ensembliste) en mettant en place des relations de corrélation sémantique entre les termes. La relation entre les nœuds peut être traduite par leur nombre de co-occurrences.

La technique de construction du profil par réseau sémantique, repose sur le même principe d'extraction de termes que celui de l'approche de construction du profil ensembliste. Ce qui différencie ces deux approches est la façon de représenter les termes. Au lieu d'ajouter des termes extraits dans un vecteur, on les ajoute sur le réseau des nœuds. Un nœud peut représenter un seul terme ou un concept et ses termes associés (par exemple le concept « Programmation » et ses termes associés : « Coder », « Développer », …).

2.2.2.3. Construction d'un profil utilisateur basé sur une représentation conceptuelle

La représentation conceptuelle d'un profil utilisateur, consiste à représenter les intérêts de l'utilisateur, par un réseau de nœuds conceptuels décrivant un domaine d'intérêts de l'utilisateur et de les relier entre eux en respectant la topologie des liens, définie dans des hiérarchies (essentiellement basée sur l'utilisation d'une l'ontologie). On obtient, un profil représenté sous forme d'une hiérarchie de concepts, grâce à l'association des intérêts de l'utilisateur aux concepts des domaines de l'ontologie (Gauch et al., 2007).

La représentation conceptuelle est similaire à la représentation par réseau sémantique, dans le sens où ces deux types de représentation sont basés sur des nœuds de termes reliés par des relations. Cependant, dans la représentation conceptuelle, les nœuds représentent des domaines abstraits plutôt que des termes spécifiques ou des ensembles de mots relatifs comme dans la représentation par réseau sémantique. De plus, les liens entre les concepts sont explicitement induits de l'ontologie concernée et le profil résultant inclura des relations informationnelles plus diverses et spécifiques. La représentation conceptuelle peut également être assimilée à la représentation ensembliste du fait que chaque concept décrivant un intérêt est représenté par un vecteur de termes pondérés où le poids traduit le degré d'intérêt de l'utilisateur pour ce concept (Gauch et al., 2007) .

Dans la littérature, plusieurs types de structures hiérarchiques et ressources sémantiques ont été définies et sont disponibles. Les plus simples sont construits sur la base d'une taxonomie de concepts ou d'un thesaurus de référence. Par exemple, les systèmes de (Guarino, Masolo et Vetere, 1999) utilisent l'ontologie *Sensus*, une taxonomie d'approximativement 70 000 nœuds, et un sous-ensemble de l'annuaire Yahoo! 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 telles que *OBIWAN (Ontology Based Informing Web Agent Navigation)* (Pretschner et Gauch, 1999), *Personae* (Tanudjaja et Mui, 2002).

Dans un contexte à grande échelle telle que le Web, la représentation conceptuelle peut engendrer certains problèmes d'hétérogénéité et de diversité des intérêts. D'ailleurs, les utilisateurs peuvent avoir différentes perceptions d'un même concept, cela peut engendrer des imprécisions lors de la représentation de l'utilisateur (Godoy, 2006).

Après avoir donné les éléments principaux des différentes techniques de construction et de représentation d'un profil utilisateur, nous allons nous intéresser, dans la section suivante, à la prise en compte de l'évolution du profil utilisateur.

2.3. Gestion de l'évolution du profil utilisateur

La gestion de l'évolution du profil utilisateur est un processus complémentaire à la construction du profil utilisateur. Elle désigne l'adaptation du profil à la variation des intérêt et aux variations des besoins en information de l'utilisateur au cours du temps (Zemirli, 2008). Dans cette section, nous nous intéressons particulièrement à l'évolution des intérêts de l'utilisateur. Ces derniers peuvent changer et devenir non pertinents dans le temps. Un intérêt jugé pertinent dans une période de temps peut devenir obsolète dans les périodes suivantes. L'adaptation du profil utilisateur à cette évolution implique des changements au niveau des intérêts décrits dans le profil qui conduisent éventuellement à la suppression d'intérêts existants, à l'émergence de nouveaux intérêts et dans tous les cas au (re)calcul des poids des intérêts.

Dans la littérature, nous pouvons distinguer deux approches principales de gestion de l'évolution du profil utilisateur. La première approche consiste à prendre en compte l'évolution des intérêts pendant l'étape de construction du profil lors de l'extraction des intérêts, donc durant la phase décrite en section 2.2.2. Cette approche revient à recalculer un nouveau profil pour prendre en compte l'évolution des intérêts de l'utilisateur. La deuxième approche consiste à gérer l'évolution des intérêts de l'utilisateur après la construction du profil utilisateur et correspond à un processus de mise à jour du profil utilisateur. Cette mise à jour est ensuite répétée. Les sections suivantes décrivent successivement ces deux approches.

2.3.1. Gestion de l'évolution des intérêts pendant l'étape de construction du profil utilisateur

Le problème de l'évolution des intérêts peut être traité en appliquant les techniques que l'on peut classifier en deux grandes approches : approche par sélection d'instance (*instance selection*) et approche pondérée (*instance weighting*). Les sections suivantes décrivent successivement ces deux approches.

2.3.1.1. Approche par sélection d'instance

L'approche par sélection d'instance (*instance selection*) sélectionne les informations pertinentes par rapport à une période de temps choisie. Ce raisonnement a été utilisé dans la construction du profil utilisateur dans plusieurs contextes. La plupart de ces travaux se basent sur l'usage d'une fenêtre temporelle (*time window*) qui décide d'un intervalle de temps dans lequel les informations sont considérées et qui ignore les autres informations (Cheng et al., 2008 ; Maloof et Michalski, 2000). Par exemple, en recherche d'information personnalisée, dans (Bennett et al., 2012), les auteurs utilisent l'historique à court terme de l'utilisateur lié à une seule session de recherche (la dernière) pour extraire ses intérêts. Les techniques utilisées dans l'approche de sélection d'instance oublient complètement les informations dépassant une date définie. Pourtant, certaines informations ignorées peuvent s'avérer pertinentes et ne pas les prendre en compte, peut entraîner une perte d'informations intéressantes. En effet, (Tan, Shen et Zhai, 2006) ont montré que l'historique de recherche à long terme est très important pour améliorer la tâche de recherche d'information dans le cas de requêtes récurrentes.

2.3.1.2. Approche pondérée

La seconde approche que l'on trouve dans la littérature, appelée approche pondérée (*instance weighting*), calcule le poids de chaque instance selon son poids de pertinence estimé en fonction du temps. C'est souvent une fonction temporelle (*time decay function*) qui est utilisée pour donner plus de poids aux informations les plus récentes. Dans ce type d'approche, toutes les

informations existantes peuvent être exploitées mais de manière différenciée. Ce raisonnement a été utilisé dans plusieurs travaux sur les systèmes de recommandation (Li et al., 2013). Cette idée peut être retrouvée également dans le contexte de la recherche d'information personnalisée comme dans (Kacem, Boughanem et Faiz, 2014) qui proposent d'appliquer une fonction temporelle pour pondérer les intérêts de l'utilisateur selon leur fraîcheur. Cette idée peut être retrouvée également dans le contexte de la construction du profil utilisateur à partir d'un réseau d'annotations comme dans (Zheng et Li, 2011) qui utilisent des fonctions temporelles pour pondérer des tags avant d'en extraire les intérêts de l'utilisateur.

Pour pondérer les informations selon leur fraicheur, plusieurs fonctions temporelles peuvent être appliquées. Nous décrivons ci-après, les trois principales fonctions de la littérature.

a. Fonction linéaire inverse

La façon la plus simple de calculer une pondération temporelle est d'appliquer une fonction linéaire inversement proportionnelle à la date d'apparition de chaque information (formule (2.6)).

$$f_{lin}(t) = \frac{1}{t+1} \qquad (2.6)$$

La valeur $t \in \mathbb{N}$ est la distance entre la date de publication de l'information et la date donnée (souvent c'est la date actuelle au moment du calcul). Par exemple, si la date du calcul est 2016, pour une information publiée en 2016, *t* vaut 0, pour celle publiée en 2015 *t* vaut 1, et ainsi de suite. En d'autres termes, *t* représente la fraicheur de l'information vis-à-vis de la date actuelle. Plus *t* est petit, plus l'information est récente.

Notons que l'addition de 1 dans l'équation permet d'éviter une division par 0, en particulier pour certaines échelles de date (par exemple le mois ou la semaine).

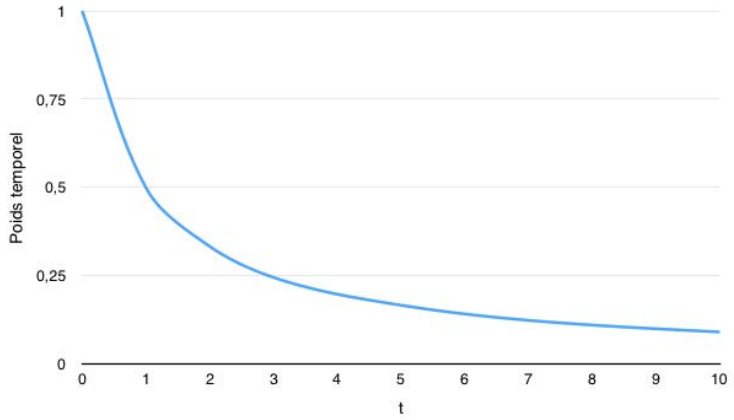


Figure 2.7 Le poids temporel de chaque valeur t en axe selon la fonction linéaire inverse

b. Fonction exponentielle

La fonction exponentielle est la fonction temporelle la plus utilisée dans les différentes applications. (Ding et Li, 2005) proposent une fonction exponentielle (formule (2.7)) pour pondérer les informations afin de les utiliser dans un système de recommandation temporelle.

$$f_{exp}(t) = e^{-\lambda t}$$

(2.7)

Comme dans la fonction linéaire inverse, la valeur $t \in N$ est la fraicheur de l'information vis-à-vis de la date actuelle. La valeur $\lambda \in [0,1]$ représente le taux de dépréciation (*Time Decay Rate*) des valeurs. Plus λ est grand, moins les informations anciennes sont importantes. λ est calculée en se basant sur la demi-vie T_0,

$$T_o = \frac{1}{2}(f(0))$$

(2.8)

T_0 représente le fait que le poids de l'information est réduit de moitié tous les T_0 instants (jour, année). λ est calculée à partir de la formule :

$$\lambda = \frac{1}{T_0}$$

(2.9)

La valeur optimale de T_0 sera définie expérimentalement (la valeur de λ est fixée expérimentalement)

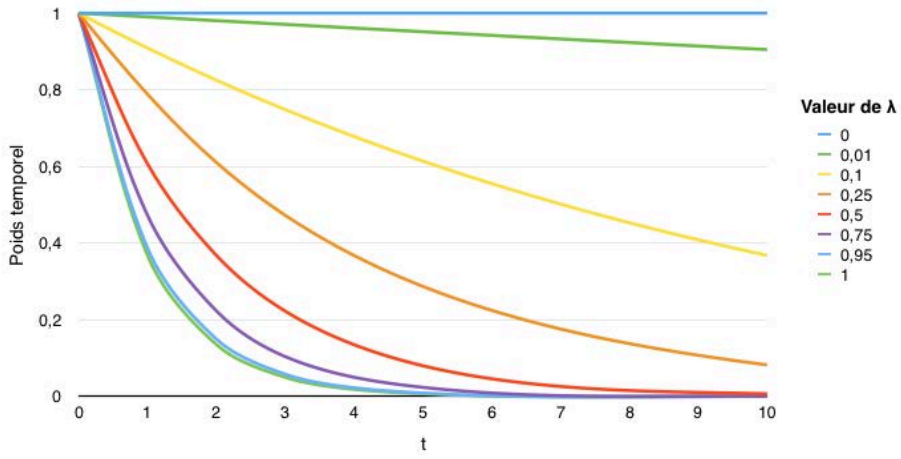


Figure 2.8 Le poids temporel de chaque valeur *t* en axe pour différentes valeurs de λ selon la fonction exponentielle

Une autre fonction temporelle basée sur la fonction exponentielle est exploitée dans le travail de (Zheng et Li, 2011). Dans le cadre de la recommandation basée sur des tags, les auteurs utilisent une fonction de score qui assigne un poids aux tags (considérés comme des intérêts) selon leur date de création. Plus le tag est récent, plus il est important. Cette fonction est représentée comme suit[5] :

[5] Nous avons modifié la notation dans la formule par rapport à la formule initiale présentée dans l'article pour unifier la représentation des fonctions temporelles dans ce mémoire

$$f_{expln}(t) = e^{\ln 2 * t/hl(u)}$$

(2.10)

Où t ∈N représente la distance entre la date où le tag a été annoté par l'utilisateur et la date actuelle, comme présenté précédemment dans la formule (2.7) En d'autres termes, *t* représente la fraicheur du tag t vis-à-vis de la date actuelle. *hl(u)* représente la demi-vie de l'utilisateur u (une vie est calculée selon la date de début et de fin de l'activité d'annotation de l'utilisateur).

c. Fonction polynomiale

La fonction polynomiale propose une décroissance polynomiale (*polynomial decay*) de la pondération calculée. Elle correspond à l'application de la formule (2.11) suivante (Cormode et al., 2009) :

$$f_{poli}(t) = (t+1)^{-\lambda}$$

(2.11)

De la même manière que dans les deux précédentes fonctions, la valeur *t* représente la fraîcheur de l'information. Pour chaque *t=i* (i∈*N*), *t=0* est considéré comme la valeur de fraicheur de l'instant le plus récent (ex : *t=0* pour l'année 2016, *t=1* pour 2015,…). La valeur λ∈[0,1] représente le taux de dépréciation (*Time Decay Rate*) des valeurs. Notons que l'addition (*t+1*) est utilisée pour assurer *f(0)* =1.

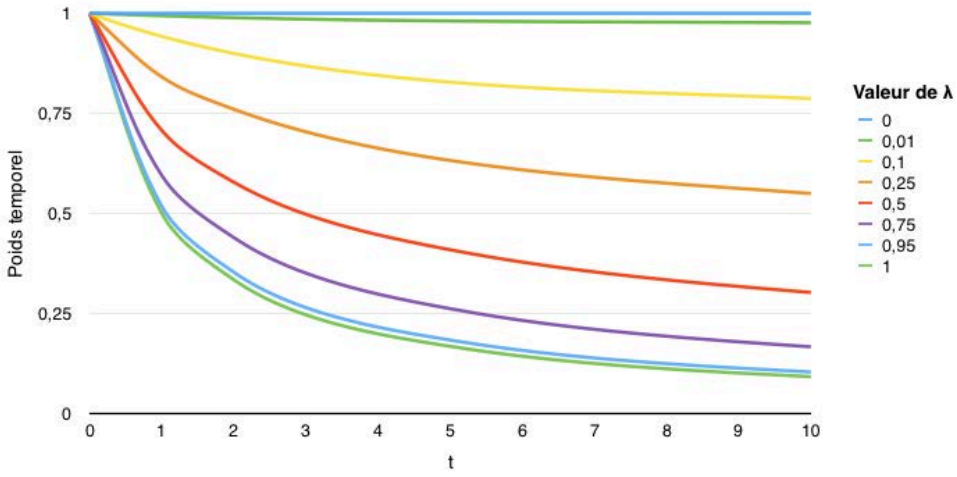


Figure 2.9 Le poids temporel en axe horizontal de chaque valeur *t* en axe vertical pour différentes valeurs de λ selon la fonction polynomiale

La section suivante présente une deuxième technique : la prise en compte de l'évolution des intérêts de l'utilisateur par une mise à jour du profil utilisateur et non plus par une (re)construction du profil.

2.3.2. Mise à jour du profil utilisateur et évolution des intérêts

L'approche par mise à jour du profil utilisateur consiste, à partir d'un profil utilisateur existant, à « ajuster » les intérêts dans le profil utilisateur selon les variations des intérêts réels ou les

besoins d'informations de l'utilisateur afin de conserver un profil à jour et pertinent. L'objectif est, sans recalculer l'ensemble du profil, de pouvoir retirer des intérêts du profil, en ajouter de nouveaux ou modifier des pondérations d'intérêts. La mise à jour du profil utilisateur permet également d'extraire de nouvelles informations et ainsi d'extraire et de compléter les intérêts existants dans le profil. Généralement, la mise à jour du profil utilisateur se fait en se basant sur le feedback de l'utilisateur de façon explicite ou implicite. Ces deux approches sont décrites dans les deux sous-sections suivantes.

2.3.2.1. Mise à jour explicite du profil utilisateur

La mise à jour explicite du profil utilisateur se base sur le feedback explicite de l'utilisateur. Plusieurs travaux (Papadogiorgaki et al., 2008 ; Pon et al., 2011 ; Wang et al., 2013) se basent sur l'algorithme de *Rocchio* (Rocchio, 1971) qui est utilisé en RI pour adapter la requête de l'utilisateur selon le retour (positif ou négatif) sur le résultat de ses précédentes recherches. Lors de la mise à jour du profil de l'utilisateur, l'ensemble de ses intérêts est actualisé selon ses jugements à propos des derniers contenus proposés. Les intérêts liés aux contenus que l'utilisateur a jugés pertinents seront ajoutés à son profil tandis que les intérêts liés aux contenus jugés non pertinents seront supprimés de son profil. Dans cette méthode, la participation de l'utilisateur est nécessaire.

2.3.2.2. Mise à jour implicite du profil utilisateur

La mise à jour implicite se fait automatiquement sans avoir besoin de feedback de l'utilisateur. Il s'agit donc de profiler l'utilisateur en prenant en compte l'aspect temporel. Les techniques citées précédemment dans la phase de construction du profil peuvent être appliquées/adaptées dans ce contexte :

- L'approche basée sur une fenêtre temporelle peut être utilisée pour sélectionner, à chaque mise à jour, seulement les informations du dernier intervalle de temps Δt, afin de ne conserver que les informations considérées les plus récentes et les plus importantes. Les informations anciennes seront donc exclues à chaque mise à jour du profil.
- L'approche pondérée peut être appliquée pour pondérer les informations dans la fenêtre temporelle choisie afin de privilégier les informations les plus récentes.

Nous citons par exemple, le travail de (Mezghani, 2015) dans notre équipe, qui propose une approche d'enrichissement temporel du profil utilisateur à partir de tags. Cette approche s'appuie principalement sur l'analyse du comportement d'annotations des utilisateurs dans une période de temps Δt pour sélectionner les tags les plus significatifs pour l'enrichissement du profil. Dans cette approche, le profil utilisateur est construit de façon implicite, en utilisant la liste des tags assignés par les utilisateurs. Le profil utilisateur est enrichi par des tags à chaque période de temps Δt. La division en période Δt a pour but de pouvoir analyser une partie des informations selon une période prédéfinie afin de réduire le spectre d'analyse et ainsi d'essayer de ne garder que les informations les plus représentatives pour une période donnée. Dans ce contexte, le choix du Δt est important. La taille de chaque Δt doit être cohérente avec la quantité de données.

L'enrichissement du profil est effectué à chaque Δt afin de refléter les intérêts actuels de l'utilisateur. Le processus d'enrichissement comprend trois principales étapes (cf. la Figure 2.10) :

- Etape 1 : calcul de la température des ressources : Le terme « ressource » dans ce contexte représente le contenu sur lequel les utilisateurs assignent (annotent) les tags.

Une ressource peut être une image, une URL, du texte. La température d'une ressource reflète sa popularité à un moment donné. Ce calcul permet de refléter l'importance d'une ressource pour un utilisateur donné dans chaque Δt. La température d'une ressource est calculée en combinant les trois paramètres suivants :

- **La fraîcheur des tags associés à la ressource** : plus les tags sont récents plus la ressource est intéressante pour l'utilisateur. La fraîcheur d'une ressource *r* est calculée avec la fonction suivante :

$$fraîcheur(r) = \frac{\sum_{i=1}^{h} \frac{1}{p1(t_i)}}{h} \quad (2.12)$$

  Où *h* représente le nombre de tags associés à la ressource *r*. $p1(t_i)$ représente la distance entre l'heure d'annotation du tag t_i et l'heure actuelle.

- **La similarité des utilisateurs** (qui ont annoté la ressource) : si deux utilisateurs ont annoté la même ressource avec des tags semblables, cela reflète leur similarité en termes d'intérêts. Ils sont donc considérés comme des personnes proches. La similarité cosinus est exploitée pour calculer la similarité entre deux utilisateurs.

- **La popularité** (de la ressource) qui est le nombre de tags associés à la ressource.

Pour une période Δt et étant donné une ressource *r*, les trois paramètres sont combinés pour obtenir la température $T_{\Delta t}(r)$ selon la formule suivante :

$$T_{\Delta t}(r) = \alpha * fraîcheur + \beta * similarité + \gamma * popularité \quad (2.13)$$

α, β et γ sont des constantes qui reflètent le degré d'influence de chaque paramètre et sont fixées dans l'expérimentation.

- Etape 2 : calcul du poids des tags. Après le calcul de la température de chaque ressource, seules les ressources dont les valeurs de température augmentent entre deux périodes successives de temps (*Δt*-1 et *Δt*), sont considérées. En fait, l'augmentation de la température reflète l'intérêt de l'utilisateur envers ces ressources. Ainsi, les auteurs proposent de garder les ressources annotées les plus pertinentes, en considérant seulement les métadonnées qui reflètent le contenu de chaque ressource telles que le titre, les mots-clés et la description de la ressource. L'étape suivante consiste à attribuer un poids pour les tags associés aux ressources. Ce poids est calculé selon le degré de correspondance de chaque tag avec les métadonnées de la ressource associée.
- Etape 3 : ajout des tags pertinents. Après le calcul du poids des tags associés aux ressources les plus pertinentes, le profil utilisateur est enrichi avec les tags les plus pertinents. Plus le tag a un poids important, plus il reflète le contenu de la ressource et donc les intérêts de l'utilisateur. Un tag est considéré comme un intérêt potentiel s'il a un poids supérieur à un certain seuil fixé lors de l'expérimentation.

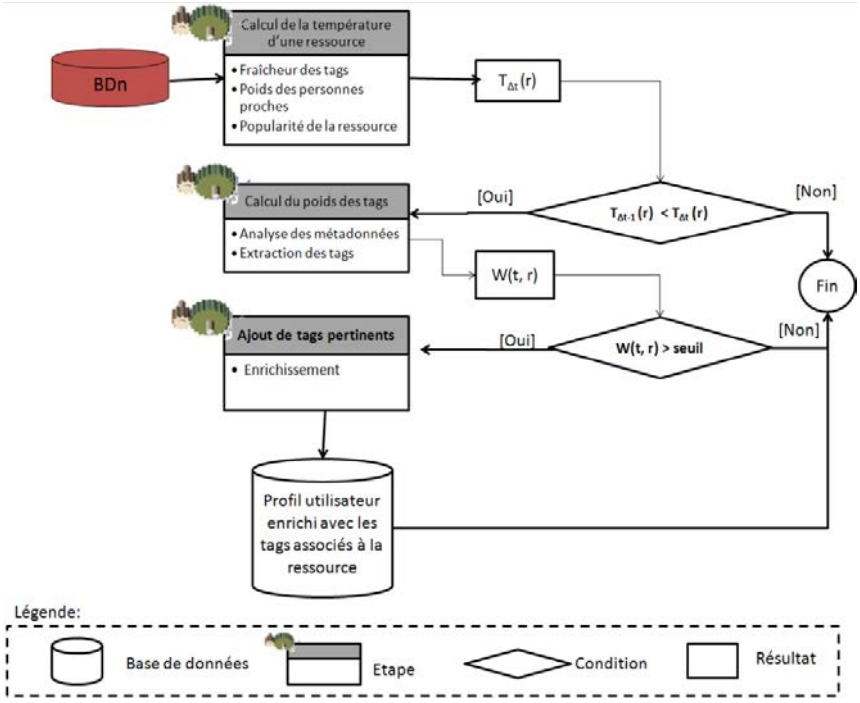


Figure 2.10 Le processus d'enrichissement du profil utilisateur social pour une période Δt (Mezghani, 2015)

Différents travaux proposent de faire cohabiter les informations récentes et les plus anciennes en privilégiant les informations liées à l'activité courante de l'utilisateur et les informations « à plus long terme » concernant l'utilisateur. Le principe est d'utiliser deux types de profils pour le même utilisateur :

- le profil à court terme qui contient les informations récentes liées à l'activité courante de l'utilisateur (dernière fenêtre temporelle),
- le profil à long terme qui contient des informations qui reflètent les besoins à long terme de l'utilisateur, généralement cumulées dans le temps.

Cette approche se retrouve souvent dans les travaux du domaine de RI qui se basent sur les sessions de recherche : considérer la dernière session de recherche pour extraire le profil à court terme et les informations des sessions précédentes pour extraire le profil à long terme (Li et al. 2007), (Zemirli et Tamine-lechani, 2007), (Sugiyama, Hatano et Yoshikawa, 2004). Pour montrer le principe, nous détaillons par la suite le travail de (Sugiyama, Hatano et Yoshikawa, 2004).

Dans le contexte de la RI, (Sugiyama, Hatano et Yoshikawa, 2004) proposent une approche pour adapter automatiquement les résultats de recherche de l'utilisateur selon ses besoins. Le principe du système proposé est de détecter automatiquement les changements de préférences de l'utilisateur, sans avoir besoin de son intervention. Pour cela, le système se base sur la sélection des résultats lors des précédentes recherches et de l'historique de navigation pour mettre à jour le profil utilisateur. Quand l'utilisateur soumet à nouveau une requête au moteur de recherche, le résultat de recherche sera adapté en se basant sur son profil utilisateur mis à jour (voir la Figure 2.11).

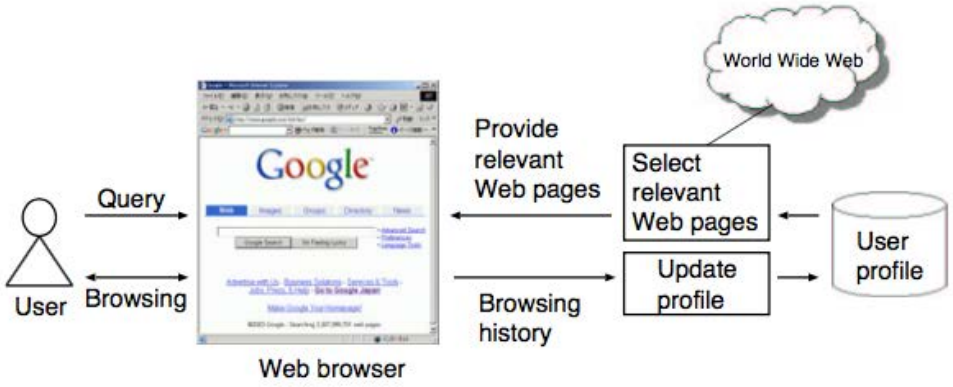


Figure 2.11 Vue globale du système de mise à jour du profil utilisateur proposé par (Sugiyama, Hatano et Yoshikawa, 2004)

Ce système distingue deux types de préférences : les préférences à long-terme et les préférences à court terme. Les préférences à long-terme sont cumulées par l'utilisateur dans le temps. Les préférences à court terme sont extraites à partir des activités de recherches actuelles. Le profil à long terme (persistant) noté P^{per} contient des préférences extraites de l'historique de navigation de l'utilisateur pendant les N jours précédents. Le poids de chacune de ces préférences est la somme des poids sur les N jours. Le poids des préférences pour chaque jour est calculé avec une fonction temporelle (*forgetting factor*) pour donner plus de poids aux préférences apparaissant plus récemment (cf. section 2.3.1.2.b). Le profil à court terme P^{today} contient des préférences à partir des activités de navigation du jour actuel. Le profil à court terme est lui, calculé en se basant sur les profils partiels construits le jour même : $P^{(cur)}$ se basent sur les activités de session actuelle et $P^{(br)}$ représente le profil contenant des préférences qui se basent sur les activités des sessions précédant la session actuelle. Finalement le profil à court terme est obtenu en combinant le $P^{(br)}$ et le $P^{(cur)}$ avec la formule (2.14) suivante :

$$P^{today} = x\,P^{br} + y\,P^{cur} \qquad (2.14)$$

Où x et y sont des constantes tels que $x + y = 1$. Les auteurs attribuent plus de poids à y qu'à x pour donner plus d'importance à la session de recherche actuelle.

Finalement, le profil à long terme P^{per} et le profil à court terme P^{today} sont combinés selon la formule (2.15):

$$P = a\,P^{per} + y\,P^{today} = a\,P^{per} + bx\,P^{br} + by\,P^{cur} \qquad (2.15)$$

2.4. Bilan

Dans cette partie, nous avons étudié la notion de profil utilisateur, son utilisation dans les systèmes de personnalisation d'informations ainsi que les approches de construction du profil utilisateur. Le défi dans la construction du profil utilisateur est d'obtenir un profil de qualité (Kobsa, 2001, 2007). La question principale est donc : comment construire un profil qui soit pertinent par rapport aux caractéristiques ou intérêts de l'utilisateur dans la vie réelle. Pour cela, des problématiques se posent au niveau des sources d'informations utilisées, des techniques de représentation et de construction employées mais également au niveau de la technique de prise en compte l'évolution des intérêts dans le profil utilisateur(Amato et Straccia, 1999). Notons que l'on ne trouve pas dans la littérature, de solution générale ayant des résultats meilleurs que les autres quelle que soit l'application, mais au contraire des combinaisons de techniques choisies ou ajustées selon l'application visée.

Dans le contexte du Web 2.0, l'utilisateur est entouré d'informations générées par lui-même mais aussi par les autres (individus similaires, individus dans son réseau social). Les informations disponibles dans l'écosystème de l'utilisateur représentent des sources d'informations qui peuvent être intéressantes à exploiter.

Dans ce mémoire, nous nous intéressons en particulier à l'utilisation des informations disponibles dans le (ou les) réseau(x) social(aux) de l'utilisateur.

A la différence de la technique qui prend en compte des utilisateurs ayant des profils similaires (filtrage collaboratif), cette technique utilise les relations que l'utilisateur a avec d'autres. Les relations de l'utilisateur sont basées sur le fait qu'il partage des intérêts ou des caractéristiques communs avec d'autres utilisateurs. L'utilisation des informations issues du réseau social de l'utilisateur permet donc, d'une part de chercher des informations complémentaires sur l'utilisateur, et d'autre part, de restreindre les sources d'informations puisqu'il n'est plus alors nécessaire d'accéder à l'ensemble des utilisateurs mais seulement aux utilisateurs qui sont en relation avec lui.

Malgré la différence des sources d'informations exploitées, l'approche de construction du profil utilisateur à partir de son réseau social, pose les mêmes questions que l'approche de construction du profil utilisateur classique (profilage classique) : comment construire un profil pertinent par rapport aux caractéristiques ou intérêts de l'utilisateur dans la vie réelle ? Comment prendre en compte l'évolution des intérêts dans le profil utilisateur ? Cependant ces problèmes ne seront plus uniquement centrés sur l'utilisateur mais également (et plutôt) sur d'autres individus qui font partie de son réseau social.

Une revue de la littérature nous a permis de constater que même si beaucoup de travaux existent dans le domaine, il reste encore de nombreuses pistes à explorer pour dériver l'information utile pour l'utilisateur à partir de son réseau social. Par exemple, comment intégrer de manière efficace le réseau social dans un profil utilisateur ? Comment sélectionner de manière efficiente les individus dans les réseaux sociaux qui seraient les plus discriminants pour caractériser l'utilisateur ? Comment exploiter de manière efficiente les informations provenant du réseau social de l'utilisateur ? Dans cette thèse, nous essayons, entre autres, de répondre à ces questions.

Nous présentons plus en détail dans le chapitre suivant ce qu'est un réseau social, l'acquisition des données dans un réseau social et la construction du profil utilisateur à partir de ces données.

3. CONSTRUCTION DU PROFIL UTILISATEUR A PARTIR DE SON RESEAU SOCIAL

Dans le chapitre précédent nous avons présenté les éléments fondamentaux du profil utilisateur ainsi que les techniques de construction du profil utilisateur dans un contexte classique (profilage classique). Dans ce chapitre, nous présentons les travaux associés à la construction du profil de l'utilisateur qui s'appuie sur les informations de son réseau social. Cette approche repose sur les travaux en sociologie concernant les liens se créant entre les utilisateurs en fonction de leurs points communs ou affinités (géographique, affiliation, intérêts).

Pour faciliter la rédaction et la compréhension, nous appelons « *profilage social* » la construction d'un profil de l'utilisateur à partir des informations contenues dans son réseau social. Ainsi, nous considérons deux termes différents pour distinguer les différents profils utilisateurs construits en fonction de la technique utilisée :

- nous appelons **profil utilisateur** ou plus précisément **profil utilisateur individuel,** le profil construit par profilage classique (cf. Chapitre 2),
- nous appelons **profil social** le profil construit par profilage social.

L'approche de profilage social ne sera pas centrée uniquement sur l'utilisateur mais également, et surtout, sur les individus de son réseau social. De ce fait, les travaux issus de cette approche reposent non seulement sur les principes classiques de la construction du profil utilisateur mais nécessitent également l'étude et l'utilisation de techniques d'analyse et de fouille de données

dans le réseau social de l'utilisateur. L'état de l'art sur cette partie portera donc sur deux axes d'étude : la construction du profil social qui s'appuie sur les méthodologies présentées dans la section précédente et l'analyse des réseaux sociaux.

Pour présenter le contexte de notre travail, nous présentons tout d'abord les éléments fondamentaux constitutifs d'un réseau social ainsi que les travaux sur l'analyse de réseaux sociaux associés à notre étude. Nous décrivons ensuite la notion de profilage social. Puis, nous présentons les travaux existants dans le domaine du profilage social avant de faire une synthèse pour amener nos positionnements et contributions par rapport aux travaux existants.

3.1. Réseau social

Afin de présenter les éléments fondamentaux des réseaux sociaux utilisés dans les travaux en profilage social, cette section donne une définition de ce qu'est un réseau social, puis étudie plusieurs typologies de réseaux sociaux, compare les réseaux sociaux numériques et les réseaux sociaux traditionnels, et finalement présente en détail les éléments qui constituent un réseau social.

3.1.1. Définitions

Réseau social : au sens large, un réseau social désigne un ensemble d'entités sociales (individus ou organisations) reliées entre elles par des liens créés lors d'interactions sociales (Wasserman et Faust, 1994).

Un exemple type de réseau social est les réseaux de chercheurs scientifiques qui sont souvent étudiés et exploités en tant qu'échantillon de test par le monde académique, et que nous utiliserons comme domaine d'expérimentation dans cette thèse. Les principaux modèles de réseaux de chercheurs scientifiques sont (Ding, 2011 ; Newman, 2001a, 2004a):

- **les réseaux de co-auteurs** qui représentent les associations entre auteurs qui publient des articles ensemble,
- **les réseaux de co-citation** qui représentent les associations entre couples d'auteurs qui sont cités dans la même publication,
- **les réseaux d'affiliation** qui contiennent les chercheurs scientifiques d'un même établissement (centre de recherche) ou d'un même événement de recherche (projet, séminaire, conférence).

Média social : selon (Attias et al., 2010), un média social est une plate-forme dont les activités intègrent trois éléments fondamentaux : la technologie, la création de contenus et les interactions sociales. De même, (Kaplan et Haenlein, 2010) définissent les médias sociaux comme un groupe d'applications en ligne qui se fondent sur la philosophie et la technologie d'internet et permettent la création et l'échange de contenus générés par les utilisateurs.

Réseau social numérique (RSN) : les réseaux sociaux numériques sont des réseaux sociaux virtuels qui gagnent de plus en plus en popularité, non seulement dans la sphère économique ou publique mais aussi dans le monde académique.

Un grand nombre de RSNs est issu du développement des nouvelles technologies et de la popularité des médias sociaux qui permettent aux utilisateurs d'interagir, de se contacter, d'échanger des informations, de partager leurs intérêts en commun en ligne de manière générale sans limite de distance ni de temps.

Les RSNs peuvent être une source concrète et très riche pour étudier les phénomènes liés aux comportements des réseaux sociaux eux-mêmes, ou bien aux comportements des utilisateurs. Dans ce mémoire, de manière générale, le terme réseau social est employé pour désigner indifféremment un réseau social traditionnel ou un réseau social numérique.

Notons que si les définitions de média social et réseau social numérique semblent très proches, les réseaux sociaux numériques se construisent quasi uniquement sur les interactions entre utilisateurs ; la création de contenu étant une possibilité et non une exigence dans ce cas.

La sous-section suivante présente une typologie des réseaux sociaux numériques.

3.1.2. Types et caractéristiques des réseaux sociaux numériques

Le travail de (Kaplan et Haenlein, 2010) propose plusieurs catégories de médias sociaux. Nous nous sommes intéressés aux catégories qui possèdent, de manière explicite ou implicite, la caractéristique d'un réseau social dit numérique. La caractéristique « réseau social » sera dite explicite lorsque les liens entre utilisateurs sont construits explicitement par eux. La caractéristique « réseau social » sera dite implicite lorsque les liens entre utilisateurs ne sont pas explicites et peuvent être construits à partir des interactions ou actions des utilisateurs (annotations, réponses, etc.). En s'appuyant sur ce travail, nous listons ci-après les catégories principales de réseau social existantes.

Site de réseautage social (social networking site) : il s'agit d'une application qui permet de créer un profil personnel, d'inviter d'autres utilisateurs qui auront accès à ce profil afin de communiquer, envoyer des messages publics ou privés. Cette application permet également de partager des contenus de ce profil sous la forme de textes, images, vidéos ou bien audio. On distingue différents types de réseaux sociaux en fonction du contexte et de leur utilisation. Les réseaux peuvent être qualifiés de :

- généralistes : ces sites permettent de créer et d'agrandir son cercle d'amis, les plus connus étant Facebook[7], Google+[8], les plus spécifiques étant les sites de rencontre (ex. Meetic[9]) ;
- professionnels : comme LinkedIn[10] ou Viadeo[11] qui sont devenus des outils indispensables dans la relation entre professionnels en permettant de construire des réseaux professionnels personnalisés (« réseautage » professionnel). Il existe aussi des réseaux sociaux professionnels spécialisés par métiers (avocat[12], marketing, finance…) ;

[7] www.facebook.com
[8] plus.google.com
[9] www.meetic.com
[10] www.linkedin.com
[11] viadeo.com
[12] www.hub-avocat.fr

- focalisés sur les intérêts : comme la musique (MySpace[13], LastFM[14], Deezer[15], SoundClound[16]), la littérature (Babelio[17], GoodReads[18]), le cinéma (IMDb[19]), … ;
- centrés sur les services et la vie quotidienne, sur sa vie de quartier (Peuplade[20])

Blog : un blog peut être considéré comme une sorte de page web personnelle sur laquelle une ou plusieurs personnes publient périodiquement des contenus. Contrairement au site web personnel, le blog bénéficie d'une structure éditoriale préexistante, sous la forme d'outils de publication plus ou moins formatés. Les utilisateurs peuvent ajouter des commentaires et entrer en conversation sur les billets (*post*) de leur blog. Les blogs ont un caractère polymorphe puisque toutes les formes d'expression sont utilisées (image, vidéo, texte, audio).

Micro-blog (*microblogging service*) : il s'agit d'une nouvelle forme de média social, dont la conception dérive de celle du blog, elle permet aux utilisateurs de publier de courts messages (*tweet*) destinés à leurs abonnées (*followers*). Le micro-blog a pour objectif de diffuser de l'information en temps réel. Il peut contenir non seulement du texte mais aussi des images, des vidéos embarquées ou bien des liens vers des sites web. Il est donc à mi-chemin entre le blog et la messagerie instantanée. Le micro-blog le plus populaire est Twitter[21] mais il existe également d'autres plateformes comme SinaWeibo[22], Soup[23].

Communauté de partage d'informations : l'objectif de ce type d'application est le partage de contenus multimédias entre utilisateurs. Dans le contexte du web 2.0, les utilisateurs peuvent créer, indexer, commenter et partager des contenus. Ce type d'application permet de partager des images (Flickr[24], Instagram[25], Pinterest[26]...), des vidéos (Youtube[27], Dailymotion[28], ...), des présentations (Slideshare[29]), etc.

Forum de discussion : un forum est un espace de discussion public qui permet aux utilisateurs d'échanger des points de vue sur les sujets qui les intéressent ou de poser des questions. Généralement les discussions dans le forum sont archivées et cela permet des communications asynchrones entre utilisateurs. Les sujets de discussion sont souvent affichés par ordre chronologique. Les discussions peuvent s'effectuer de manière privée ou publique. Il existe plusieurs forums de discussions en ligne orientés sur différents centres d'intérêt de l'utilisateur

[13] myspace.com
[14] last.fm
[15] www.deezer.com
[16] soundclound.com
[17] www.babelio.com
[18] www.goodreads.com
[19] www.imdb.com
[20] www.peuplade.fr
[21] www.twitter.com
[22] www.weibo.com
[23] www.soup.io
[24] www.flickr.com
[25] www.instagram.com
[26] pinterest.com
[27] www.youtube.com
[28] www.dailymotion.com
[29] www.slideshare.net

comme par exemple Reddit[30], 4chan[31], Usenet[32]. Nous pouvons également citer les sites de questions/réponses (Q&A) comme Quora[33] ou StackExchange[34] qui rassemblent plusieurs forums de discussion spécialisés (par exemple, StackOverflow[35] qui est orienté sur la programmation, MathOverflow[36] qui traite de problèmes en mathématiques).

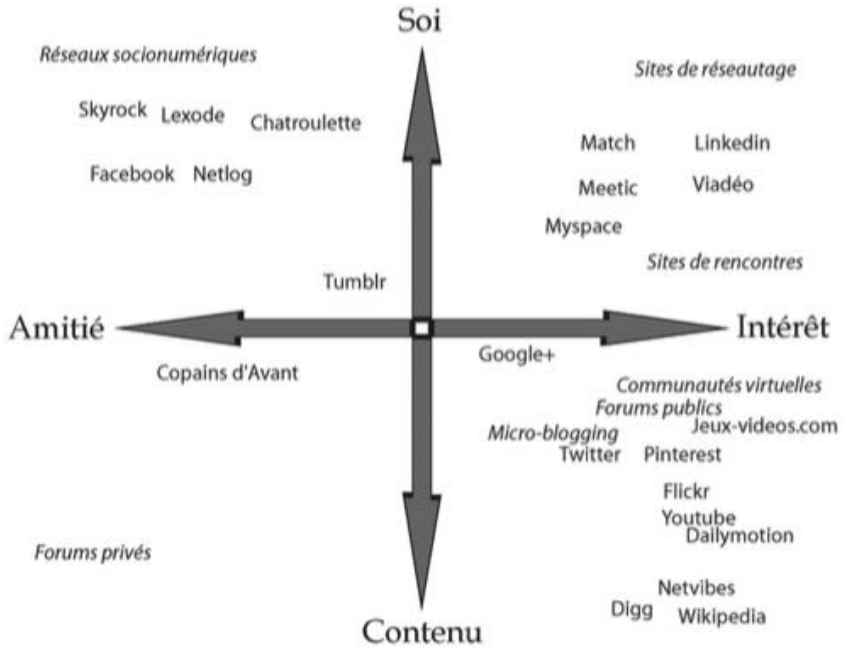
Les réseaux de projet collaboratif : ce type de média social permet la création de contenus simultanément par plusieurs utilisateurs (multi-utilisateurs). On peut distinguer 2 sous-catégories de projet collaboratif. La première rassemble les sites qui permettent aux utilisateurs d'ajouter, de modifier ou de supprimer du contenu. On appelle ce genre d'application des « wikis ». Un « wiki » très connu est Wikipedia[37], une encyclopédie en ligne disponible en plus de 230 langues. La deuxième sous-catégorie représente les sites de marque-pages sociaux (social bookmarking), qui permettent de partager des liens de sites web intéressants. Ceux-ci peuvent être « votés » par les internautes du site s'ils les trouvent également intéressants. Les liens web partagés dans l'application seront classés par rapport au nombre de votes. Ces mécanismes amènent le partage et l'évaluation collaborative de contenus multimédias. Par exemple le site web Delicious[38] permet aux utilisateurs de partager et faire connaitre leurs marque-pages qui peuvent par la suite être classés.

Avec ce type de classification, on attribue en général une catégorie d'usage générique à chaque plateforme alors que certains médias sociaux relèvent souvent de plusieurs de ces catégories. La classification des média sociaux est un problème ouvert compte tenu de la diversité des fonctionnalités offertes par chaque média social. Ces fonctionnalités peuvent être jugées plus ou moins importantes et donc mises plus ou moins en avant selon les objectifs et finalités de l'application. Par exemple, l'application YouTube a pour objectif principal de permettre à l'utilisateur de partager des vidéos en intégrant des commentaires ce qui la classe dans la catégorie des communautés de partage d'informations. En même temps, elle permet également à l'utilisateur de construire son profil (chaîne) et d'indiquer ses données personnelles (nom, description, site web personnel). Elle permet aussi de suivre d'autres chaînes YouTube et d'attribuer des mentions (*like*, *dislike*) aux vidéos regardées. Ces fonctionnalités relèvent à la catégorie de réseautage social. Quant à l'application Facebook, sa fonctionnalité la plus importante est la communication et les relations entre utilisateurs (réseautage social). Cependant, elle possède également une fonctionnalité pour partager des vidéos et permettre aux utilisateurs d'attribuer la mention (*like*) dans les « posts » ce qui relève de la catégorie communauté de partage d'informations.

(Coutant et Stenger, 2013) s'appuient sur une approche sociotechnique et ethnographique pour proposer une analyse des médias sociaux en considérant les fonctionnalités offertes par les plateformes et les pratiques des utilisateurs. Une cartographie des médias sociaux dans ce contexte est alors proposée selon deux axes : le genre de participation et la visibilité (Figure 3.1). « Les genres de participation se basent sur l'activité des utilisateurs, leur engagement, les formes d'apprentissage et les contextes » (Coutant et Stenger, 2013). Parmi ces genres, on peut distinguer la participation liée à un intérêt et la participation liée à l'amitié. Cette dernière est

[30] www.reddit.com
[31] www.4chan.org
[32] un forum de discussion décentralisé
[33] www.quora.com
[34] stackexchange.com
[35] www.stackoverflow.com
[36] mathoverflow.net
[37] www.wikipedia.org
[38] delicious.com (fermeture en 2010)

très populaire auprès des jeunes car cela leur permet de passer du temps avec leurs amis et de s'amuser. La participation basée sur les centres d'intérêt prend en compte les affinités qu'ont les utilisateurs sur différents sujets (technologie, musique, jeux vidéo, etc.). La visibilité se base sur la manière d'utiliser le média social pour partager les informations qui peuvent être soit du contenu personnel (son quotidien, ses compétences, ses relations, ses goûts, etc.), soit du contenu provenant de tiers (texte, image, vidéo, etc.)



Légende : *italique* : catégorie de média social, normal : exemple de média social

Figure 3.1 Une cartographie des média sociaux de (Coutant et Stenger, 2013)

Les réseaux sociaux dits numériques sont beaucoup étudiés dans le domaine du marketing. Plus généralement, les médias sociaux permettent aux organisations et aux entreprises d'avoir une interaction directe avec leurs clients. L'objectif poursuivi peut être professionnel, par exemple, les campagnes de marketing qui promeuvent des produits à travers les réseaux sociaux (ex. partager une photo de publicité d'un produit pour gagner un cadeau, partager des tags sur les produits, événements créés par l'organisation) ou encore en politique (ex. lors de l'élection du président des Etats Unis, Barack Obama a utilisé Twitter et Facebook pour présenter sa campagne). Ces types de campagne peuvent être appelés « Buzz ». Un buzz désigne un événement ou un phénomène qui attire l'attention des utilisateurs dans les media sociaux et implique un partage et une diffusion de l'information. Le buzz peut aussi induire la création de nouveaux liens entre personnes (rencontres autour du buzz, …).

Après avoir donné une typologie rapide des réseaux sociaux numériques, la sous-section suivante établit une comparaison, et surtout une différenciation, entre réseaux sociaux numériques et traditionnels.

3.1.3. Comparaison des réseaux sociaux numériques avec les réseaux sociaux traditionnels

Les réseaux sociaux numériques, généralement issus des médias sociaux, sont différents des réseaux sociaux traditionnels en de nombreux points (Arnaboldi et al., 2013 ; Guille, 2014) que nous listons ci-après.

- **Hétérogénéité** : nous distinguons deux niveaux d'hétérogénéité :
    - hétérogénéité inter-réseaux qui désigne l'hétérogénéité entre différents réseaux sociaux. Différents types de réseaux sociaux peuvent posséder différents types de nœuds, nature de relations, nature des interactions (réseau de connaissance, réseau de similarité, réseau de partage d'information…), orientation de relations et interaction (réseau orienté ou réseau non-orienté) et informations partagées (texte, tweet, image, vidéo, tag, …).
    - hétérogénéité intra-réseau qui désigne l'hétérogénéité dans le même réseau social. Un réseau social peut posséder différents types de nœuds, les nœuds peuvent être connectés par différents types de relations ou interactions, mais également les types d'informations partagées peuvent être différents.
- **Volume** : alors que les réseaux sociaux traditionnels reposent généralement sur un petit nombre d'acteurs, la plupart des réseaux sociaux numériques possède un grand nombre d'utilisateurs, chacun d'entre eux publiant plus ou moins régulièrement des messages. Sur Twitter il y a plus de 500 millions d'utilisateurs inscrits, plus de 400 millions de tweets envoyés par jour, et plus de 300 millions d'utilisateurs actifs chaque mois en 2016 (Statista, 2016)
- **Rapidité** : la grande force des réseaux sociaux numériques est l'immédiateté de l'interaction et de la publication. Les utilisateurs peuvent interagir ou publier les contenus et les partager à n'importe quel moment et instantanément (il n'y a généralement aucun filtrage immédiat sur le contenu publié). En termes de relations, les utilisateurs peuvent se connecter entre eux même s'ils ne se connaissent pas dans la vie réelle.

Après cette typologie des réseaux sociaux, nous étudions, en détail les éléments qui constituent un réseau social dans la section suivante.

3.1.4. Présentation et éléments d'un réseau social

Un réseau social est généralement représenté par un graphe orienté ou non orienté. Nous représenterons un réseau social par un graphe $G = (V, E)$ où V est l'ensemble des nœuds représentant les entités sociales (acteurs sociaux) et E est l'ensemble des associations entre les nœuds dans V tel que $E \subseteq V \times V$. Soient v_i et v_j deux nœuds du réseau tels que $v_i, v_j \in V$, si $e = (v_i , v_j) \in E$, alors il existe une liaison entre le nœud v_i et le nœud v_j dans G. Les nœuds v_i et v_j sont dits adjacents, ou encore connectés ou voisins. Dans ce mémoire, pour un nœud v_i, nous utilisons le terme « voisin social » pour appeler les nœuds $v_j \in V$ connectés à v_i. Le nombre total de nœuds dans le réseau est désigné par le cardinal de l'ensemble V, noté N. Ce dernier est souvent utilisé pour désigner la taille du réseau.

Nous détaillons ci-dessous les éléments fondamentaux caractérisant le graphe d'un réseau social, successivement les nœuds, les liens, les groupes et, enfin, les graphes de contenu social.

Les notations proposées ici sont basées essentiellement sur (Boccaletti et al., 2006 ; Wasserman et Faust, 1994).

3.1.4.1. Nœuds

Un nœud dans un graphe de réseau social représente une entité sociale, également appelée « acteur ». Les acteurs peuvent être des individus (appelés aussi dans ce mémoire utilisateurs) ou des groupes d'individus (organisations). Aux nœuds du graphe du réseau social peuvent être attachées des informations propres à chaque nœud. Certains travaux utilisent le terme **libellé** (*label*) pour désigner ces informations (Bhagat, Cormode et Muthukrishnan, 2011 ; Kajdanowicz, Kazienko et Doskocz, 2010). Dans certains travaux, le terme **attribut** (*attribute*) a été adopté pour assigner ces mêmes informations (Kim et Leskovec, 2010). C'est ce dernier terme que nous utiliserons dans ce mémoire. Les attributs des nœuds peuvent appartenir à différentes catégories : données démographiques (ex. âge, genre, adresse, emplacement), intérêts, loisirs, affiliation, préférences. On peut trouver aussi l'historique des activités de l'acteur ; les types d'activités que l'on trouvera en historique sont liés au réseau social sous-jacent.

Les attributs des nœuds peuvent être de différents types : simple (énuméré, numérique, textuel, etc.), par exemple genre (masculin ou féminin), âge, poids, taille, description ou structuré comme des vecteurs (ex. intérêts) ou sous forme arborescente. Certains attributs ne possèdent qu'une seule valeur (âge, genre) alors que d'autres peuvent avoir plusieurs valeurs possibles (groupes de musiques préférés, sports préférés, …).

Généralement, un réseau social est constitué d'acteurs homogènes qui ont le même statut ou rôle dans le réseau (one-mode network). Les nœuds peuvent être associés entre eux sans restriction. On peut également trouver des réseaux sociaux composés de différents types d'acteurs (*many-mode network*). Le type de réseau le plus connu et le plus étudié dans cette catégorie est un réseau composé de deux types de nœuds (*two-mode network*), également connu sous le terme réseau biparti.

Le réseau biparti est un réseau à partir duquel on peut partitionner les nœuds en deux sous-ensembles V_1 et V_2 tels que chaque lien du réseau ait une extrémité dans V_1 et l'autre dans V_2 (Borgatti, 2012). Un réseau biparti est représenté par un graphe $G = (V_1, V_2, E)$ où V_1 et V_2 représentent deux ensembles indépendants et $E \subseteq V_1 x\, V_2$. Le graphe biparti peut être transformé en graphe uni-parti $G = (V_1, E_1)$ ou $G = (V_2, E_2)$ en se basant sur leurs liens vers les mêmes nœuds en commun pour construire les associations entre nœuds. Cependant, cette approche de transformation peut impliquer une perte importante d'informations comme le montre la Figure 3.2.

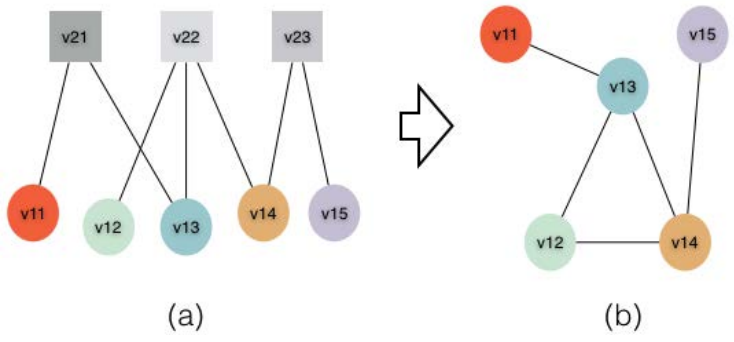


Figure 3.2 Transformation d'un réseau biparti (a) vers un réseau uni-parti des nœuds V_1 en se basant sur les liens en commun vers le nœud V_2 (b)

3.1.4.2. Liens entre nœuds

Les liens dans un graphe permettent d'associer des nœuds par paires. Dans le contexte d'un réseau social, l'association peut être faite grâce aux relations ou interactions sociales. En effet, les deux termes relations et interactions peuvent avoir un sens plus ou moins similaire. Dans ce mémoire, nous différencions le terme « relation » et le terme « interaction » de la manière suivante :

Relation désigne le fait que deux acteurs sont liés par l'un des liens sociaux suivants :

- connaissance : être membre de la famille, amis, collègues du travail etc.,
- proximité géographique : être dans la même zone géographique (ex. quartiers, village, etc.),
- association, affiliation : être dans la même association (ex. club de sport, club de musique, …) ou le même établissement (ex. école, université, …),
- similarité sociale : s'intéresser au même sujet, partager les mêmes intérêts.

Interaction : l'interaction peut être considérée comme un type de relation entre les acteurs ; l'interaction fait naître une relation. L'interaction sociale est la communication ou l'échange entre deux utilisateurs (ex. discuter, envoyer des messages, …). Les interactions peuvent être exploitées de deux manières différentes. La première génère un lien entre les acteurs qui sont en interaction. La seconde permet de définir des mesures entre des acteurs. Par exemple, dans certains réseaux sociaux, le comportement des interactions (nombre, fréquence) peut être utilisé pour désigner le niveau de confiance entre deux acteurs (Gilbert et Karahalios, 2009 ; Granovetter, 1973).

Les relations dans un réseau social peuvent être caractérisées par différents aspects, que nous classifions comme décrits dans les paragraphes suivants : l'orientation de liens, leur pondération, le caractère explicite ou implicite des relations sociales, plusieurs types de relations (multidimensionnel).

a. Orientation des liens

Les liens dans les réseaux sociaux peuvent être réciproques ou non. Un exemple de lien réciproque est une relation de connaissance, d'amitié : si Bob et Alice sont amis, alors Alice connaît Bob et Bob connaît Alice. Un exemple de lien non réciproque est une relation comme « suivre », « être fan de quelqu'un » : Bob peut être fan de Zidane un joueur de foot alors que ce dernier ne le connait pas forcement.

Partant de cet aspect, on peut distinguer deux grandes familles de réseaux sociaux : les réseaux non-orientés et les réseaux orientés (cf. Figure 3.3).

Un **réseau non-orienté** est un réseau dont les relations sont bidirectionnelles. Le sens des relations n'est donc pas pris en compte dans ce type de réseau ; les relations sont considérées comme réciproques. Un réseau non orienté se représente sous forme d'un graphe dont l'ensemble des liens E regroupe des couples de nœuds non ordonnés, appelés liens non-orientés. Pour un réseau contenant N nœuds, le nombre de liens maximal est alors de $N * (N-1) / 2$.

Un **réseau orienté** est un réseau dont le sens des relations est pris en compte. Il se représente sous forme d'un graphe dont l'ensemble des liens E regroupe des couples de nœuds ordonnés, appelés liens orientés. Dans un graphe orienté, la présence d'un lien $e1 = (v_i , v_j)$ entre les nœuds v_i et v_j n'implique pas nécessairement l'existence d'un lien $e2 = (v_j, v_i)$. Dans de tels types de réseaux, l'orientation des liens est généralement représentée graphiquement par une flèche

indiquant la direction du lien. Pour un réseau contenant N nœuds, le nombre de liens maximal est de $N * (N - 1)$.

b. Pondération de liens

Un lien dans un réseau social peut être pondéré ou non (Newman, 2004b). Un lien pondéré est affecté d'un nombre réel positif appelé poids de ce lien. Ce poids sert à différencier deux liens lors d'une exploitation du graphe et sa signification sera relative au calcul effectué pour déterminer ce poids (par exemple la date de dernière interaction, le nombre d'interactions, etc.). On peut donc distinguer deux grandes familles de réseaux : les réseaux non-pondérés et les réseaux pondérés. La Figure 3.3 ci-dessous illustre la différence entre ces deux types de réseau.

Un **réseau non pondéré** est un réseau dans lequel chaque lien n'a pas de poids. Lors de l'exploitation du graph, les liens existants ont tous la même importance.

Un **réseau pondéré** est un réseau dans lequel chaque lien $e = (v_i,v_j)$ est caractérisé par un poids $w(v_i,v_j)$ qui correspond à une valeur affectée au lien. Dans un réseau non-orienté si le lien $e = (v_i,v_j)$ appartient à E, on a $w(v_i,v_j) = w(v_i,v_j)$.

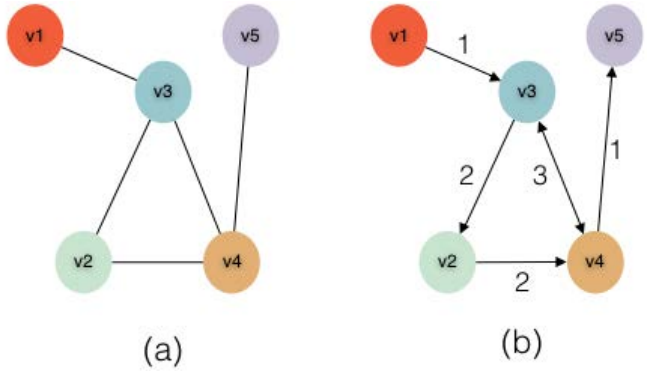


Figure 3.3 Réseau non-orienté non pondéré (a) et réseau orienté et pondéré (b)

c. Réseau social explicite versus implicite

Nous pouvons distinguer les types de réseaux sociaux selon la manière dont les liens entre les individus sont créés : on parle alors de réseau social explicite ou de réseau social implicite.

Dans un **réseau explicite,** les relations entre les individus sont créées explicitement par les individus eux-mêmes, sont connues d'eux-mêmes et des autres. Dans le contexte des RSNs, les réseaux explicites sont les réseaux où les utilisateurs déterminent explicitement qui sont les utilisateurs avec lesquels ils veulent se connecter (amis, collègues du travail, famille). Par exemple sur Facebook, Myspace et LinkedIn, la connexion se fait par la demande d'ajout de contact. Sur Google+, un utilisateur peut s'abonner au compte d'autres utilisateurs en les mettant dans des « cercles ». Sur Twitter ou Instagram, un utilisateur peut suivre les autres utilisateurs directement si le compte de ces utilisateurs est public ou via la demande de suivi si le compte de ces utilisateurs est privé. Dans tous les cas, la topologie de ce type de réseau reflète le choix des utilisateurs de se connecter ou non avec d'autres personnes et reflète souvent aussi les liens qui existent dans la vie réelle (Frey, Jégou et Kermarrec, 2011).

Dans un **réseau implicite**, les relations ne sont pas créées par les utilisateurs mais sont extraites implicitement à partir des interactions des utilisateurs ou des informations données, par exemple, les réseaux collaboratifs de chercheurs (extraits depuis les co-publications entre chercheurs ou laboratoires, ou les participations aux conférences). Dans le contexte des RSNs,

on trouve souvent ce type de relations dans les réseaux de partage d'informations dans lesquels on extrait les intérêts des individus à partir des données disponibles (partage, diffusion). Par exemple, sur *Delicious*, on peut extraire un réseau d'utilisateurs qui annotent les mêmes contenus, sur Twitter on peut extraire le réseau des utilisateurs qui mettent les mêmes « *hashtags* » sur les informations qu'ils partagent. Sur les forums de discussion, on peut extraire les relations provenant des utilisateurs réagissant sur le même fil de discussion (« *thread* »). Le réseau social n'est pas connu des utilisateurs mais est construit par analyse de données, à des fins d'analyse du réseau par exemple.

d. Réseau social multidimensionnel

Un individu peut établir plusieurs types de relations avec plusieurs types de personnes. Par exemple, Bob est ami avec Alice Carol et Eve, il est collègue de travail de Dave dans l'entreprise E-Corp et est abonné au même club de foot qu'Eve. Dans ce cas, on peut définir pour Bob trois types de relations dans son réseau social : ami (Alice, Carol et Eve), collègue de travail (Dave), et abonné au même club (Eve).

A partir de ces caractéristiques du réseau, un **réseau multidimensionnel** (*Multidimensional network*), appelé aussi réseau multiplexe, est un réseau qui permet de définir plusieurs types de liens entre deux ou plusieurs individus (Berlingerio et al., 2013 ; Tian Dai, Chong Tat Chua et Lim, 2012). Chaque lien est donc qualifié par le type de relation qu'il définit. Ainsi, deux individus peuvent être reliés par plusieurs liens, chacun qualifié par un type de relation, par exemple, Bob est ami de Eve et Bob est dans le même club que Eve.

Les nœuds et les relations entre nœuds permettent de donner naissance à un niveau d'analyse intéressant dans un réseau social : le groupe, notion abordée ci-après.

3.1.4.3. Groupes

Les travaux en analyse des réseaux sociaux peuvent s'appliquer à différents niveaux dans le réseau comme décrit Figure 3.4. Une **Dyade** représente une paire de nœuds et leurs relations. Une **Triade** est un sous-graphe composé de 3 nœuds et de leurs inter-relations. **Un Groupe, ou cluster** peut être globalement défini comme un ensemble de nœuds fortement connectés entre eux et plus faiblement connectés au reste du réseau. Les **communautés,** également appelées « *component* », sont des groupes de nœuds qui peuvent partager des propriétés communes et/ou jouer un même rôle dans le réseau. Nous pouvons trouver par exemple des communautés basées sur la famille, sur les connaissances, sur les collègues de travail, sur les personnes d'un même quartier etc.

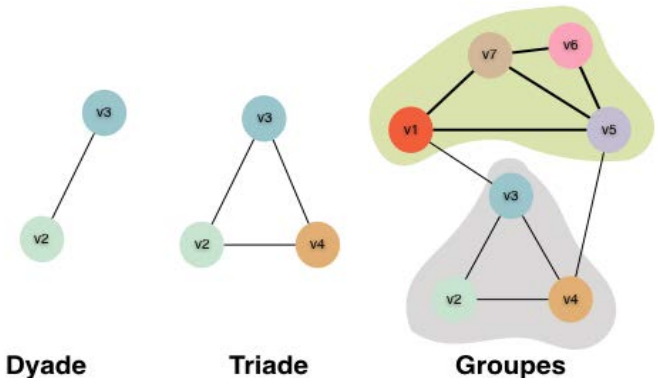


Figure 3.4 Dyade, triade et groupes ou clusters

3.1.4.4. Graphe de contenu social

Comme vu précédemment, les RSNs sont des réseaux riches en informations : les relations entre utilisateurs et les contenus générés par les utilisateurs. De ce fait, (Yahia, Benedikt et Bohannon, 2007) définissent un graphe de contenus sociaux comme un graphe biparti représentant la création et la consommation d'informations partagées dans le réseau social. Ce graphe est composé de 2 types de nœuds correspondant aux individus (utilisateurs) et aux contenus (ex. texte, photo, vidéo). Les relations peuvent avoir des libellés associés (tags). Un exemple d'un tel graphe est présenté dans la Figure 3.5.

La sémantique des relations possibles entre les nœuds varie selon le type du nœud source et celui du nœud cible. On peut distinguer 4 types de relations :

- **Personne-Personne (***Person-to-Person***)** représente les relations sociales entre des individus (amis, collègues de travail, etc.).
- **Personne-Contenu (***Person-to-content*) désigne les relations que les individus ont vis-à-vis d'un contenu au travers des interactions que ce contenu génère comme partager, aimer, commenter.
- **Contenu-Contenu (***Content-to-content***)** désigne les relations entre des contenus partagés par exemple des liens hypertexte, les relations entre deux commentaires sur un contenu.
- **Contenu-Personne (***Content-to-Person***)** représente les relations d'un contenu vers une personne et que l'on peut déduire du contexte de génération de ce contenu, par exemple un article ou un commentaire est écrit par des individus.

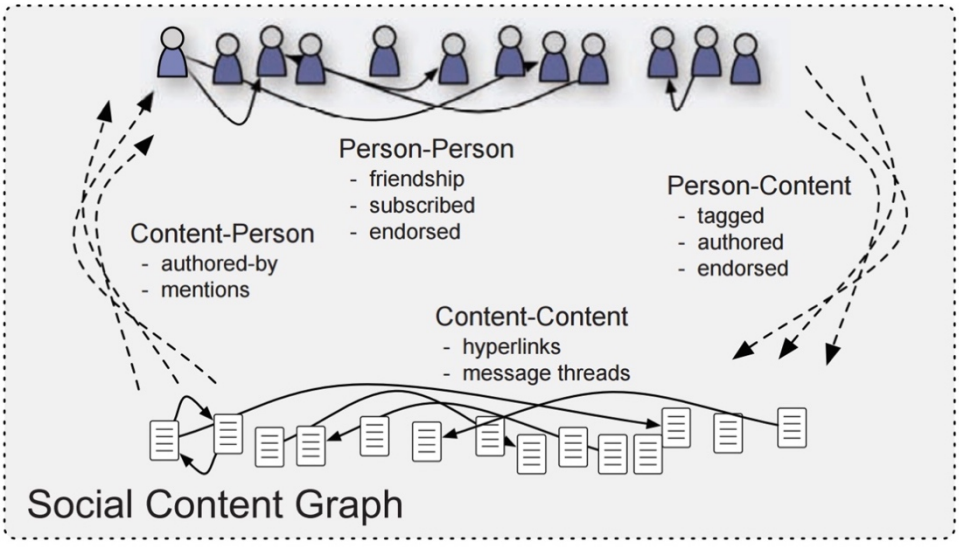


Figure 3.5 Graphe de contenu social (Yahia, Benedikt et Bohannon, 2007)

Après avoir étudié les éléments constituant un réseau social, nous montrons dans la section suivante comment ces éléments peuvent être utilisés pour analyser les réseaux et en déduire ou calculer des propriétés.

3.2. Analyse des réseaux sociaux

L'analyse des réseaux sociaux est menée dans le domaine des sciences sociales depuis les années 1930 (Breslin et Decker, 2007). Cette analyse vise à identifier les structures sociales présentes dans les réseaux et à expliquer le comportement des individus au sein de ces structures sociales, en appliquant des modèles mathématiques (théorie des graphes) ou des éléments issus de la sociométrie. L'accessibilité de plus en plus croissante des données sociales des utilisateurs avec l'explosion du Web 2.0 et des RSNs a ouvert la voie à des expérimentations sociales ou automatisées beaucoup plus importantes (Mehra, 2005 ; Wasserman et Faust, 1994). L'aspect numérique et donc la disponibilité des RSNs attirent l'attention du monde de la recherche pour leur aspect « stockable » et « traçable ». Les travaux en analyse des réseaux sociaux exploitent les données des RSNs pour en démontrer de manière empirique les théories ou propriétés.

Dans cette section, nous présentons tout d'abord des éléments de sociologie liés à l'analyse de réseaux sociaux puis nous détaillons les différents aspects des réseaux étudiés qui nous ont servi dans nos travaux.

3.2.1. Eléments de sociologie pour l'analyse des réseaux sociaux

Avant l'avènement des réseaux sociaux numériques, l'essentiel des travaux en analyse des réseaux sociaux a été mené en sciences sociales. Les différentes problématiques abordées dans ces études sont très vastes. Nous présentons ici uniquement les éléments que nous jugeons importants pour nos travaux : les analyses socio-centrées ou égocentrées, le capital social, la corrélation sociale et l'influence sociale, et, enfin, la force des liens.

3.2.1.1. Analyse socio-centrée et analyse égocentrée

L'analyse de réseaux sociaux peut être divisée en deux grandes approches selon le niveau d'analyse : l'analyse socio-centrée et l'analyse égocentrée.

a. Analyse socio-centrée

L'analyse socio-centrée porte sur le réseau entier. Un point important est ici de clairement marquer la frontière du réseau (la frontière peut être claire dans une entreprise, mais pas dans un groupe) en définissant des critères de sélection des nœuds et des relations. Les éléments mathématiques de la théorie des graphes sont très souvent utilisés dans ce type d'analyse. Elle est utile pour détecter par exemple les structures sociales (groupes, clusters, …) ainsi que leurs relations dans le réseau. Le problème principal de ce type d'analyse est la nécessité, et parfois la difficulté, d'accéder aux données du réseau entier. De plus, elles nécessitent le traitement de très grands volumes de données lorsqu'elles sont appliquées à des réseaux sociaux numériques réels et publics (ex. Facebook, LinkedIn, Twitter…).

b. Analyse égocentrée

L'analyse égocentrée est centrée sur un individu en particulier et peut être répétée sur plusieurs individus. Elle porte sur le réseau personnel de l'individu, appelé réseau égocentrique (*egocentric network* ou *ego network* en anglais). Un **réseau égocentrique** représente la cartographie de l'ensemble des relations directes d'un individu focal (appelé « *égo* »). Il s'agit d'un graphe composé des relations entre les individus (appelés « *alters* ») situés à distance 1 (directement reliés) de l'égo, ce dernier étant bien entendu exclu de ce graphe. Cette notion peut

être généralisée pour prendre en compte les utilisateurs situés à distance $k \in \mathbb{N}$ de l'égo dans le réseau social ; on a alors des réseaux k-égocentriques.

Nous pouvons représenter un réseau égocentrique sous la forme d'un graphe : dans un réseau social représenté par le graphe non orienté G = (V, E), pour un individu u, son réseau égocentrique est représenté par un graph G'(u) = (V', E') avec V' $\subseteq$ V et E' $\subseteq$ E où

- V' est l'ensemble de nœuds qui sont directement connectés à u : $\forall$ v $\subseteq$ V, e = (u, v) $\in$ E => v $\in$ V'.
- E' est l'ensemble des relations entre les nœuds dans V' : $\forall$ $v_i \in V'$ et $v_j \in V'$, e = (v_i, v_j) $\in$ E => e $\in$ E'

Le réseau égocentrique est donc un sous-graphe du graphe complet comme le montre la Figure 3.6. Notons que la définition peut être adaptée dans le cas des graphes orientés en choisissant pour les définitions de V' et E' les orientations d'arcs souhaitées. Le choix des orientations d'arcs retenues dépendra du sens porté par les arcs.

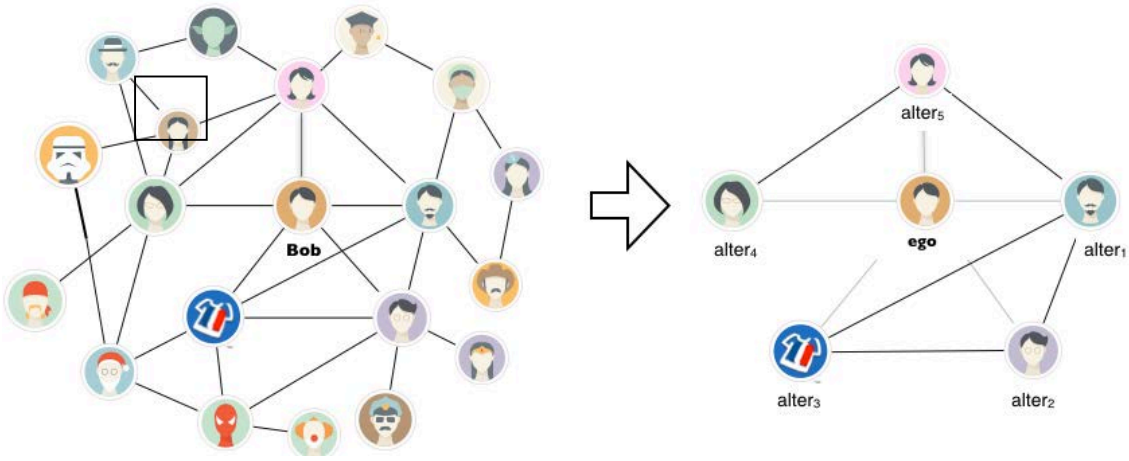


Figure 3.6 A gauche un réseau social complet dont Bob fait partie, à droite, le réseau égocentrique de Bob

En analyse égocentrée, les difficultés liées au volume pour traiter les données sont bien moindres qu'en analyse socio-centrée car la taille du réseau étudié est réduite drastiquement par construction.

La sous-section suivante aborde la notion de capital social qui permet de donner une valeur à un réseau social.

3.2.1.2. Capital social

En sociologie des réseaux sociaux, le capital social est une ressource associée à un réseau. La valeur de cette ressource dépend de la taille du réseau personnel d'un individu et du volume et de la richesse en ressources du réseau. Elle mesure le degré de facilité d'accès à des ressources (informations) par les individus en s'appuyant sur leurs relations sociales (Lin, 1995). Les auteurs soulignent que le capital social résulte de l'investissement d'un individu dans ses relations avec d'autres. En d'autres termes, le capital social représente la facilité avec laquelle les individus ont accès aux informations via leurs relations sociales.

Selon (Borgatti, Jones et Everett, 1998) , il existe plusieurs méthodes de calcul du capital social, certaines prennent en compte les facteurs culturels, d'autres le mesurent au niveau des individus (Fukuyama, 1996 ; Putnam, 1995) tandis que certains ne l'interprètent qu'à partir des structures

internes (Burt, 1978 ; Lin, 1995) ou externes des groupes d'individus (Everett et Borgatti, 1999).

La sous-section suivante aborde l'étude de la compréhension des relations dans un réseau social.

3.2.1.3. Corrélation sociale et influence sociale

Comprendre les relations entre les nœuds et les liens dans un réseau social est un des sujets de recherche les plus actifs dans le domaine de la sociologie et de l'analyse des réseaux sociaux. Les principales questions concernant ce sujet sont : pour un couple d'individus connectés, ont-ils des points en commun ? Et est-ce que leur relation affecte leur comportement ? Ceci est un sujet lié à la corrélation sociale et à l'influence sociale entre les deux nœuds connectés.

La **corrélation sociale** est le fait qu'il puisse y avoir un lien de causalité entre les actions d'un individu et les affiliations qu'il entretient dans le réseau social. Les causes de cette corrélation sont diverses, par exemple l'influence sociale et l'homophilie.

L'influence sociale (*social influence*) est un phénomène bien connu des réseaux sociaux. Ce phénomène désigne le changement de comportement d'individus affectés par les autres individus du réseau en interaction avec eux (Sun et Tang, 2011).

Un autre phénomène associé à ce sujet de recherche est l'homophilie. Ce phénomène désigne la tendance qu'ont les individus à se connecter avec d'autres individus qui ont des similitudes ou des points communs avec eux (partager les mêmes intérêts, travailler dans le même établissement, …). (Singla et Richardson, 2008) ont étudié les relations entre les utilisateurs de la messagerie électronique *Messengers* et ont montré que les utilisateurs qui communiquent sur le réseau de cette messagerie électronique ont plus de chance d'être similaires que des paires d'utilisateurs pris aléatoirement. La similarité entre les utilisateurs est mesurée selon l'âge, le genre, le code postal de leur résidence, les requêtes de recherche effectuées sur le web. De plus, ils ont également montré que la similarité augmente avec le temps de discussion dans la messagerie (nombre d'échanges, …).

Les études comme celles de (Crandall et al., 2008) et (Aiello et al., 2013) sur le phénomène d'homophilie dans le contexte des media sociaux ont montré, d'une part, que la similarité entre les utilisateurs qui se connectent entre eux a tendance à augmenter au fil du temps grâce à l'influence sociale et, d'autre part, que les utilisateurs ont tendance à se connecter à d'autres utilisateurs qui leur sont similaires (principe de l'homophilie), ce qui a tendance à amplifier l'effet de l'influence sociale.

Plusieurs études se basent sur ce concept de corrélation sociale et d'influence sociale pour déduire les comportements ou les caractéristiques inconnus des personnes dans un réseau social. Elles tentent de répondre à la question : en ayant des informations sur seulement quelques individus dans un réseau, comment peut-on déduire les comportements des autres individus du réseau que l'on ne connait pas encore ? (Leenders, 2002 ; Singla et Richardson, 2008 ; Wen et Lin, 2010). Ce sont ces études qui sont à l'origine de nos travaux de recherche.

Selon (Aral et Walker, 2013 ; Shi, Adamic et Strauss, 2007)[54], la force des liens est un facteur important qui peut impacter l'influence sociale entre les individus. Nous présentons en détail

[54] En fait, (Aral et Walker, 2013) considère deux facteurs différents qui peuvent faire varier le taux d'influence : la force des liens qui représente la signification et l'intensité des relations et le niveau d'ancrage dans le réseau (embeddedness) qui représente la quantité d'amis en commun.Nous considérons ici que la force des liens peut être caractérisée par le niveau d'ancrage.

dans la section qui suit, la force des liens ainsi que les mesures qui permettent d'évaluer cette force entre les individus.

3.2.1.4. La force des liens

La force des liens (*tie strength*) est un concept introduit dans le travail de (Granovetter, 1973) largement reconnu : « *The Strength Of Weak Ties* ». Selon Granovetter, *« la force des liens entre deux individus est une combinaison entre la quantité de temps passé ensemble, l'intensité émotionnelle, l'intimité (confiance mutuelle) et la réciprocité des services qui caractérisent le lien entre ces deux individus »*. Il distingue deux types de liens, les liens forts et les liens faibles. Les liens forts d'un individu sont les personnes proches avec qui il partage beaucoup de confiance et avec qui il entretient des échanges réguliers et qui sont le plus souvent dans les mêmes cercles sociaux que lui. Souvent, ce sont les individus qui sont similaires et qui ont beaucoup de liens entre eux. Les liens faibles sont, à l'opposé, ses connaissances avec qui il n'a juste que de brefs contacts occasionnels. L'auteur démontre que les liens faibles d'un individu sont ceux qui sont les plus susceptibles de lui apporter de nouvelles informations qui sont les plus inédites, les plus difficiles d'accès, les plus originales et sont par conséquent plus utiles que ses liens forts pour accéder à de nouvelles informations. Cela démontre l'utilité des liens faibles dans les relations sociales. L'auteur a aussi montré que les liens faibles offrent beaucoup plus d'opportunités dans la recherche d'emplois par exemple.

Beaucoup de travaux utilisent le concept de la force des liens pour étudier les comportements des individus ou des organisations. L'une des principales questions est de savoir comment évaluer la force d'un lien (fort ou faible). Selon (Gilbert et Karahalios, 2009), il existe 7 dimensions pour mesurer la force des liens. Les quatre premières dimensions ont été évoquées dans (Granovetter, 1973) : la quantité de temps passé ensemble, l'intensité émotionnelle, l'intimité (confiance mutuelle) et la réciprocité des services. Des travaux plus récents étendent cette liste. (Burt, 2004) propose d'utiliser les facteurs structurels comme la topologie du réseau pour calculer la force des liens. (Wellman et Wortley, 1990) supposent que le support émotionnel entre les personnes peut montrer leur lien fort (ex. conseil de famille par rapport à un problème familial). Enfin, (Lin, Ensel et Vaughn, 1981) ont montré que la distance sociale, caractérisée par le statut socio-économique, le niveau d'éducation, l'affiliation politique, la race et le genre, peuvent influencer la force des liens entre les personnes.

Dans la pratique, des indicateurs (informations) relatifs à ces dimensions ont été adoptés comme mesure et modèle pour évaluer la force des liens : la topologie du réseau, la réciprocité de la communication (Friedkin, 1980), le fait de posséder des amis en commun (Shi, Adamic et Strauss, 2007), la date de la dernière communication (Lin, Dayton et Greenwald, 1978), la fréquence de communication (Bond et al., 2012 ; Gilbert, Karahalios et Sandvig, 2008).

(Gilbert et Karahalios, 2009) étudient les mesures de force des liens dans le contexte des média sociaux. En utilisant Facebook comme terrain d'étude, ils ont défini de nouvelles mesures propres aux fonctionnalités de cette application (par exemple le nombre de mots dans les posts échangés sur le mur de l'utilisateur, le nombre d'échanges de messages privés, le nombre de jours depuis la dernière communication, le nombre de groupes en commun, la liste des intérêts en commun, etc). Au total, 74 variables de Facebook ont été étudiées comme mesures de force des liens. L'expérimentation sur 2000 relations sur Facebook a montré que les mesures étudiées ont plus de 85% de pertinence pour prédire la force des liens des participants.

Après avoir introduit quelques éléments de sociologie permettant d'appréhender et de comprendre les réseaux sociaux, nous détaillons ci-après les enjeux de l'analyse des réseaux sociaux.

3.2.2. Différents aspects de l'analyse des réseaux sociaux

L'analyse des réseaux sociaux est l'étude de ces réseaux afin de mettre en exergue des propriétés du réseau social. C'est un vaste domaine et nous nous restreindrons à une partie d'état de l'art en lien direct avec notre étude pour analyser les relations entre individus et informations en prenant en compte la dynamique sous-jacente. Nous allons nous intéresser, dans cette section, aux mesures et propriétés des réseaux sociaux, à l'analyse de la dynamique d'un réseau social, à la prédiction de liens et enfin à la détection de communautés.

3.2.2.1. Propriétés des réseaux sociaux et mesures associées

Les propriétés des réseaux sociaux peuvent être étudiées au niveau local ou au niveau global. Les mesures locales s'intéressent uniquement aux propriétés des nœuds et des liens, alors que les mesures globales considèrent l'ensemble du réseau à travers des propriétés statistiques calculées sur l'ensemble de la structure (Boccaletti et al., 2006). Ces deux aspects sont étudiés dans les deux sous-sections qui suivent. Les formules présentées dans cette section se basent principalement sur le travail de (Boccaletti et al., 2006) et de (Burt et Minor, 1983).

a. Propriétés locales

Les propriétés locales peuvent se distinguer selon le niveau des entités sociales qu'elles décrivent (granularité). On peut étudier les propriétés au niveau des nœuds seuls ou bien au niveau des groupes de nœuds (communautés). Ces deux aspects sont étudiés dans les sous-paragraphes suivants.

❖ Propriétés des nœuds

- **Degré** : le degré d'un nœud v_i dans un graphe $G = (V, E)$, noté k_{vi} est le nombre de liens dans lesquels intervient le nœud v_i. Dans un graphe orienté, on distinguera le degré entrant k_{vi}^{in} (nombre de liens entrants) et le degré sortant k_{vi}^{out} (nombre de liens sortants) (Boccaletti et al., 2006) . On utilise le degré comme mesure pour étudier la connectivité d'un nœud dans le réseau. Il permet par exemple de déterminer le rôle des nœuds dans le réseau (nœud influenceur, nœud populaire, nœud isolé, …).
- **Distance (taille du plus court chemin)** : la distance est une mesure qui fournit une propriété locale entre deux nœuds. La distance entre un nœud v_i et un nœud v_j, noté $d_{vi, vj}$, désigne le plus petit nombre de liens qu'il faut parcourir pour joindre ces deux nœuds (c'est la taille du plus court chemin entre ces deux nœuds également appelée distance géodésique) (Boccaletti et al., 2006).
- **Centralité** : la centralité désigne la position (plus ou moins centrale) d'un nœud relativement aux autres nœuds dans le graphe. La mesure de centralité est souvent utilisée pour mesurer le capital social des individus (Burt, 1978). Il existe plusieurs mesures de centralité dans la littérature. Nous présentons ici uniquement les trois mesures les plus exploitées (Freeman, 1978) :
    - **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 d'un acteur dans un graphe. Avec cette mesure, l'acteur qui occupe la position la plus centrale dans un graphe est celui qui possède le plus grand nombre de connexions directes dans le graphe. Dans un graphe $G = (V, E)$, le degré de centralité d'un nœud v_i, noté $C_d(v_i)$, est le nombre de connexions directes (degré) de v_i noté k_{vi} normalisé par le nombre maximal de connexions directes qu'un nœud peut avoir (formule (3.1)).

$$C_d(v_i) = \frac{k_{vi}}{N-1}$$

(3.1)

- **La centralité de proximité (*closeness centrality*)** est une mesure qui repose sur la distance entre les nœuds. Selon cette mesure de centralité, un individu est considéré plus central s'il peut accéder facilement aux autres nœuds dans le réseau, c'est-à-dire que sa distance avec les autres nœuds est faible. Elle peut être considérée comme une fonction inverse de la distance moyenne du nœud avec les autres nœuds. La centralité de proximité d'un nœud v_i, notée $C_p(v_i)$, d'un graphe *G = (V, E)* est calculée par la somme de ses distances avec les autres nœuds $v_j \in V$, normalisée par le minimum de cette somme $(N-1)$ pour le cas où v_i est directement lié à tous les autres individus v_j du graphe *G*) (3.2).

$$C_p(v_i) = \frac{N-1}{\sum_{j=1}^{N} d(v_i, v_j)}$$

(3.2)

- **La centralité d'intermédiarité (*betwenness centrality*)** est une mesure qui repose sur la position intermédiaire des nœuds d'un graphe. D'un point de vue conceptuel, (Freeman, 1978) a défini cette mesure comme la capacité que les individus ont à assurer un rôle de coordination et de contrôle. L'hypothèse est la suivante : plus un individu se trouve dans une position intermédiaire entre plusieurs individus, plus il aura la capacité à contrôler la circulation de l'information entre les autres nœuds. Autrement dit, un nœud v_i d'un graph *G* = (*V*, *E*) est considéré important s'il se localise sur plusieurs chemins entre d'autres nœuds du graphe. La centralité d'intermédiarité de v_i , notée $C_i(v_i)$, est définie par la somme des nombres de chemins les plus courts entre deux nœuds v_j et$v_k \in V$ qui passent par le nœudv_i, normalisée par le nombre de chemins les plus courts entre toutes les paires v_j et$v_k \in V - \{v_i\}$. Soit Pv_jv_k le nombre de chemins les plus courts (géodésiques) entre deux nœuds v_j etv_k, la centralité d'intermédiarité du nœud v_i est définie par la somme des chemins les plus courts entre v_j et v_k passant par v_i notés $Pv_jv_k(i)$ (avec $i \neq j \; et \; i \neq k$), normalisée par le nombre total des chemins les plus courts entre v_j et v_k qui n'incluent pas v_i (3.3).

$$C_i(v_i) = \sum_{v_j, v_k \in V-\{v_i\}, i \neq j, i \neq k} \frac{Pv_jv_k(v_i)}{Pv_jv_k}$$

(3.3)

- **Coefficient de clustering** : le coefficient de clustering désigne la probabilité que deux voisins v_j et v_k du nœud v_i soient eux-mêmes voisins. Considérons une disposition « géométrique » des nœuds dans laquelle v_i a deux voisins v_j et v_k et dans laquelle v_j et v_k sont aussi voisins. Ils forment ainsi un triangle. t_{vi} est le nombre de triangles dont le nœud v_i fait partie. Le calcul du coefficient de clustering est donné par la formule (3.4):

$$coefficient\ de\ clustering(v_i) = \frac{2*t_{vi}}{k_{vi}*(k_{vi}-1)}$$

(3.4)

Pour rappel, k_{vi} est le degré du nœud v_i (cf. ci-avant).

❖ Propriétés des groupes (ou communautés)

Les propriétés des groupes de nœuds peuvent être dérivées des mesures existantes liées aux nœuds (individus) présentées précédemment mais appliquées à un niveau de granularité moins fin. Ci-après nous présentons les principales propriétés des groupes de nœuds.

- **Degré** : le degré d'un groupe c_i, noté k_{c_i}, dans un graphe *G = (V, E)*, est le nombre d'individus extérieurs au groupe qui sont liés à au moins un membre du groupe c_i.
- **Distance** : la distance entre deux groupes c_i et c_j est le plus petit nombre de liaisons (lien) nécessaires qu'il faut parcourir pour joindre les membres des deux groupes. La façon de calculer la distance de communauté est plus variée que celle du calcul de la distance des individus. Pour calculer la distance entre un groupe (ensemble d'individus) et un individu externe dans un graphe, plusieurs moyens de calcul peuvent être considérés : par exemple, la moyenne des distances entre 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.
- **Centralité** : la centralité désigne la position (plus ou moins centrale) d'un groupe relativement aux autres groupes dans le graphe. Nous présentons ici uniquement les trois mesures de centralité les plus importantes (Borgatti, 2012).
    - **La centralité de degré (*degree centrality*)** est une mesure qui repose sur le nombre de connexions vers le groupe dans un graphe. Avec cette mesure, le groupe qui occupe la position la plus centrale dans un graphe est celui qui possède le plus grand degré. Dans un graphe *G = (V, E)*, la centralité de degré d'un groupe c_i, noté $C_d(c_i)$, est la somme de connexions directes de tous les nœuds $v_i \in c_i$ divisée par le nombre des nœuds extérieurs (3.5).

$$C_d(c_i) = \frac{K_{c_i}}{N-|c_i|}$$

(3.5)

    - **La centralité de proximité *(closeness centrality)*** est une mesure qui repose sur la distance entre les groupes de nœuds dans un graphe. Selon cette mesure de centralité, un groupe est considéré plus central que les autres groupes si les nœuds dans ce groupe sont accessibles plus facilement par les autres groupes du réseau. Elle est définie comme l'inverse normalisé de la somme des distances du groupe vers tous les nœuds externes notés *x*. La centralité de proximité d'un groupe c_i , noté $C_p(c_i)$, d'un graphe *G = (V, E)* est calculée par la somme des distances entre tous les nœuds membres du groupe vers tous les nœuds externes *x*, normalisée par le maximum de cette somme (N-$|c_i|$) obtenue lorsque tous les individus *x* se situent à distance 1 de c_i. Le calcul est donc donné par la formule (3.6).

$$C_p\ (c_i) = \frac{N-|c_i|}{\sum_{x\ \in\{V-ci\}} d_f(x,c_i)}$$

(3.6)

$d_f(x, c_i)$ est une fonction qui se base sur la distance entre le nœud *x* et le groupe c_i (cf. définition de la distance de groupe ci-dessus).

- **La centralité d'intermédiarité (*betwenness centrality*)** mesure la proportion des chemins les plus courts (géodésiques) passant par le groupe entre les paires d'individus non membres du groupe. Si c_i est un groupe d'individus, g($v_{i,}$ v_j) le nombre de chemins les plus courts entre v_i et v_j n'appartenant pas à c_i, g(v_i,$v_{j,}$ c_i) le nombre de chemins géodésiques entre v_i et v_j passant par au moins un membre de c_i, la centralité d'intermédiarité du groupe c_i est la proportion de chemins géodésiques passant par c_i, normalisée par le nombre de chemins les plus courts entre toutes les paires d'individus non membres de c_i (calculé par $\frac{(|V-c_i|)*(|V-c_i|-1)}{2}$). Le calcul est donc donné par la formule (3.7).

$$C_i(c_i) = \frac{2 * \sum_{v_i,v_j\in\{V-c_i\}} \frac{(v_i, v_j, c_i)}{g(v_i, v_j)}}{(|V - c_i|) * (|V - c_i| - 1)}$$

(3.7)

Après l'étude des propriétés locales des nœuds et des communautés, nous nous intéressons ci-après aux propriétés globales des réseaux sociaux.

b. Propriétés globales

Les propriétés globales apportent une information sur l'ensemble de la structure du réseau. Nous citons ici, les propriétés fondamentales et que nous retenons pour nos études.

- **Densité** : la densité représente la connectivité globale à l'intérieur du réseau. La densité d'un graphe *G*=(*V*, *E*), est calculée par le nombre de liens *M*, divisé par le nombre de liens possibles $Nbliens_{max}$ (formule (3.8)).

$$Densité = \frac{M}{Nbliens_{max}}$$

(3.8)

Pour un graphe orienté, le nombre de liens possibles est calculé par la formule (3.9) :

$$Nbliens_{max} = N * (N - 1)$$

(3.9)

Pour un graphe non orienté, le nombre de liens possible est calculé par la formule (3.10):

$$Nbliens_{max} = \frac{1}{2} * N * (N - 1)$$

(3.10)

- **Degré moyen** : le degré moyen d'un graphe $G = (V, E)$ correspond à la moyenne des degrés individuels d_{vi} de chaque nœuds v_i (formule (3.11)).

$$degré\ moyen = \frac{1}{N}\sum_{v_i \in V} k_{v_i} \tag{3.11}$$

Le degré moyen d'un graphe donne une information globale sur la connectivité des nœuds. Il est également utilisé comme mesure de référence pour déterminer comment un nœud donné est connecté par rapport à la moyenne.

- **Distance moyenne** : la distance moyenne d'un graphe $G = (V, E)$ correspond à la distance moyenne séparant deux nœuds quelconques dans le réseau. Cette mesure fournit une information sur la proximité des nœuds dans le réseau ainsi que sur leur facilité à communiquer et échanger. Elle est obtenue en faisant la moyenne des distances moyennes de chaque nœud v_i (3.12).

$$Distance\ moyenne = \frac{\sum_{v_i \in V} distance\ moyenne(v_i)}{N} \tag{3.12}$$

Dans la formule (3.12), la distance moyenne(vi) est la moyenne des distances de vi avec tous les autres nœuds du graphe.

Les mesures et propriétés présentées dans cette sous-section permettent de donner des informations sur le réseau social mais permettent une analyse sur une sorte de « photographie » du réseau à un instant donné. Elles ne prennent pas en compte les modifications ou changements dans le temps liés aux informations et aux relations entre individus, c'est-à-dire l'évolution du réseau. La section suivante aborde cette problématique.

3.2.2.2. Analyse de la dynamique d'un réseau social

Pendant plusieurs années, l'étude des propriétés et des caractéristiques des réseaux sociaux (densité, degré de distribution, classification, composants connexes, communautés, etc.) a été un secteur très actif de recherche. Cependant, la plupart des études ont été conduites avec une vision globale et finalement assez statique des réseaux alors qu'un RSN est considéré comme un réseau qui évolue au fil du temps. Un domaine de recherche associé à l'évolution du réseau social est l'analyse de la dynamique des réseaux *(Dynamic Network Analysis* ou *DNA*). L'étude de la dynamique des réseaux date de la fin des années 1990 avec les fameux articles de (Barabási et Albert, 1999 ; Watts et Strogatz, 1998) qui étudient la propriété des réseaux de petit monde (*small world networks*) et des réseaux sans échelle (*scale free networks*). Ces découvertes ont ensuite été appliquées dans diverses disciplines : réseaux de collaborations scientifiques, réseaux sociaux, réseaux de télécommunication, réseaux biologiques. Différents axes de recherche ont été abordés, on peut distinguer deux principaux même s'ils se recouvrent parfois : l'analyse de la dynamique du réseau et la visualisation du réseau.

L'analyse de la dynamique du réseau est une étude sur les changements qui se produisent dans le réseau à travers le temps (Moody, McFarland et Bender-deMoll, 2005). L'étude se base donc sur les données disponibles dans le réseau et reliées à une information temporelle (dimension temporelle). Les études dans ce domaine sont plus liées aux activités des acteurs et à leurs

relations plutôt qu'à l'étude des propriétés du réseau comme le font les études sur l'analyse statique du réseau (Trier, 2008).

L'analyse de la dynamique d'un réseau social se fait en représentant le graphe du réseau social à différents moments sur un axe temporel (*graphe dynamique*). Dans la littérature, le graphe dynamique d'un réseau social est désigné par différents noms : « *temporal network* », « *evolving graph*», «*time-varying graph*», «*dynamic graph*», *« link streams* », « *timestamped graph* » (Albano, 2014 ; Cattuto et al., 2013 ; Grindrod et Higham, 2010 ; Holme et Saramäki, 2012). Le graphe dynamique peut être défini comme suit :

> Pour un graphe $G=(V,E)$, un ordre de séquences de ce graphe S est défini par : $S=\{G_1=(V_1, E_1), G_2=(V_2, E_2), \ldots, G_k=(V_k, E_k)\}$ où chaque G_t est le sous graphe de G à différent instants t avec $t_i < t_{i+1}$. Un instant t peut être une estampille temporelle (*timestamp*) spécifiée ou un intervalle de temps. Chaque V_t et chaque E_t représentent donc l'ensemble des nœuds et des relations à différentes valeur de t tels que $V = V_1 \cup V_2 \cup \ldots \cup V_k$ et $E = E_1 \cup E_2 \cup \ldots \cup E_k$. Notons que la granularité de temps pour observer des changements dépend du réseau social étudié (heure, jour, mois année, …).

L'analyse de la dynamique consiste à observer le graphe dans la série d'instants de temps $t_1,\ldots,t_{n-1}, t_n$. Entre l'instant t_{i-1} et t_i on peut observer l'addition/la suppression des nœuds ou des liens. L'objectif est alors de construire un modèle de ces évolutions afin de comprendre comment le réseau évolue mais aussi afin de prédire les processus qui vont se produire par la suite.

On peut distinguer deux approches différentes pour l'analyse de la dynamique du réseau social (Aggarwal et Subbian, 2014 ; Spiliopoulou, 2011) :

- la première approche tente d'étudier comment le réseau social évolue dans un intervalle de temps. Cette méthode observe les changements du réseau dans un intervalle de temps constitué de plusieurs instants avec un début et une fin, puis on construit un modèle correspondant aux mécanismes de changements observés durant cet intervalle. Autrement dit, cette méthode tente d'étudier le modèle d'évolution du réseau à partir d'informations relatives à une série de temps donnée. Le modèle explique ce qui s'est passé durant l'intervalle de temps.
- la deuxième approche étudie le modèle construit à chaque instant. On utilise le(s) modèle(s) résultant des instants précédents pour adapter et définir le(s) modèle(s) suivant(s). Cette méthode est adaptée pour l'exploration de données à fréquence rapide et qui sont volumineuses comme les flux de données. Généralement, l'exploration de flux de données se fait au travers d'une fenêtre temporelle (*sliding window*) : pour étudier le modèle, seules les informations qui sont dans la fenêtre temporelle sont considérées, les données antérieures ne sont pas accessibles.

Les différences entre ces deux approches peuvent être données selon différents aspects :

- la première méthode permet d'utiliser toutes les informations disponibles dans le réseau pour la construction du modèle sachant que ces informations n'évoluent pas durant la construction du modèle. A l'inverse, dans la deuxième méthode, on ne peut pas savoir comment le réseau va évoluer ni quelle quantité de données va arriver et à quel moment. Il faut donc adapter le modèle à chaque arrivée de données.
- En termes d'objectif d'utilisation, la première méthode permet plutôt de modéliser et anticiper l'évolution d'un réseau social à partir d'un intervalle de temps donné alors que la deuxième méthode permet plutôt de surveiller le réseau en temps réel.

L'analyse de la dynamique du réseau social permet, d'une part, de comprendre le mécanisme qui produit l'évolution dans le réseau social, autrement dit, de définir des modèles de l'évolution du réseau social, et, d'autre part, d'utiliser le mécanisme d'évolution observé ou les modèles étudiés pour prédire les actions ou événements qui peuvent se produire dans le réseau dans le futur.

Dans la section suivante, nous présentons les travaux liés à l'analyse de la dynamique du réseau social. Nous séparons ces travaux en deux catégories : l'analyse dynamique de la structure du réseau (la dynamique du réseau) et l'analyse dynamique de la diffusion d'informations (dynamique sur le réseau ou à l'intérieur du réseau). Ces deux points sont abordés dans les deux sous-sections suivantes. La dernière sous-section présente des travaux qui, tout en se basant sur l'une des catégories précédentes (ou les deux), ajoutent l'analyse du contenu échangé.

a. Analyse de la dynamique de la structure du réseau social

Les travaux sur la dynamique de la structure du réseau social portent sur la dynamique des **relations** entre les individus dans le réseau qui modifie donc la **structure** du réseau. Cette dynamique est liée à la création et/ou à la suppression de nœuds ou de liens mais aussi à la persistance des liens déjà existants (en particulier à la variation de la pondération des liens dans les réseaux pondérés).

Plusieurs travaux portant sur l'évolution du réseau au niveau global (macroscopique) (Kumar, Novak et Tomkins, 2006 ; Lin et al., 2009). (Kumar, Novak et Tomkins, 2006) ont étudié l'évolution de la structure des RSNs (Flickr[56] et Yahoo! 360[57]). Ils ont constaté une évolution similaire des deux réseaux qui se caractérisent par une croissance rapide des relations et des informations, suivie d'une diminution puis une reprise de croissance lente mais régulière. Dans cette étude, les auteurs ont pu identifier trois groupes d'utilisateurs :

- les « *singletons* », qui s'inscrivent dans le réseau mais ne participent pas à l'activité du réseau social (pas d'interaction ni de connexion avec les autres utilisateurs),
- les « *linkers* » qui se trouvent dans le groupe « *GiantComponent* » comprenant les utilisateurs les plus actifs et sociables. Ils se connectent soit via des liens (invitations) soit directement par la recherche d'amis via la plateforme.
- les utilisateurs du groupe « *MiddleRegion*» qui représentent les petites communautés isolées qui n'ont que des interactions internes à leur groupe (et pas avec les utilisateurs des autres groupes).

Cette étude a montré qu'avec le temps, les groupes « *MiddleRegion* » ont tendance à s'unir avec des groupes « *GiantComponent* ».

Plusieurs travaux portent sur l'évolution au niveau local (microscopique) des liens entre utilisateurs (Leskovec, Huttenlocher et Kleinberg, 2010 ; Rivera, Soderstrom et Uzzi, 2010). Selon (Rivera, Soderstrom et Uzzi, 2010), nous pouvons distinguer différents mécanismes fondamentaux qui peuvent générer la dynamique des liens (création/suppression) :

- le mécanisme « *TriadicClosure* » se base sur le principe d'amis en commun : deux individus indépendants qui ont des amis en commun ont une grande probabilité de devenir amis,

[56] www.flickr.com

[57] www.360.yahoo.com (clôturé depuis juillet 2009)

- le mécanisme « *Preferential Attachment* » établit qu'un individu a tendance à se connecter avec d'autres individus qui ont déjà beaucoup de connexions (populaires) dans leur réseau (degré élevé),
- le mécanisme « *Homophily* » établit le fait que des personnes ayant les mêmes goûts ont plus de probabilités de se connecter entre elles. Par exemple, des chercheurs travaillant sur la même thématique ont tendance à collaborer et à échanger des informations,
- le mécanisme « *Global Connection* » s'appuie sur le fait qu'un individu peut créer des contacts au-delà de ses réseaux ou de ses proches. Par exemple, les professionnels se contactent pour collaborer et créent ainsi des nouveaux réseaux professionnels,
- le mécanisme « *Random* » se fonde sur le fait que l'individu peut se connecter avec d'autres individus de façon totalement aléatoire.

L'étude de la dynamique de la structure du réseau social est une facette de la dynamique des réseaux sociaux. La sous-section suivante aborde l'autre facette complémentaire : la dynamique des informations.

b. Analyse de la dynamique de la diffusion d'informations

Un phénomène de diffusion consiste en la transmission d'un objet (virus, information, etc.) d'une entité à une autre. Les exemples de diffusion sont nombreux et variés : diffusion de maladie, diffusion de produits innovants, de virus informatique.

Dans le contexte des réseaux sociaux, il s'agit de la diffusion d'informations. Dans tous les cas, l'objet qui se propage lors d'une diffusion circule d'un contact vers un autre, ce qui signifie que la diffusion se produit sur un graphe : la diffusion se fait quand un nœud mentionne ou copie des informations depuis un nœud voisin dans le réseau. L'information apparait dans un nœud du réseau et est diffusée vers d'autres nœuds qui peuvent également transmettre aux autres nœuds voisins et ainsi de suite.

La structure du réseau a donc beaucoup d'impact sur la dynamique des informations dans le réseau social. Les individus qui sont considérés proches dans le réseau ont une probabilité plus grande d'échanger des informations. Et inversement, la dynamique de la diffusion d'informations peut également être à son tour, un facteur de changement de la structure du réseau (Stattner, Collard et Vidot, 2013 ; Weng et al., 2013).

Afin d'étudier les phénomènes de diffusion et de mieux les comprendre, des travaux portent sur les modèles pour reproduire le comportement d'une diffusion et décrire de façon formelle comment s'effectue la propagation. Dans la littérature il existe différents modèles de diffusion dans les graphes statiques et les graphes dynamiques. Ces méthodes sont hors du contexte de cette thèse et nous ne les détaillerons pas ; pour plus de détails sur ce modèle consulter (Albano, 2014 ; Masuda et Holme, 2013).

L'étude sur la diffusion d'informations peut être utilisée, par exemple, pour détecter les événements importants dans le réseau, pour détecter les individus les plus influents (Guille, 2014) ou bien pour trouver les techniques permettant de diffuser efficacement des informations dans le réseau social (Jiang, Chen et Liu, 2014).

La sous-section suivante s'intéresse à l'ajout de l'analyse des contenus dans l'analyse de la dynamique du réseau.

c. Incorporer les contenus dans l'analyse de la dynamique du réseau

Plusieurs travaux sur la dynamique d'un réseau social ne prennent pas en compte les contenus (informations) circulant dans le réseau comme un facteur important pour l'analyse de la dynamique du réseau (Stattner, Collard et Vidot, 2013). Cependant, ces contenus fournissent souvent des informations qui peuvent être exploitées pour aider à déduire ou extraire des connaissances sur le réseau social étudié. Dans la plupart des cas, le contenu et la structure évoluent en parallèle, et la dynamique de l'évolution peut être déduite de ces deux aspects (Aggarwal et Subbian, 2014).

Dans le contexte des RSNs, ce dernier point s'avère important car dans la plupart des cas, les réseaux sociaux numériques reposent sur les relations entre plusieurs utilisateurs mais aussi sur les contenus partagés. Par exemple, dans le réseau Twitter, quand un utilisateur mentionne d'autres utilisateurs dans un tweet, le tweet contient la liste des utilisateurs pour qui le tweet est envoyé mais aussi son contenu. Dans ce cas, on pourrait par exemple détecter les événements sous-jacents au réseau social étudié en combinant les informations disponibles dans le contenu et la structure. Les travaux de (Aggarwal et Subbian, 2012) définissent ce type de modèle dans lequel la structure et le contenu sont utilisés pour déterminer les événements clefs qui expliquent le flux d'informations dans le réseau social.

Nous présentons dans les sections qui suivent quelques domaines de recherche qui exploitent l'analyse des réseaux sociaux (analyse statique ou analyse de la dynamique). Nous commençons par la prédiction de liens.

3.2.2.3. Prédiction de liens

La prédiction de liens est un axe de recherche sur l'évaluation de la force potentielle des liens entre nœuds qui ne se sont pas encore connectés. Elle s'appuie sur l'analyse du réseau et se focalise sur la possibilité de formation de relations (liens) entre les nœuds dans le réseau. La prédiction de liens peut être définie globalement de la façon suivante : étant donnés deux nœuds x et y d'un réseau qui ne sont pas reliés à l'instant t_1, quelle est la probabilité d'avoir un lien entre ces deux nœuds dans le futur à l'instant t_i avec $i>1$.

Dans cette section, nous explorons successivement les domaines d'application potentiels de la prédiction de liens, les méthodes de calcul principales existantes, et enfin la prédiction de liens prenant en compte le temps, et donc l'évolution du réseau.

a. Domaines d'application

Dans la littérature, la prédiction peut être utilisée pour répondre à plusieurs objectifs :

- le calcul de la probabilité pour prédire la connexion des nœuds : la prédiction de liens permet de découvrir des liens que l'on ne peut pas observer directement dans le réseau mais qui pourraient bien compléter et enrichir sa structure dans l'avenir. Un exemple d'application de cette approche est le cas d'un réseau des co-auteurs scientifiques (Pavlov, 2007). Il est possible que deux auteurs scientifiques qui n'ont jamais travaillé ensemble auparavant puissent être amenés à collaborer, notamment quand ils sont considérés comme proches dans le réseau (ils s'intéressent aux mêmes sujets de recherche, ils ont des collaborateurs communs, ...),
- l'observation de liens cachés : la prédiction de liens peut également avoir pour objectif l'observation et la découverte des liens cachés dans le réseau, c'est-à-dire des liens qui ne sont pas visibles dans le réseau mais qui existent ou ont de grandes chances d'exister dans la vie réelle. Par exemple, les recherches en sécurité utilisent

des méthodes de prédiction de liens pour surveiller les réseaux de terroristes. Dans ce contexte, on peut découvrir que des individus travaillent souvent ensemble même s'ils n'y a pas de lien direcs entre eux (Krebs, 2002),

- l'observation de liens anormaux : on peut également utiliser la prédiction de liens pour observer des liens qui se forment de manière anormale. On trouve souvent ce type d'application dans le domaine de la sécurité pour chercher les informations biaisées, des logiciels malveillants (Rattigan et Jensen, 2005),
- enfin, la prédiction de liens pourrait également être considérée comme un autre aspect pour calculer la force des liens correspondant aux relations déjà existantes dans un réseau (Tylenda, Angelova et Bedathur, 2009 ; Xiang, Neville et Rogati, 2010). Dans ce cas, la prédiction de liens s'applique au sous-graphe ne contenant pas le lien étudié. Le calcul de prédiction donne en résultat une probabilité d'existence au lien enlevé et permet, par là même, de lui donner une pondération.

b. Méthodes de calcul pour la prédiction de liens

Généralement, les méthodes de prédiction de liens dans les réseaux sociaux s'appuient sur les techniques utilisées en théorie des graphes et en analyse des réseaux sociaux. Etant donné un réseau social, on détermine pour chaque paire de nœuds non reliés s'il existe une possibilité de formation de lien entre eux. Le principe est de calculer le score de similarité ou la vraisemblance entre les deux nœuds en question. Selon la valeur obtenue, on obtient la probabilité que ces nœuds se connectent entre eux dans le futur : plus le score de similarité est élevé plus les nœuds ont une probabilité importante de se connecter.

Nous présentons ci-après les indicateurs utilisés pour calculer le poids de similarité entre les nœuds en différenciant successivement deux types de mesures : les mesures basées sur la topologique du réseau et les mesures basées sur les attributs des nœuds.

❖ Mesures basées sur la topologie du réseau

La topologie définit les connexions et donc les chemins entre les nœuds connectés dans le réseau. Les indicateurs topologiques permettent de mesurer les propriétés topologiques entre les nœuds (ex. le nombre de relations, la distance entre les nœuds). Les indicateurs topologiques pour la prédiction de liens ont été introduits dans (Liben-Nowell et Kleinberg, 2003) afin de calculer les scores de similarité entre deux nœuds et la probabilité que ces nœuds soient reliés dans le futur. Ces indicateurs topologiques se basent soit sur les informations des nœuds voisins (le nombre de voisins des nœuds étudiés, le nombre de voisins communs) soit sur les chemins entre les nœuds. Différentes mesures sont exploitées dans la littérature :

- les mesures basées sur le nombre de voisins considèrent que la probabilité qu'un nœud y ait une relation avec un nœud x dépend du nombre de voisins du nœud x. Ceci est lié au concept d'attachement préférentiel cité précédemment. Le travail de (Newman, 2001b) utilise ce principe dans le contexte d'un réseau de collaborateurs et il montre que la probabilité de collaboration entre deux personnes x et y peut s'évaluer par le produit du nombre de collaborateurs de x par le nombre de collaborateurs de y,
- les mesures basées sur les voisins communs s'appuient sur l'idée que pour une paire de nœuds $(x,y) \in G$, les deux nœuds x et y ont une plus grande probabilité d'avoir un lien entre eux si leurs ensembles de voisins $\Gamma(x)$ et $\Gamma(y)$ possèdent une partie de recouvrement importante (ils ont un nombre important de nœuds voisins en commun). Dans la vie réelle, il parait logique que deux personnes ayant beaucoup

d'amis en commun aient tendance à se connaître également ou ont une forte probabilité de se rencontrer. On peut citer par exemple la mesure *Jaccard* (Jaccard, 1901) qui est, à la base, largement utilisée pour calculer la similarité de documents dans la recherche d'informations. Dans le contexte de la prédiction de liens, elle peut être considérée comme une méthode « voisins communs » normalisée. Il s'agit de calculer la probabilité que les nœuds *x* et *y* aient le nœud *z* en commun pour tous les nœuds *z* choisis aléatoirement à partir de l'ensemble des voisins de *x* et *y*. Nous citons également la mesure *Adamic/Adar* (Adamic et Adar, 2003) qui se base sur le nombre de voisins communs des nœuds *x* et *y*, en considérant que plus les voisins communs possèdent peu de voisins, plus il est probable que *x* et *y* se connectent. Nous présenterons plus en détail cette mesure dans la partie contribution,

- les mesures basées sur les chemins entre les nœuds se basent sur l'ensemble des chemins pour aller d'un nœud *x* à un nœud *y*. Les méthodes de calcul dans ce type de mesure sont basées généralement sur le plus court chemin qui s'intéresse à la distance la plus petite entre deux nœuds. On peut citer par exemple, la méthode *Katz* (Katz, 1953)qui calcule le nombre de chemin les plus courts dans l'ensemble des chemins existants entre ces deux nœuds.

❖ Mesures basées sur les attributs des nœuds

Dans ce type de mesure, on utilise les informations qui se trouvent sur les nœuds du réseau pour prédire la probabilité qu'ils se forment des liens entre eux. Ces informations peuvent être des descriptions du nœud lui-même ou des données qu'il partage avec d'autres nœuds. Dans le cadre de réseaux de publications scientifiques, le travail de (Yin et al., 2010) propose d'exploiter les mots-clés présents dans les publications des auteurs scientifiques pour prédire leur collaboration. (Leroy, Cambazoglu et Bonchi, 2010) proposent d'utiliser les informations données dans le profil de l'utilisateur pour calculer la similarité à partir des intérêts communs qui donne une prédiction de liens entre ces utilisateurs. Sur les réseaux Flickr et Last.fm, (Schifanella et al., 2010) ont prouvé que les utilisateurs qui partagent des contenus similaires ont tendance à se connecter. Ils proposent une prédiction de liens en se basant sur la similarité des métadonnées (tags) partagées.

Dans le contexte du micro-blog Twitter (Rowe, Stankovic et Alani, 2012) s'intéressent, en plus de s'appuyer sur la topologie du réseau, à l'analyse de la similarité des tags partagés par les utilisateurs pour prédire leurs futures relations. Le but est de recommander de nouveaux contacts aux utilisateurs (nouveaux « *followings* »).

Après avoir présenté les méthodes de calcul pour la prédiction de liens, la sous-section suivante aborde la prise en compte du temps dans la prédiction de lien.

c. Prédiction de liens temporelle

Il existe plusieurs méthodes de prédiction de liens temporelles. Ces méthodes peuvent être utilisées de plusieurs manières et pour des objectifs différents. (Tylenda, Angelova et Bedathur, 2009) utilisent des méthodes de prédiction de liens : *Adamic-Adar*, *PageRank*, *Local Probabilistic Models*, en y ajoutant la notion de temps. Par exemple, dans la méthode *Adamic-Adar* qui se base sur le nombre de voisins communs, on considèrera aussi la fraîcheur de la dernière interaction entre les deux nœuds étudiés et leurs voisins communs. Les interactions anciennes deviennent moins importantes que les interactions récentes. Elles pourraient même être oubliées. Cette idée se trouve également dans (O'Madadhain, Hutchins et Smyth, 2005).

(Munasinghe et Ichise, 2011) proposent la méthode *Time Score* qui prend en compte la fraîcheur de la dernière interaction entre deux nœuds étudiés avec leur voisins communs. D'une part, les

interactions sont pondérées par une fonction temporelle polynomiale (poids temporel de l'interaction) : plus les interactions sont récentes plus elles ont de l'importance. D'autre part, ils considèrent également que si les deux nœuds ont interagi avec leurs voisins communs pendant un même intervalle de temps court, ils ont une probabilité plus importante de se connecter. La mesure proposée dans ce travail se base donc sur ces deux facteurs : le poids temporel de la dernière interaction entre deux nœuds avec leurs voisins communs et l'intervalle de temps de leurs d'interactions avec leurs voisins communs.

Après avoir étudié des éléments essentiels des réseaux sociaux et de la prédiction de liens, la section suivante dresse un état de l'art rapide des travaux dans le domaine de la détection de communautés.

3.2.2.4. Détection de communautés

La détection de communautés est un des problèmes largement étudiés dans l'analyse des réseaux sociaux. Ce problème peut être vu comme un problème de partitionnement des nœuds (clustering) où l'ensemble des nœuds est partitionné en différents groupes (communautés). Les nœuds de chaque groupe partagent des propriétés communes et sont fortement connectés entre eux, et faiblement connectés aux nœuds à l'extérieur du groupe (Fortunato, 2010). On peut caractériser les techniques de détection de communautés selon quatre facettes complémentaires :

- **les sources d'informations utilisées pour la détection des communautés** : de façon similaire à la prédiction de liens, il existe deux grandes approches de détection de communautés selon les sources d'informations utilisées pour partitionner les nœuds (Yang, McAuley et Leskovec, 2013). L'approche basée sur la topologie des nœuds s'appuie sur la structure du réseau (liens entre les nœuds) pour identifier les nœuds à regrouper (groupe de nœuds fortement connectés). L'approche basée sur les attributs des nœuds identifie les groupes de nœuds en s'appuyant sur la similarité entre les attributs des nœuds dans le réseau. Pour être plus optimaux, certains travaux combinent l'usage de la topologie du graphe et les attributs des nœuds lorsque les deux types d'information sont disponibles (Cruz, Bothorel et Poulet, 2011),
- **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 principales techniques : la modularité, la marche aléatoire et les k-cliques. Les détails sur ces algorithmes peuvent être retrouvés dans (Fortunato, 2010),
- **la gestion du recouvrement de communautés** : un algorithme prenant en compte le recouvrement des communautés est un algorithme pour lequel un même nœud peut appartenir à plusieurs communautés en même temps (ex. algorithme *CFinder* (Palla et al., 2005), *iLCD* (Cazabet, Amblard et Hanachi, 2010)). C'est très souvent le cas dans les réseaux sociaux. D'autres algorithmes par contre n'autorisent pas le recouvrement (ex. *Infomap*(Rosvall et Bergstrom, 2007)).
- **la prise en compte de la dynamique des réseaux :** comme déjà indiqué précédemment, les réseaux sociaux sont évolutifs. De nouveaux nœuds et liens apparaissent et disparaissent au fil du temps. L'étude des réseaux dynamiques est un domaine assez nouveau mais qui attire aujourd'hui de plus en plus d'intérêt. Il existe plusieurs approches de détection de communautés dynamiques par exemple dans les travaux de (Cazabet, Amblard et Hanachi, 2010 ; Lancichinetti, Fortunato et Radicchi, 2008 ; Lin et al., 2009 ; Palla, Barabási et Vicsek, 2007 ; Rosvall et Bergstrom, 2007). Pour la plupart des algorithmes proposés, l'évolution de la

topologie du graphe implique la régénération des communautés. Ceci peut s'avérer couteux en temps d'exécution et ressources exploitées, surtout sur de très grands graphes. Pour pallier cet inconvénient, de nouveaux algorithmes tels que *iLCD* (Cazabet, Amblard et Hanachi, 2010) ainsi que ceux de (Li et al., 2012 ; Shang et al., 2012) proposent de détecter des communautés qui seront dynamiquement mises à jour au fil de l'évolution de la structure du réseau (apparition de nouveaux liens, suppression de liens existants). Cette approche réduit la complexité de calcul en évitant la régénération des communautés à chaque modification de la structure du réseau social. Les communautés détectées à court terme s'avèrent être pertinentes mais il est difficile d'assurer la cohérence des communautés détectées à long terme**.**

Nous ne détaillerons pas plus avant cette partie car notre but est seulement de comprendre les enjeux et les principes de la détection de communautés. La détection de communautés n'est pas l'enjeu de cette thèse mais un outil utilisé dans la recherche réalisée. Nous utiliserons dans nos travaux des algorithmes de ce type déjà validés.

Cette section a permis d'étudier des éléments du vaste domaine de l'analyse des réseaux sociaux. Nous nous sommes limités aux éléments intéressants, nécessaires et/ou utilisés pour construire un profil utilisateur à partir de son réseau social (profilage social), objet de la section suivante.

3.3. Profilage social

Dans les systèmes de personnalisation d'informations, une nouvelle approche qui exploite les informations à partir du réseau social de l'utilisateur s'est développée dans plusieurs travaux (Cabanac, 2011 ; Davis Jr. et al., 2011 ; Ren et al., 2010 ; Tchuente, 2013 ; Wen et Lin, 2010, 2011 ; Zeng, Yao et Zhong, 2009). A partir de la littérature, nous pouvons distinguer trois directions d'études dans ce contexte : le filtrage social d'information (système de recommandation sociale, RI sociale), la déduction des informations caractérisant l'utilisateur à partir de ses voisins dans un réseau (ex. adresse, âge, profession, établissement, intérêts, …) et enfin, la construction d'un profil social générique de l'utilisateur. Nous regroupons les travaux relatifs à ces trois domaines sous le terme commun de « **profilage social ».** Le profilage social désigne le fait d'extraire ou d'enrichir des informations sur les individus ou groupes d'individus en utilisant leur environnement social comme source d'information.

Dans l'état de l'art, la plupart des techniques actuelles de profilage social ne font pas une séparation claire entre la construction du profil et son exploitation par les mécanismes associés. Le réseau social de l'utilisateur et éventuellement son profil individuel sont directement exploités par les mécanismes (Figure 3.7 A). Il serait toutefois intéressant que le profil utilisateur puisse être perçu comme une donnée ou une information qui peut être représentée indépendamment des mécanismes de personnalisation d'informations qui l'utilisent. Ceci rendrait génériques et par conséquent réutilisables les profils construits. Notre objectif est de pouvoir définir un profil utilisateur générique et construire un véritable profil social, c'est-à-dire indépendant d'un mécanisme précis de sorte que plusieurs types de mécanismes puissent l'exploiter (Figure 3.7 (B)).

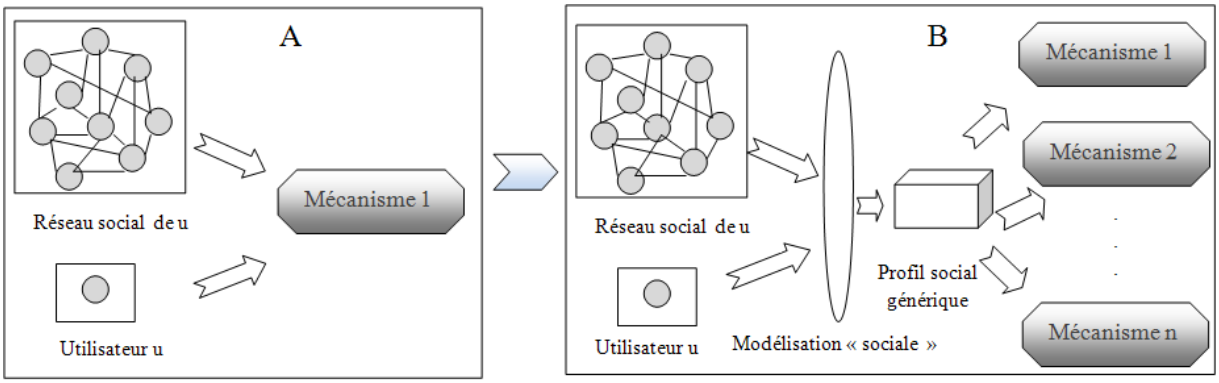


Figure 3.7 Utilisation des informations du réseau social avec et sans la construction du profil social (Tchuente, 2013)

Dans la littérature, mise à part les travaux de (Tchuente, 2013), il n'existe pas, à notre connaissance, de recherche qui évoque le terme profil social ou qui porte spécifiquement sur la construction de ce type de profil. Les travaux sur le filtrage social ne font pas forcément une séparation claire entre la construction du profil et son exploitation par les mécanismes associés, et de fait rendent le profil inutilisable par d'autres mécanismes de personnalisation d'informations. Les techniques de déduction des informations de l'utilisateur à partir des informations de ses voisins sociaux ne séparent pas forcement les informations extraites du réseau social de celles propres à l'utilisateur (profil utilisateur individuel). (Tchuente, 2013) propose un modèle du profil utilisateur qui sépare les intérêts propres à l'utilisateur dans une « dimension utilisateur » et les intérêts extraits depuis les informations partagées par ses voisins sociaux dans une « dimension sociale ».

Dans nos travaux, le terme **profil social** s'appuie sur la dimension sociale présentée dans le travail de (Tchuente, 2013) et désigne un profil dans lequel les informations sont extraites à partir des informations des voisins sociaux de l'utilisateur (réseau égocentrique). Le profil social de l'utilisateur peut contenir plusieurs types d'attributs tout comme dans le profil utilisateur (ex. localisation, professionnel, intérêts). Notons que dans ce travail nous nous intéressons en particulier à l'extraction des intérêts de l'utilisateur (ses intérêts sociaux). Le profil social peut être exploité comme un profil complémentaire au profil utilisateur existant en particulier dans le cas de profil utilisateur pauvre ou vide (problème de démarrage à froid). La manière d'exploiter le profil social dépendra de chaque mécanisme : un mécanisme pourra par exemple, recourir à l'information contenue dans le profil social si le profil utilisateur ne contient pas assez d'information. Un autre mécanisme pourra aussi par exemple, intégrer systématiquement les informations de ce profil social en plus de celles du profil utilisateur dans le processus de personnalisation d'informations. Dans cette thèse, nous n'aborderons pas cette partie mais seulement la construction du profil social.

La construction du profil social s'appuie sur le réseau social de l'utilisateur. La section suivante définit les éléments fondamentaux des réseaux sociaux utilisés dans les travaux du domaine.

Les approches de profilage social reposent généralement sur l'analyse du réseau social de l'utilisateur. Le réseau social utilisé dans les travaux associés peut être soit le réseau entier, soit une partie du réseau social extrait selon la nature de liens (réseau de familiarité basé sur les connaissances entre personnes (réseau social explicite de l'utilisateur), réseau de similarité basé sur les activités dans le système (réseau social implicite)).

Nous observons que dans la plupart des travaux, le réseau social utilisé repose sur la notion de réseau égocentrique et sur l'analyse égocentrée dans lesquelles l'utilisateur est un élément

central. Les études portent donc principalement sur ce qui se passe dans le réseau de l'utilisateur en question pour déterminer ses caractéristiques (intérêts) en utilisant la propriété d'influence sociale entre les utilisateurs et cet utilisateur central dans le réseau social. L'hypothèse sous-jacente est que le réseau social de l'utilisateur central va permettre de trouver des informations le concernant. Nos travaux s'appuient sur cette hypothèse.

Nous présentons dans ce qui suit, un état de l'art sur le profilage social selon trois orientations : profil social en filtrage social d'information, déduction d'attributs du profil de l'utilisateur à partir de son réseau social, et, enfin, modélisation générique du profil social de l'utilisateur.

Nous adoptons les notations suivantes pour la suite :

Utilisateur : dans cette section, nous utilisons le terme utilisateur pour désigner l'utilisateur central pour lequel on cherche à construire le profil social.

Individu : le terme individu est utilisé ici pour désigner les membres du réseau social de l'utilisateur.

Voisin social : désigne un individu en contact direct avec l'utilisateur (il existe un lien entre l'utilisateur et cet individu).

3.3.1. Filtrage social d'information

Dans le contexte de la recherche d'information, (Carmel et al., 2009) proposent un système de recherche d'information sociale qui réordonne les résultats de recherche en tenant en compte des informations du réseau social de l'utilisateur. Pour un utilisateur, son profil est construit lors de sa connexion au système. Le profil est défini d'une part, par une liste *T(u)* de mots clés associés aux contenus qu'il partage et d'autre part, par la liste *N(u)* des individus dans son réseau social. Les individus du réseau social de l'utilisateur peuvent être extraits à partir de ses contact directs (réseau de familiarité) et des individus qui partagent les mêmes activités que lui (réseau de similarité) : annoter la même ressource, être membre de la même communauté, commenter le même article de blog, etc.

Le score personnalisé d'un document retourné à partir d'une requête donnée par l'utilisateur est calculé en combinant d'une part, le score de ce document sans tenir compte de son profil utilisateur et d'autre part, le score basé sur son profil utilisateur comme présenté dans la formule suivante :

$$S_p(q,e|P(u)) = \alpha S_{np}(q,e) + (1-\alpha)[\beta \sum_{v \in N(u)} w(u,v) \cdot w(v,e) + (1-\beta) \sum_{t \in T(u)} w(u,t).w(t,e)] \quad (3.13)$$

Où $S_p(q,e|P(u))$ est le score personnalisé d'un document *e* pour la requête *q* avec le profil *P(u)* de l'utilisateur *u* donné. $S_{np}(q,e)$ est le score non personnalisé de e pour la requête *q*. *w(u,v)* est le poids de force de relation entre l'utilisateur *u* et un individu $v \in N(u)$. *w(v, e)* est le poids associé à *e* vis-à-vis de l'individu *v*. De la même manière, $w(u,t)$ représente le poids de *t* vis-à-vis de l'utilisateur *u* et $w(t,e)$ représente le poids de relation entre le terme *t* et le document *e*, donné par le système de recherche.

(Zeng, Yao et Zhong, 2009) proposent un système de recherche d'information sociale d'articles scientifiques sur les réseaux de co-publications scientifiques de la librairie DBLP. Les utilisateurs du système (auteurs) sont caractérisés par deux ensembles d'intérêts : leurs intérêts individuels calculés à partir des titres de leurs publications (i.e. profil utilisateur individuel) et

leurs intérêts calculés à partir des publications des individus de leur réseau social appelé « *group interests* » (i.e. profil social).

Le point intéressant de ce travail est qu'il s'intéresse à la dynamique des intérêts des utilisateurs. Les intérêts dans le profil utilisateur et les intérêts dans le profil social peuvent être calculés et ordonnés par des mesures temporelles par exemple :

- **le poids cumulé (*cumulative interests*)** est calculé en se basant sur le nombre d'apparitions des intérêts dans un intervalle de temps. Pour un auteur donné, si *i* est un intérêt (extrait à partir des termes dans les titres des publications), le poids cumulé de l'intérêt *i, TRI(i),* est calculé avec la formule ci-dessous :

$$TRI(i) = \sum_{j=1}^{n} m(i,j) \qquad (3.14)$$

  où *n* est le nombre d'intervalles (années) donné, et $m(i,j)$ est le nombre de publications relatives à l'intérêt *i* pendant l'année *j*.

- **le poids de rétention cognitive** se base sur le fait qu'une personne peut être amenée à changer le domaine de ses intérêts à long terme (phénomène d'oubli déjà observé en psychologie cognitive). Le calcul du poids de rétention est basé sur la fonction exponentielle ou celle de la loi de Poisson (*power law function*). Les deux fonctions donnent un poids plus important à des intérêts plus récents. Le poids de rétention cognitive exponentiel $RI_{Exp}(i)$ d'un intérêt *t(i)* dans un intervalle de temps est calculé avec la formule (3.15).

$$RI_{Exp}(i) = \sum_{j=1}^{n} m(i,j) * Ae_i^{-bT_i} \qquad (3.15)$$

  Le poids de rétention cognitive polynomial $RI_{Pow}(i)$ d'un intérêt *i* est calculé avec la formule (3.16)

$$RI_{Pow}(i) = \sum_{j=1}^{n} m(i,j) * AT_i^{-b} \qquad (3.16)$$

  Pour les deux formules, *n* est le nombre d'intervalles (années) donné. T_i représente la durée en année de l'apparition de *i*. $m(i,j) * AT_{t(i)}^{-b}$ représente la rétention totale de l'intérêt *i* pendant l'intervalle *j*. Les paramètres *A* et *b* sont déterminés expérimentalement. Pour la fonction exponentielle *A* et *b* sont fixés à 0.535 et 0.382 respectivement. Pour la fonction polynomiale A est fixé à 0.855 et *b* est fixé à 1.295.

Les intérêts dans le profil utilisateur et le profil social sont utilisés pour raffiner les résultats des requêtes de l'utilisateur sur le dataset *SwetoDLP* contenant plus de 615 000 auteurs via l'interface de recherche **DBLP Search Support Engine (DBLP-SSE)** développé par les auteurs. La Figure 3.8 représente la capture d'écran de cette interface.

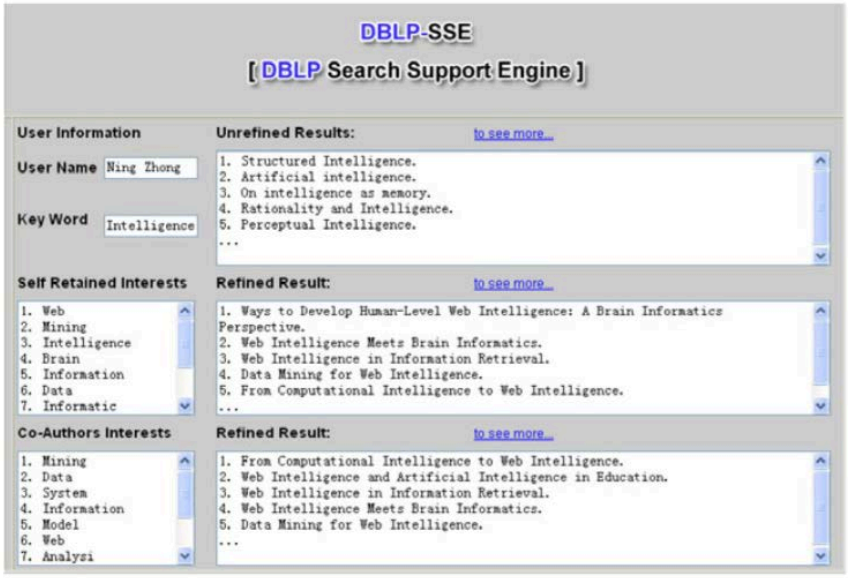


Figure 3.8 Capture d'écran de l'interface de recherche d'information personnalisée sur DBLP : DBLP search Support Engine (DBLP-SDE) (Zeng, Yao et Zhong, 2009)

Concernant l'évaluation par des retours utilisateurs faisant suite à ce travail, (Ren et al., 2010) montrent que 100% des participants trouvent que les résultats de recherche personnalisée en appliquant le poids de rétention des intérêts sont meilleurs que les résultats de recherche non personnalisée. 83.3% des participants trouvent que la personnalisation s'appuyant sur le profil utilisateur fournit de meilleurs résultats alors que 16,7% des participants trouvent que la personnalisation s'appuyant uniquement sur le réseau social fournit les meilleurs résultats. Les profils utilisateur paraissent plus pertinents pour les utilisateurs que le profil social (ce qui est normal car les informations partagées par l'utilisateur sont plus reliées à ses intérêts que celles partagées par ses co-auteurs). Cependant, le résultat de la personnalisation basée sur le réseau social nous montre que le profil social peut être réellement utile et pertinent en cas d'absence d'information dans le profil utilisateur individuel.

(Cabanac, 2011) propose un système de recommandation sociale de publications scientifiques pour les auteurs d'articles de recherche. Pour un auteur, ce système lui recommande des auteurs scientifiques qui sont similaires dans le but de parcourir les publications de ces auteurs similaires. Pour calculer la liste des auteurs similaires, deux mesures de similarité sont utilisées : la mesure sociale basée sur l'interaction sociale (*social similarity measure*) et la mesure basée sur la sémantique des termes représentant les intérêts des auteurs (*topical similarity measure*).

La mesure sociale est calculée à partir de deux types de graphe social : le graphe de co-auteurs (Figure 3.9, gauche) et le graphe d'affiliation calculé à partir des participations aux mêmes conférences (Figure 3.9, droite). Ce dernier permet d'exploiter les interactions potentielles entre auteurs dans la vie réelle lorsqu'ils peuvent se croiser en conférence. Une liste d'auteurs (liste sociale) similaire est calculée à partir de ces deux graphes sociaux. Pour qualifier les auteurs pertinents (similaires), trois mesures différentes sont utilisées :

- **la proximité** dans le graphe de co-auteurs est utilisée pour mesurer la proximité entre deux chercheurs. Cette mesure utilise l'inverse de la longueur du plus court chemin entre deux chercheurs.
- **la connectivité** dans le graphe de co-auteurs désigne la possibilité pour un chercheur d'accéder à d'autres chercheurs par un nombre d'intermédiaires le plus restreint

possible. Cette mesure est donnée par le nombre de chemins les plus courts entre deux chercheurs.

- **la probabilité de rencontre** dans le graphe de participation à des événements communs. Cette mesure est relative au nombre de participations à des événements communs entre deux chercheurs.

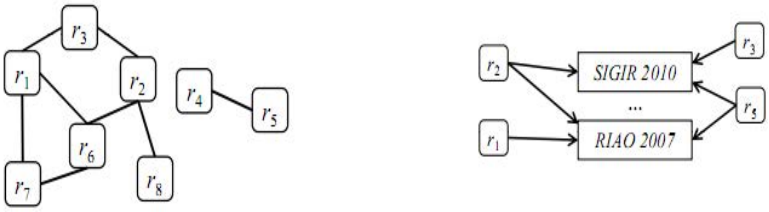


Figure 3.9 Graphe de co-auteurs (à gauche) et graphe de participations aux événements communs (à droite) (Cabanac, 2011)

Ces trois mesures sont combinées en utilisant les techniques de combinaison (*CombMNZ*) de scores décrits dans (Shaw et al., 1994) afin de définir la liste sociale d'auteurs.

La mesure basée sur la sémantique des termes permet, en parallèle de la mesure sociale, de construire une « liste d'auteurs similaires » à partir des auteurs dont le profil est similaire à l'utilisateur en utilisant les termes représentant les intérêts des auteurs (*Topical Similarity Measure*). Les profils des auteurs sont calculés par extraction des termes dans les titres de leurs publications (pondération via la mesure *tf*idf*). Pour déterminer si deux auteurs sont similaires, la fonction classique cosinus est utilisée.

Enfin, la liste finale des auteurs à recommander est finalement déterminée en combinant la liste sociale et la liste des auteurs similaires suivant les techniques de combinaison (*CombMNZ*) de scores définis dans (Shaw et al., 1994).

La Figure 3.10 présente les deux étapes de combinaison de la liste basée sur la mesure sociale et de la liste basée sur la mesure de sémantique des termes. La fonction intersection utilisée permet d'enlever les auteurs de la liste sociale qui n'ont pas d'intérêts en commun avec l'utilisateur avant de combiner les deux listes en utilisant la fonction *CombMNZ*.

Figure 3.10 Combinaison de la recommandation sociale (Cabanac, 2011).

Cette section s'est intéressée aux travaux existants dans le profilage social dans le domaine du filtrage social de l'information. Le but est de déduire des intérêts de l'utilisateur. La section suivante s'intéresse à une autre catégorie de travaux dans lesquels l'objectif est de déduire, à partir du réseau social, des attributs descriptifs de l'utilisateur.

3.3.2. Déduction d'attributs du profil de l'utilisateur

(Davis Jr. et al., 2011) proposent un algorithme pour déterminer la localisation (*location*) des utilisateurs sur Twitter en se basant sur la localisation de ses voisins directs. L'efficacité de ce

travail a été montrée empiriquement. Dans le même contexte, (Jurgens, 2013) utilise l'approche de propagation de label (*Label Propagation Approach*) pour prédire la localisation des utilisateurs sur Twitter et Foursquare[60]. Cette approche est itérative. A l'état initial, il existe des nœuds (utilisateurs) qui sont libellés par des attributs publics (localisation, post, ...) et ceux qui ne le sont pas (nœuds vides). Ensuite, chaque nœud vide se voit assigner le label le plus populaire parmi ces voisins directs et ce, jusqu'à ce que tous les nœuds du réseau soient libellés. Cependant, dans ces travaux, un seul attribut du profil (localisation) est considéré.

En se basant sur le même concept que (Jurgens, 2013), (Li, Wang et Chang, 2014) proposent une méthode de propagation pour déduire les attributs dans le profil de l'utilisateur (profession, éducation et localisation) sur le réseau LinkedIn à partir des relations dans le réseau social de l'utilisateur. Ce travail considère qu'il peut y avoir des bruits dans l'information qui peuvent impacter la pertinence de la méthode. Par exemple, concernant la déduction de l'attribut « métier », il se peut qu'un utilisateur se connecte avec des relations qui n'ont pas le même métier que lui. En se basant sur leur étude, sur 19 000 relations sociales étudiées sur LinkedIn, seulement 11% des utilisateurs connectés travaillent dans la même entreprise et seulement 18% d'entre eux ont étudié dans la même université. Pour pallier ce problème, les auteurs s'appuient sur le type de relations entre utilisateurs pour sélectionner le type d'attribut à propager de l'un à l'autre. Ils définissent ainsi des règles de propagation. Par exemple, pour un nœud v1, on ne doit lui propager l'attribut « métier » que depuis ses voisins qui ont la relation de type « collègue ». Les expérimentations réalisées par (Li, Wang et Chang, 2014) ont montré l'efficacité de leur proposition et montrent, dans un premier temps, l'avantage d'utiliser les informations à partir des voisins sociaux de l'utilisateur pour déduire les attributs manquants dans le profil utilisateur. Cependant, notons que nos objectifs sont principalement liés à l'extraction des intérêts de l'utilisateur et que les travaux cités précédemment ne l'ont pas abordée de façon concrète.

(Wen et Lin, 2010) se positionnent sur le problème de la qualité (pertinence) des intérêts extraits depuis le réseau social de l'utilisateur : compte tenu du volume d'information qui circule dans le réseau social de l'utilisateur, à quel niveau est la pertinence des intérêts extraits à partir de ses voisins sociaux ? Les auteurs proposent les deux contributions principales suivantes :

- combiner différents types de réseaux sociaux pour extraire les intérêts au lieu d'utiliser un seul réseau social comme dans la plupart de travaux. Cela permet de pallier le manque d'informations dans certains réseaux sociaux (problème de confidentialité et d'accès aux informations),
- proposer une technique qui permet de prédire la qualité des intérêts extraits à partir du réseau social de l'utilisateur. Cette prédiction permet de décider quand on peut prendre en considération des intérêts extraits à partir du réseau social de l'utilisateur et quand il ne le faut pas.

L'étude porte sur les contenus sociaux des 400 000 employés (utilisateurs) d'une très grande entreprise : leurs données de communication (mail, messagerie instantanée), leurs contenus partagés sur la plateforme de partage de contenu social (réseau social de marque-pages, blogs, fichiers partagés en public, …) et les informations produites dans le cadre de leur travail.

Les auteurs proposent, dans un premier temps, d'examiner à la fois les intérêts implicites et les intérêts explicites. Les intérêts explicites de chaque utilisateur sont collectés depuis la liste des intérêts renseignés par l'utilisateur dans son profil explicite dans la plateforme de réseau social.

[60] https://fr.foursquare.com/

Ces intérêts sont stockés sur un vecteur de termes avec un poids =1 par défaut car généralement les termes indiqués dans le profil des utilisateurs ne sont ni ordonnés ni pondérés.

Les intérêts implicites sont extraits à partir des contenus partagés par l'utilisateur. Pour chaque contenu, la technique d'allocation de Dirichlet latente (*Latent Dirichlet Allocation : LDA*) est appliquée pour en extraire des « topics » et construire un vecteur de « topics » pondérés. Les « topics » de différents vecteurs sont ensuite agrégés en prenant en compte l'importance de la source de contenus utilisée. Par exemple, si on considère que le poids d'une information issue d'un mail est considéré plus important que celui d'une information issue du réseau de partage de fichiers, on donne un poids plus important à l'information issue d'un mail (par exemple 0.8) par rapport au poids de l'information issue du partage de fichiers (par exemple 0.2). Supposons que le poids initial d'un « topic » *a* extrait du contenu d'un mail de l'utilisateur ait pour poids 0.5 et celui du topic *c* extrait du réseau de partage de fichiers ait un poids de 0.8, le poids final du topic *a* vaudra donc 0.5*0.8 = 0.4 et le poids final du topic *c* vaudra donc 0.8*0.2 = 0.16. Enfin, une matrice *S (UxT)* est construite pour stocker les poids agrégés des *T* « topics » extraits pour les *U* utilisateurs.

Ensuite, dans le but de déduire les intérêts depuis les voisins de chaque utilisateur, les auteurs proposent d'appliquer un modèle d'auto-corrélation de réseau (*Network Autocorrelation Model*) (Neville, Simsek et Jensen, 2004) pour ne sélectionner que les « topics » associés aux voisins sociaux considérés significatifs pour l'utilisateur. Seuls les voisins qui sont au maximum à distance 3 de l'utilisateur sont pris en compte. Pour cela, une matrice *Z (UxN)* est construite pour stocker le poids des relations entre les *U* utilisateurs et les *N* « topics » extraits de ses voisins. Pour un utilisateur *i*, le poids $z_{i,j}$ d'un topic *j* est calculé en se basant, d'une part, sur le poids $s_{k,j}$ de chaque voisin *k* sur *j* et d'autre part, sur le poids d'influence $w_{k,i}$ entre *k* et l'utilisateur *i*. Notons que *i* représente le $i^{ème}$ utilisateur tel que $i \in [O; U]$ et j représente le $j^{ème}$ topic tel que $j \in [O; N]$. Le calcul de poids $z_{i,j}$ est présenté dans la formule (3.17).

$$z_{i,j} = \sum_{k=1}^{U} w_{k,i} \cdot s_{k,j}$$

(3.17)

Le poids d'influence $w_{k,i}$ est défini en fonction de la distance sociale entre *k* et *i*. Selon la caractéristique du réseau, différents types de fonctions pour calculer le poids d'influence peuvent être utilisés. Ici, le poids $w_{k,i}$ est calculé en appliquant une fonction exponentielle de la distance sociale entre *k* et *i* (3.18):

$$w_{k,i} = exp^{-distance(k,j)}$$

(3.18)

La distance sociale entre *k* et *j* est calculée en se basant le nombre d'interactions entre les nœuds qui font partie du chemin le plus court entre *k* et *i*. La distance sociale est définie par la formule (3.19).

$$distance(i,j) = \sum_{k=1}^{K} \frac{1}{strength(v_k, v_{k+1})}$$

(3.19)

Les $v_{1,\ldots,}v_k$, représentent les nœuds qui font partie du plus court chemin entre *i* et *j*. La valeur de $strenght(v_k, v_{k+1})$ est calculée par le nombre de communication entre *k* et *k*+1 normalisé par le nombre maximum d'interactions dans le réseau (3.20).

$$strenght(v_k, v_{k+1}) = \frac{log\ (nbcomm(v_k, v_{k+1})}{max_{k+1} log(nbcomm(v_k, v_{k+1})}$$

(3.20)

Ils proposent ensuite un modèle de prédiction de qualité des intérêts extraits à partir des voisins sociaux de chaque utilisateur en prenant en compte cinq facteurs différents sur l'utilisateur lui-même et sur l'ensemble de ses voisins sociaux : la vigueur (nombre de contenus partagés), le degré d'entrée, le degré de sortie, la centralité d'intermédiarité et le rôle de management dans l'organisation. Le modèle est construit en appliquant la technique de régression par machines à vecteurs de support (*SVM*). Le modèle peut être exploité pour fournir les informations aux mécanismes qui peuvent bénéficier de cette méthode : le mécanisme ne peut appliquer cette méthode que lorsque le score de prédiction donne une bonne estimation de la qualité.

Dans la continuité de leur travail, les auteurs proposent dans (Wen et Lin, 2011), une méthode pour améliorer leur méthode de déduction des intérêts extraits. Ils s'appuient sur l'hypothèse qu'un utilisateur n'est influencé par des voisins sociaux que sur les « topics » qui l'intéressent. Pour cela, en plus d'appliquer la corrélation sociale pour déduire les « topics » depuis ceux des voisins de l'utilisateur, ils appliquent la corrélation des attributs (topics) pour ne sélectionner seulement que ceux qui sont appropriés pour l'utilisateur en appliquant le coefficient de corrélation de Pearson (*Pearson Correlation Coefficient*) (Benesty et al., 2009). Les expérimentations sur les mêmes datasets utilisés dans (Wen et Lin, 2010) montre que la méthode de corrélation des attributs améliore de 76% la méthode existante qui applique seulement la corrélation sociale.

Les deux sections précédentes décrivent des travaux de profilage de l'utilisateur. Néanmoins, elles sont généralement mises en œuvre dans un contexte particulier pour lequel elles sont efficaces. De ce fait, elles sont difficilement généralisables ou réutilisables directement dans d'autres contextes. Comme déjà précisé en début de section 3.3, les mécanismes de construction de profil et les mécanismes qui exploitent ce dernier sont intimement liés. En particulier, le profil répond aux besoins du mécanisme qui l'exploite et est construit pour celui-ci spécifiquement. La section suivante aborde des travaux de notre équipe concernant la construction de profils sociaux utilisateurs indépendants des mécanismes d'utilisation de ces profils, et donc génériques.

3.3.3. Construction de profil utilisateur générique

Les travaux de (Tchuente et al., 2013), réalisés dans notre équipe, portent sur l'étude de la dérivation du profil utilisateur à partir des communautés de son réseau égocentrique. Dans un but de généricité et de prise en compte efficace du réseau social de l'utilisateur, ce travail propose un modèle générique de profil utilisateur constitué de deux dimensions, elles-mêmes constituées d'un ensemble d'attributs :

- **la dimension utilisateur** contient les éléments construits à partir des informations et interactions propres de l'utilisateur avec le système. Cette dimension peut être construite par les approches classiques de construction de profil présentées au chapitre 2. Cette dimension est considérée comme la plus importante du profil et doit

être utilisée en priorité par les mécanismes de personnalisation d'informations, de recommandation … ,

- **la dimension sociale** contient les éléments construits à partir des informations et interactions des communautés dans le réseau égocentrique de l'utilisateur. 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 mêmes mécanismes. Elle peut être utilisée, par exemple, lorsque le profil utilisateur manque d'information sur certaines catégories d'intérêts, ou bien lorsque l'on souhaite élargir les intérêts de l'utilisateur (apport d'intérêts nouveaux par rapport au profil individuel) ou bien au contraire les restreindre (intersection des deux profils).

Pour la construction de la dimension sociale, qui est la partie qui nous intéresse en particulier, (Tchuente et al., 2013) ont proposé un processus de construction du profil social CoBSP (*Community Based Social Profiling*) à partir des communautés du réseau égocentrique de l'utilisateur. Le choix d'utiliser le réseau égocentrique comme source d'informations se justifie par des travaux existants en sciences sociales (Bhattacharyya, Garg et Wu, 2011 ; Sinha et Swearingen, 2002) qui ont démontré l'intérêt du réseau égocentrique d'un utilisateur (égo) pour lui faciliter l'accès à de nouvelles informations et accroître ainsi son capital social. Le réseau égocentrique dans le contexte de ce travail se base sur les relations existantes entre l'utilisateur (égo) et ses alters en considérant qu'ils se connaissent également dans la vie réelle.

Alors que la plupart des travaux dans ce contexte s'appuient sur la sélection des individus qui ont des liens forts avec l'utilisateur (approches qualifiées de « autoritaires »), ce travail suppose que les individus qui ont des liens faibles avec l'utilisateur peuvent également fournir des informations potentiellement intéressantes pour l'utilisateur. Ils proposent donc une approche d'extraction des intérêts sociaux basée sur les communautés (approche qualifiée de « affinitaire »). De façon globale, dans ce processus, on extrait les intérêts de chaque communauté dans le réseau égocentrique de l'utilisateur avant de les agréger pour obtenir les intérêts de la dimension sociale. Le profil de la communauté est déduit en réunissant (agrégeant) les informations à partir de la dimension utilisateur des profils des individus de la communauté. Le poids de chaque intérêt est calculé en se basant sur la caractéristique de la communauté qui peut être caractérisée, d'un côté, par sa structure et, d'autre part, par une « dimension sémantique ». Notons que dans ce travail, le poids de la structure d'une communauté et le poids de la sémantique de cette communauté dans le profil de l'utilisateur peuvent varier et doivent être évalués (fixés) expérimentalement (détaillé ci-après).

L'étude propose ensuite un processus de dérivation du profil social de l'utilisateur à partir des communautés extraites de son réseau égocentrique. Nous expliquons brièvement le processus *CoBSP* en quatre étapes successives décrites ci-après et illustrées Figure 3.11:

- la première étape consiste à extraire, à partir du réseau égocentrique d'un utilisateur, les communautés de ce réseau en utilisant l'algorithme *iLCD* proposé par (Cazabet, Amblard et Hanachi, 2010). Cet algorithme se base sur la structure du réseau pour extraire des communautés et prend en compte le recouvrement des communautés. Il a été évalué et considéré adapté au problème (Tchuente et al., 2013),
- la deuxième étape consiste à calculer le profil de chaque communauté détectée dans la première étape (vecteur de termes/éléments pondérés). Le profil d'une communauté peut être calculé par agrégation des informations de tous les individus qui en font partie (ici, agrégation des poids associés aux éléments présents dans la dimension utilisateur de chaque individu),

- la troisième étape consiste à pondérer le profil de chaque communauté en se basant sur une caractérisation structurelle et/ou une caractérisation sémantique. La caractérisation structurelle d'une communauté se base sur la centralité de degré (*degree centrality*). Selon cette mesure, la communauté qui possède le plus grand nombre de connexions directes dans le réseau est caractérisée comme étant la plus importante par rapport aux autres communautés. La caractérisation sémantique d'un élément de profil d'une communauté consiste à rechercher sa spécificité par rapport aux autres communautés en se basant sur la mesure de pondération telle que *tf* ou *tf-idf* entre les intérêts des communautés. L'objectif est d'obtenir une caractérisation différenciée entre communautés. Les deux caractérisations sont combinées pour obtenir une caractérisation unique, appelée *sémantico-structurelle*. La caractérisation *sémantico-structurelle* de chaque élément *e* du profil d'une communauté c_i sera calculée par la formule (3.21) suivante :

$$caractérisation\_finale(e, c_i) = \alpha * struct(c_i) + (1 - \alpha) * sem(e, c_i) \quad (3.21)$$

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

- la quatrième étape permet de dériver les intérêts du profil social par combinaison des différents poids associés à un intérêt à partir de toutes les communautés en utilisant une fonction de combinaison linéaire *Lin_CombMNZ* (Hubert, Loiseau et Mothe, 2007).

Les expérimentations menées dans les travaux de (Tchuente et al., 2013) sur le réseau de publications scientifiques DBLP et le réseau Facebook ont montré l'efficacité de cette technique par rapport aux techniques d'extraction des intérêts qui calculent le poids des intérêts par une approche individuelle. Sur le réseau DBLP, les meilleurs résultats sont trouvés en particulier dans le cas des utilisateurs qui possèdent beaucoup d'individus dans leur réseau égocentrique (au moins 50 co-auteurs).

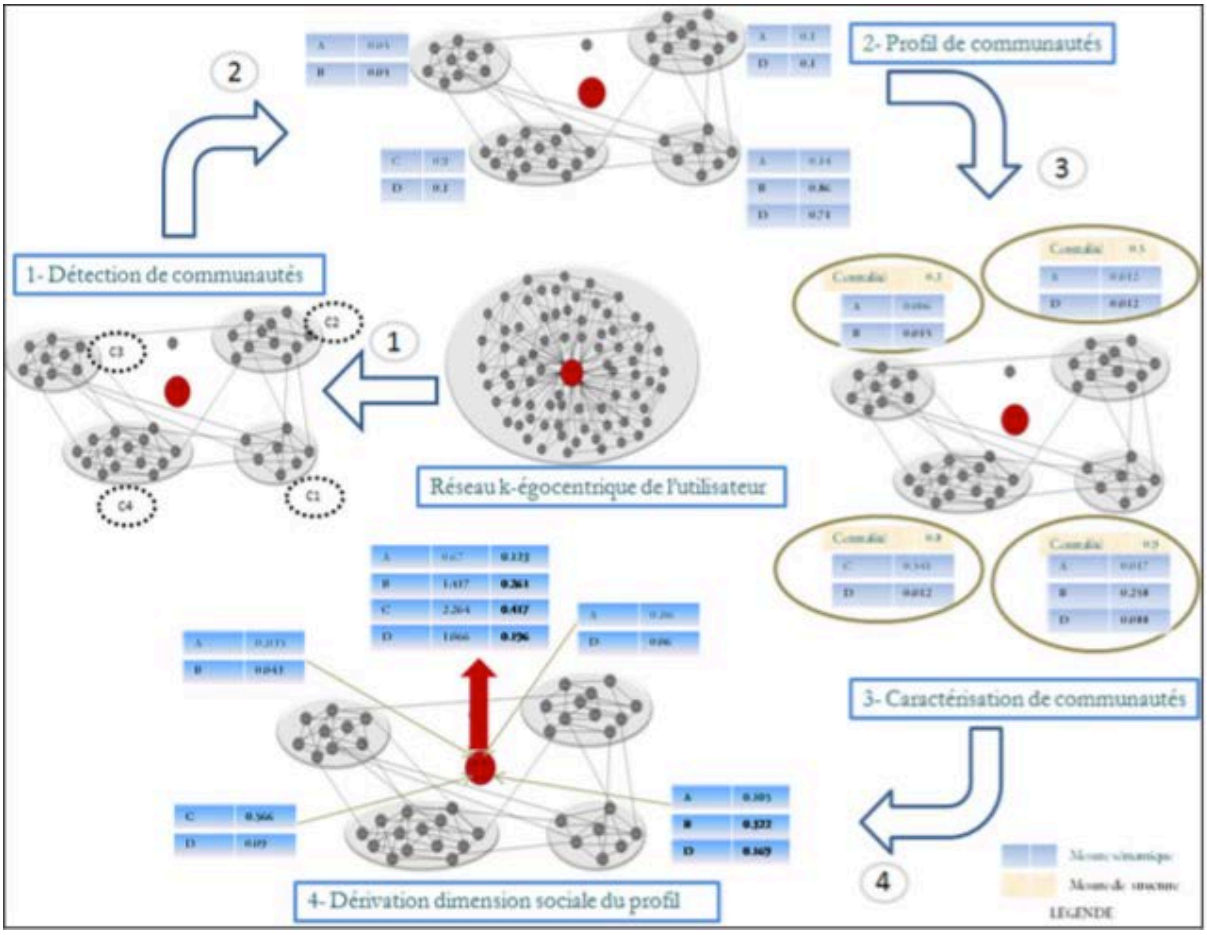


Figure 3.11 Illustration du processus de dérivation de la dimension sociale à partir de communauté dans le réseau égocentrique de l'utilisateur (Tchuente et al., 2013)

3.4. Synthèse

Dans ce chapitre, nous avons présenté les éléments fondamentaux liés au profilage social : types et caractéristiques des réseaux sociaux, éléments de réseaux sociaux d'un point de vue sociologique et mathématique, travaux en analyse de réseaux sociaux ainsi que travaux existant en profilage social. Ces travaux montrent que les informations du réseau social de l'utilisateur peuvent être un moyen de déduire les intérêts de l'utilisateur pour compléter des informations manquantes dans le profil utilisateur et qui peuvent être utiles dans le cas d'utilisateurs peu actifs mais aussi pour fournir des informations supplémentaires sur l'utilisateur en se basant sur les intérêts de ses voisins sociaux.

La popularité sans cesse croissante des médias sociaux nous amène à considérer l'utilisation des réseaux sociaux numériques comme source d'information pour extraire les intérêts de l'utilisateur. En effet, ce type de réseau social devient une source très riche pour étudier et extraire les connaissances sur l'utilisateur. Nous nous positionnons dans ce contexte afin de relever les défis liés au profilage social dans le contexte d'un réseau social numérique. Ces défis sont présentés ci-après.

3.4.1. Défis dans le profilage social par rapport aux caractéristiques des réseaux sociaux

En se basant sur les caractéristiques des réseaux sociaux décrites dans la section précédente, nous pouvons mettre en avant différents défis associés au profilage social et détaillés ci-après.

- **Fiabilité des sources d'informations**

Les informations exploitées pour construire le profil social ne proviennent pas de l'utilisateur lui-même mais de ses voisins sociaux ou individus de son réseau social. Cela soulève la problématique suivante : ces informations sont-elles pertinentes vis-à-vis des intérêts réels de l'utilisateur ? On peut formuler ce problème en termes de fiabilité des relations et des informations.

En ce qui concerne les relations, les utilisateurs ont plusieurs types de relations qu'ils maintiennent dans le réseau social. Certains types de relations peuvent être plus ou moins importants pour l'utilisateur. Par exemple, Alice et Carol sont toutes les deux les amies de Bob. Bob communique plus souvent et partage beaucoup de temps avec Alice. Il paraît évident dans ce cas, qu'Alice aura plus d'influence sur Bob que Carol. Les informations rattachées à Alice paraissent plus significatives pour inférer des informations liées aux intérêts de Bob. Si on ne prend pas en compte cette observation, on risque d'inférer des informations à partir de celles de Carol qui s'avèrent moins pertinentes. La pertinence est ici déduite du fait que l'intensité des relations entre Bob et Alice montrent qu'ils partagent activement des intérêts ; cela relève du phénomène d'homophilie (cf. section 3.2.1.3) et des liens forts (cf. section 3.2.1.4).

Le problème de la fiabilité des sources d'informations peut être observé particulièrement, dans le contexte des media sociaux où les utilisateurs ne sont pas obligés de se connaître dans la vie réelle pour communiquer entre eux. Même si les études ont prouvé que les relations créées dans les RSNs sont liées aux intérêts réels ou comportements sociaux de l'utilisateur dans la vie réelle, il faut également noter qu'il peut exister des relations qui vont être acceptées par hasard. On risque alors d'induire du bruit d'information lors de la création du profil social.

En termes de fiabilité d'information, les utilisateurs ont tendance à avoir plusieurs relations dans le réseau avec un objectif particulier pour chaque relation. Ceci signifie qu'un utilisateur n'est pas obligé de s'intéresser à toutes les informations auxquelles ses voisins s'intéressent. Si on reprend l'exemple de Bob, il aime partager des informations sur les restaurants avec Alice mais ne s'intéresse pas forcement aux sites de shopping comme Alice. En même temps, il s'intéresse aux informations sur le foot que Carol partage mais il n'aime pas forcement les autres sports auxquels s'intéresse Carol.

- **Evolution des relations et des informations**

Nous constatons que les intérêts d'une personne dans la vie réelle sont amenés à évoluer dans le temps, en particulier, dans le cas des relations entretenues et des informations consultées dans les RSNs. En effet, dans de tels réseaux, l'information partagée évolue sans cesse du fait des interactions sociales en ligne (partage, échange d'informations) qui génèrent (plus facilement) un volume important d'informations volatiles. Ainsi, les intérêts qui sont extraits à un moment donné peuvent ne plus être significatifs ultérieurement.

Dans ce contexte, et comme évoqué dans la section précédente, un problème qui se pose est la pertinence des liens de l'utilisateur avec les membres de son RSN ainsi que la pertinence des informations qu'il partage mais aussi de celles que partagent les membres de son réseau. Dans le cas des RSNs, les utilisateurs peuvent créer des contacts en ligne sans forcément connaître les personnes dans la vie réelle. Il se peut également qu'un utilisateur ne supprime pas des contacts dans ses RSNs même si ces contacts deviennent moins importants pour lui (Pempek,

Yermolayeva et Calvert, 2009). Dans ce cas, on risque donc d'avoir des « bruits » dans les relations dans le réseau social de l'utilisateur qui peuvent impacter l'efficacité du processus de construction du profil social. De la même manière, pour les informations partagées dans un réseau social, avec un grand volume d'informations disponible, certaines peuvent devenir au fil du temps obsolètes pour l'utilisateur. On risque d'avoir là aussi des « bruits » dans les informations que l'on utilise pour extraire les intérêts pour construire le profil social.

Cela montre que l'on ne peut pas prendre en compte (ou donner la même importance à) toutes les informations existant dans les RSNs pour refléter les intérêts d'un utilisateur à un moment donné. Par rapport à la section précédente, à la question de la fiabilité des sources d'informations s'ajoute ici la question de la fraîcheur ou de l'obsolescence de ces sources par rapport à leur date d'utilisation. En d'autres termes, il faut prendre en compte l'évolution des sources d'informations (dans notre contexte, il s'agit de l'évolution du réseau social de l'utilisateur) dans le processus de construction du profil social de l'utilisateur : évolution des relations d'une part et évolution des informations d'autre part. Pour cela, il est nécessaire d'étudier des modèles et méthodes qui permettent de prendre en compte ces deux types d'évolution afin d'extraire les intérêts les plus pertinents pour l'utilisateur. Il faudra d'une part, qualifier les sources d'informations (individus) les plus pertinentes pour l'utilisateur. D'autre part, à partir des informations partagées par ces individus sélectionnés, il ne faudra sélectionner que les informations les plus pertinentes vis-à-vis de l'utilisateur. Le chapitre 1 a déjà donné des indications pour la prise en compte de l'évolution dans la cadre de la construction d'un profil utilisateur. Ces approches devront être étendues ou ajustées pour s'appliquer dans le cadre de construction du profil social.

- **Hétérogénéité des informations utilisées**

Avec la diversité des RSNs, comme précisé section 3.1.2, il existe différents types de réseaux sociaux selon les objectifs d'utilisation, le type de relations mis en place entre individus, le type d'informations échangées ou même selon les caractéristiques de l'évolution du réseau.

En ce qui concerne les objectifs d'utilisation, certains réseaux sociaux sont utilisés dans un contexte général, les informations partagées sont alors générales (micro-blogs, sites de réseautage social, …). A l'inverse, dans d'autres réseaux sociaux, les informations sont plus ciblées et restreintes à un domaine (réseau de publications scientifiques, forums de discussion spécialisés, réseaux professionnels, …). Les utilisateurs possèdent également des comportements différents. Par exemple, sur Twitter, certains utilisateurs partagent des informations correspondant à leurs intérêts, ou à leur contexte de travail, ou à des informations politiques, alors que d'autres utilisent ce réseau dans un objectif personnel et ont tendance à partager des informations de leur vie quotidienne.

En ce qui concerne les types des relations, ceux-ci varient selon le réseau social. Par exemple sur un réseau de partage d'information comme Twitter ou Instagram, nous pouvons trouver les relations de « suiveur d'informations ». Un utilisateur établit une relation (suivre) avec un autre utilisateur s'il veut suivre les informations que l'autre partage. La relation n'est pas forcement réciproque donc cela ne montre pas vraiment la confiance de chaque utilisateur envers l'autre. A l'inverse, sur un réseau de réseautage tel que Facebook, les relations sont plus restreintes et sont généralement créées en se basant sur la connaissance qu'a chaque utilisateur de l'autre. La création de lien se fait par la demande d'ajout d'amis et doit être acceptée par la personne demandée. La relation créée est réciproque (bidirectionnelle). Dans un réseau de publications scientifiques par exemple, la relation se fait quand deux utilisateurs publient ensemble ce qui montre bien le fait qu'ils ont des intérêts en commun.

En ce qui concerne le type des informations échangées, chaque RSN offre la possibilité aux utilisateurs de partager ou d'annoter différents types d'informations (image, vidéo, texte …).

Par exemple, dans Facebook, les utilisateurs peuvent partager du texte, des images, des vidéos. Ils peuvent également attribuer la mention « aimer » (*like*) ou laisser des commentaires dans les contenus partagés par d'autres personnes. Depuis les dernières versions, les utilisateurs peuvent même exprimer leur émotion sur un post (aimer, adorer, mécontent, triste). Dans Twitter, les utilisateurs peuvent partager des textes courts (tweets), images, vidéos, annoter les *posts* par des *hashtags*. Cela dénote une grande hétérogénéité des informations accessibles.

En ce qui concerne les caractéristiques de l'évolution du réseau, l'évolution des relations et des informations dans les RSNs est différente selon le type de réseau social. Sur les sites de partage d'informations comme Twitter, les informations sont partagées sous la forme de flux d'informations. Les utilisateurs actifs partagent fréquemment leurs billets (tweets). Cela peut être toutes les heures, ou bien toutes les minutes s'il s'agit d'un compte de diffusion d'informations en direct. De ce fait, dans ce type de réseau social, l'évolution des informations est généralement rapide. Les « topics » de discussion dans le réseau peuvent changer très vite. L'évolution des liens dans Twitter (création, suppression) est moins rapide que celle des informations compte tenu du type de réseau qui est plutôt un réseau de partage d'information. Cependant, cette évolution peut montrer des irrégularités ou des à-coups suite à des événements importants. Par exemple, la coupe du monde de football qui entraîne les abonnements de fans de foot aux comptes des chaînes des journaux sur la coupe du monde, aux comptes officiels d'une équipe de foot ou bien aux comptes personnels des joueurs. Dans un réseau social de collaborations scientifiques, l'évolution des « topics » de recherche ainsi que des créations de liens est plutôt lente : ces créations de liens se produisent quand deux scientifiques publient ensemble ou participent aux mêmes conférences (avec éventuellement une longue période entre 2 événements) (Aggarwal et Subbian, 2014).

De par l'hétérogénéité des réseaux sociaux, il est difficile de trouver un processus de construction du profil social qui puisse être appliqué à tous les types de réseaux. Une technique qui donne de bons résultats sur un réseau social pourrait ne pas marcher pour d'autres. Il s'avère donc important de prendre en compte le type et les caractéristiques du réseau social de l'utilisateur dans le processus de construction du profil social.

3.4.2. Point sur les techniques existantes

Nous présentons le Tableau 3.1 dans lequel nous comparons des travaux associés aux défis de la construction du profil social et présentés dans ce chapitre.

Travaux	Objectif	Prise en compte des défis				
		Pertinence des sources d'informations		Evolution		Hétérogénéité
		Informations	Relations	Informations	Relations	
Wen et Lin 2010	Déduction des attributs	✗	✓	✗	✗	✓
Wen et Lin 2011	Déduction des attributs	✓	✓	✗	✗	✗
Zeng et al, 2009	Recherche d'information sociale	✓	✗	✓	✗	✗
Cabanac, 2011	Recommandation sociale	✓	✓	✗	✗	✗
Tchuente, 2013	Profil social générique	✓	✓	✗	✗	✗

Tableau 3.1 Comparaison des travaux associés à la construction du profil social

Nous constatons que la gestion de la pertinence de sources d'informations est un problème qui a déjà été abordé. Nous nous intéressons plus particulièrement aux problèmes de la prise en compte de l'évolution du réseau social de l'utilisateur dans le processus de construction du profil social et de la prise en compte de l'hétérogénéité des types et des caractéristiques des réseaux sociaux beaucoup moins traités dans la littérature.

L'ensemble des travaux menés dans le domaine du profilage social ont montré empiriquement leur efficacité. Cependant, nous constatons que la plupart de ces travaux sont mis en œuvre sans prendre en compte l'évolution du réseau social et restreints à un seul type de réseau (un seul réseau en général). Les problèmes de l'évolution des intérêts de l'utilisateur et l'hétérogénéité des informations ne sont pas beaucoup abordés. Seul le travail de (Zeng, Yao et Zhong, 2009) a traité le problème de l'évolution des intérêts de l'utilisateur et propose d'appliquer une fonction temporelle pour mesurer l'importance des informations en fonction de la date de leur publication. Toutefois, l'évolution des relations n'est pas abordée dans ces travaux.

Pour traiter le problème d'hétérogénéité des informations et des relations dans le réseau social de l'utilisateur, les travaux de (Wen et Lin, 2010) étudient l'efficacité de leur méthode d'inférence des intérêts à partir du réseau social selon plusieurs caractéristiques du réseau social de l'utilisateur. Cependant, cette étude repose seulement sur les caractéristiques topologiques d'un seul réseau social. La méthode n'est pas évaluée sur d'autres réseaux sociaux. De plus, celle-ci est effectuée sans prendre en compte l'évolution du réseau social.

Nous présentons, dans le chapitre suivant, notre contribution visant à apporter des éléments de réponses à ces limitations.

4. CONTRIBUTION : METHODE TEMPORELLE POUR LA CONSTRUCTION DU PROFIL SOCIAL DE L'UTILISATEUR

Dans ce chapitre, nous présentons, dans un premier temps, notre positionnement vis-à-vis des travaux existants, ensuite nous présentons les notions et concepts du profil social de l'utilisateur sur lesquels s'appuie notre travail. Enfin, nous présentons notre méthode de prise en compte de l'évolution du réseau social dans la construction profil social ainsi que l'intégration de cette méthode dans les processus existants.

4.1. Positionnement

L'objectif de départ de nos travaux repose sur l'étude de méthodes et techniques de profilage social permettant d'extraire les intérêts et construire un profil social de l'utilisateur pertinent. Celui-ci pourra être exploité comme un profil complémentaire au profil utilisateur existant :

- pour compléter les informations manquantes dans le profil utilisateur et qui peuvent être utiles dans le cas d'utilisateurs peu actifs,

- pour fournir des informations supplémentaires sur l'utilisateur en se basant sur les intérêts de ses voisins sociaux.

Comme décrit dans la partie l'état de l'art, les techniques de profilage social ne sont pas centrées uniquement sur l'utilisateur mais également sur les contenus et les caractéristiques de son réseau social.

Nous nous positionnons dans le contexte des réseaux sociaux numériques qui possèdent un caractère évolutif et hétérogène. Le caractère évolutif des réseaux sociaux peut poser le problème de la pertinence des intérêts du profil social construit en particulier si l'on ne prend pas en compte l'évolution de ces réseaux, on risque d'extraire des intérêts obsolètes pour l'utilisateur. Le caractère hétérogène peut poser problème dans l'utilisation d'une technique spécifique de profilage social : une technique peut être efficace dans un type de réseau social mais pas sur d'autres types de réseaux.

D'après les études de l'état de l'art dans la section 3, l'ensemble des travaux menés dans le domaine du profilage social ont montré empiriquement leur efficacité. Cependant, nous constatons que la plupart de ces travaux ne prennent pas en compte l'évolution du réseau social comme un facteur important et sont restreints à un type particulier (expérimentations effectuées sur un seul réseau en général).

L'objectif principal des travaux dans cette thèse est de dépasser ces limites pour proposer des méthodes et techniques pour construire un profil social de l'utilisateur pertinent et à jour, et cela, dès sa première construction. Nous envisageons d'améliorer l'efficacité des processus de construction du profil social existants en intégrant la prise en compte de la caractéristique évolutive du réseau social de l'utilisateur.

Comme l'efficacité des processus de profilage social dans la littérature a déjà été prouvée, nous reprenons les processus existants et nous proposons d'y intégrer une méthode de prise en compte du temps de l'évolution du réseau social pour améliorer la pertinence des résultats obtenus. Plus précisément, nous distinguons deux approches différentes d'extractions des intérêts de l'utilisateur à partir de son réseau social : l'approche basée sur les individus et l'approche basée sur les communautés. La méthode de prise en compte du temps proposée sera intégrée dans les deux approches pour en évaluer l'efficacité.

Pour pallier le problème de caractéristiques hétérogènes du réseau social, nous faisons en sorte que la méthode proposée soit la plus générique possible pour pouvoir l'appliquer dans différents types de réseaux. Nous mettons en œuvre la méthode proposée sur deux réseaux sociaux différents : *DBLP* qui est un réseau de publications scientifiques et *Twitter* qui est un réseau de micro-blogs. Ces deux réseaux sociaux possèdent différentes caractéristiques en termes d'objectif d'utilisation, de type d'informations partagées, de type de relations et interactions entre les individus et enfin de caractéristique évolutive. Ces expérimentations permettent d'étudier l'impact des caractéristiques du réseau social sur la pertinence du processus de construction du profil social proposé.

La pertinence du profil social peut être mesurée en comparant les intérêts contenus dans le profil social construit avec les intérêts réels de l'utilisateur. Autrement dit, une technique de profilage social efficace doit être capable de retourner un profil social qui se rapproche le plus du profil réel de l'utilisateur. Nous admettons dès le départ que c'est un défi difficile voire même impossible de produire un profil social qui soit à 100% pertinent pour l'utilisateur. Il est de même difficile de construire un profil social aussi pertinent que le profil individuel de l'utilisateur dont les intérêts sont extraits à partir de ses propres activités et qui reflètent donc, de façon plus significative, ses préférences et/ou intérêts. L'objectif est donc de produire un profil social qui soit le plus proche possible du profil individuel de l'utilisateur.

Nous allons tout d'abord donner une définition générale du profil social.

4.2. Définition générale du profil social

Avant de présenter notre méthode de prise en compte des caractéristiques évolutives et hétérogènes du réseau social dans le processus de construction du profil social, nous définissons d'abord ce qu'est un profil social dans notre contexte : le modèle considéré, la représentation du profil, les notations nécessaires pour la suite et, enfin, le principe général du processus de construction du profil inspiré de (Tchuente, 2013).

4.2.1. Modèle et représentation du profil social

Nous présentons dans la Figure 4.1, le modèle générique du profil utilisateur et du profil social inspiré de (Tchuente, 2013). Comme considéré dans le travail de (Tchuente, 2013), ce modèle de profil social est générique. Cette généricité permet d'abstraire ce modèle des différents mécanismes de personnalisation d'informations (filtrage social) et des différents types de réseaux sociaux existants. Cette abstraction lui permet d'être réutilisable indépendamment du mécanisme qui l'exploitera et du type de réseaux sur lequel il sera appliqué.

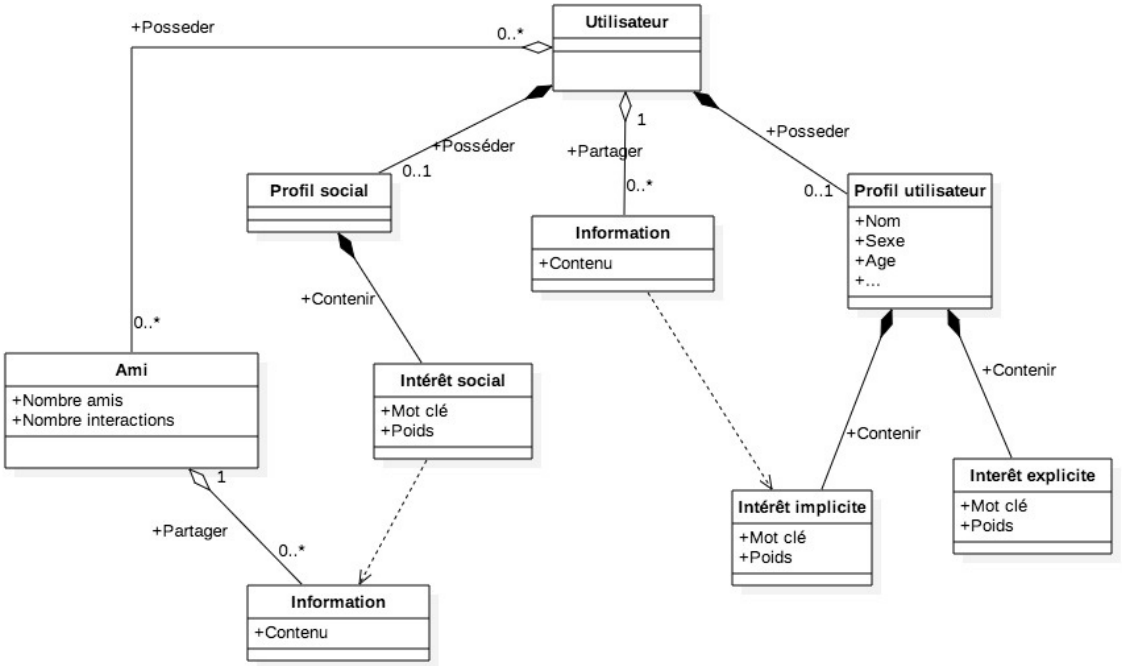


Figure 4.1 Modèle du profil utilisateur et du profil social

Le profil utilisateur représente les informations propres à l'utilisateur. Ces dernières peuvent contenir différents types d'attributs, comme par exemple, des informations géographiques ou démographiques (nom, âge, genre), ainsi que la liste des intérêts explicites indiqués par l'utilisateur ou les intérêts implicites extraits à partir des contenus qu'il partage.

Le profil social contient uniquement la liste des intérêts extraits depuis les informations du réseau social de l'utilisateur. Plus précisément, le réseau social considéré est le réseau égocentrique de l'utilisateur (ego) qui ne prend en compte que les voisins sociaux (alters) à distance 1 de l'utilisateur (ses voisins directs). Les intérêts contenus dans le profil social sont

donc extraits à partir des informations partagées par ses voisins sociaux. L'unité d'intérêt de base correspond à un élément (terme extrait d'un titre, d'un document, etc.). Le profil social est donc représenté par une liste d'intérêts sous forme d'un vecteur de termes pondérés.

Nous présentons ci-après les différentes approches de construction du profil social qui seront utilisées dans nos travaux.

4.2.2. Approches de construction du profil social

Nous distinguons les techniques de construction du profil social de l'utilisateur selon deux grandes approches : celle basée sur les individus et celle basée sur les communautés.

- **Approche basée sur les individus** : dans l'approche de construction du profil basée sur les voisins sociaux, les intérêts d'un utilisateur donné sont extraits à partir des informations de ses voisins sociaux considérés de manière individuelle. Cette approche est qualifiée d'autoritaire. A partir des informations de chaque individu du réseau social de l'utilisateur, les intérêts sont extraits et calculés en se basant sur les caractéristiques propres de chaque individu. Le poids accordé à un individu dans le calcul est lié à ses caractéristiques propres, indépendamment des autres individus. Les travaux associés à cette approche qui nous paraissent importants sont : (Zeng, Yao et Zhong, 2009), (Wen et Lin, 2010), (Wen et Lin, 2011), (Cabanac, 2011).
- **Approche basée sur les communautés** : dans l'approche de construction du profil basé sur les communautés, les intérêts d'un utilisateur sont extraits à partir de chaque communauté extraite depuis son réseau social. Les intérêts sont donc extraits et calculés en se basant sur les caractéristiques de chaque communauté en non pas de chaque individu. Nous pouvons citer le travail de (Tchuente et al., 2013) issu de cette approche.

4.2.3. Définition des termes et notations

Tout au long de ce chapitre, nous présentons nos contributions en utilisant les termes et les notations suivants :

4.2.3.1. Définition des termes utilisés

Etant donnée ***u*** représentant l'utilisateur central dont on veut construire le profil social (ego), son réseau égocentrique G'(*u*) est composé de ses voisins sociaux ainsi que des relations et des interactions entre eux.

Dans le graphe *G'(u)*, *u* est le nœud central relié avec l'ensemble des nœuds de ses voisins sociaux que nous appelons les **individus** « ***INDIVS*** ». Notons que ***u*** est considéré comme un individu central et donc est exclu de ***INDIVS***.

$\boldsymbol{INFOS_u}$ représente l'ensemble des informations que ***u*** partage ou qui sont indiquées explicitement dans son profil. De la même manière, pour chaque individu $\boldsymbol{indiv} \in \boldsymbol{INDIVS}$, l'ensemble des informations qu'il partage est décrite par « $\boldsymbol{INFOS_{indiv}}$ ». Les informations $\boldsymbol{INFOS_{indiv}}$ sont traitées selon le réseau social étudié : dans le cas de réseaux de publications scientifiques, les informations pourraient être les titres de publications ou bien les mots-clés associés aux publications. Dans le cas de Twitter, les informations pourraient être les textes contenus dans les *tweets* partagés ou bien les *hashtags* associés aux tweets partagés. Nous détaillerons nos choix pour chaque réseau social étudié dans la partie expérimentations (Chapitre 5).

Les individus ***INDIVS*** sont également reliés entre eux par des relations sociales. Notons que nous distinguons dans nos travaux les **relations** et les **interactions** dans le réseau.

Les **relations** sont utilisées pour extraire et représenter les liens entre l'utilisateur central et les individus dans son réseau égocentrique (***u et indiv$_i$***) ainsi que ceux entre les individus (***indiv$_i$* et *indiv$_j$***). Les liens peuvent être orientés ou non-orientés selon le type de réseau social.

Les **interactions** sont utilisées pour désigner les actions entre deux nœuds : (***u et indiv$_i$***) ou (***indiv$_i$* et *indiv$_j$***).

Les relations et les interactions dans les réseaux seront définies selon le type de réseau social étudié : dans le cas de réseaux de publications scientifiques, les liens entre les individus pourraient être définis par la publication d'au moins un article ensemble. Les interactions pourraient être représentées par la co-publication des articles. Dans le cas de Twitter, les liens entre deux nœuds seront définis par le fait de suivre (*follows*) quelqu'un d'autre et les interactions peuvent être définies par le partage de *tweet*(*s*) de quelqu'un d'autre, de réponse au(x) *tweet(s)* de quelqu'un d'autre ou par le fait de mentionner quelqu'un dans son *tweet*. Nous les détaillerons nos choix pour chaque réseau social étudié dans la partie expérimentations (section 5).

Pour chaque $\boldsymbol{info_{indiv}} \in \boldsymbol{INFOS_{indiv}}$, on peut extraire l'ensemble de mots clés ou éléments $\boldsymbol{E}$. Dans les calculs qui suivent, nous utilisons le terme « *élément* » : $e \in E$, pour désigner chaque mot clé[62].

La Figure 4.2 est l'illustration de la représentation du profil social et du profil utilisateur à partir des informations du réseau social de l'utilisateur en respectant les termes et notations proposés.

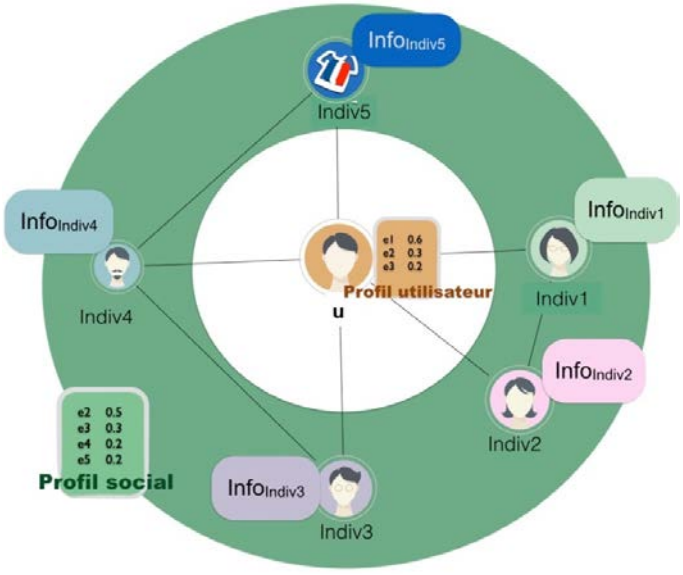


Figure 4.2 Profil social vs Profil utilisateur

4.2.3.2. Définition des formules utilisées

La fonction $\boldsymbol{combinaison}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{p})$ permet de calculer une combinaison des valeurs *A* et *B* avec la combinaison linéaire de la formule (4.1) :

[62] La notation de « élément » à la place de « terme » est utilisée pour éviter l'ambiguïté d'abréviation : *t* de terme avec *t* de temps

$$p * A + (1 - p) * B$$

(4.1)

Où p ∈ [0,1] est la proportion de *A* par rapport à *B*. Cette formule permet de faire varier l'importance entre *A* et *B* pour trouver une combinaison optimale de ces deux éléments dans un calcul.

4.2.4. Définition du processus général de construction du profil social

A partir des travaux sur le profilage social décrits dans la partie 3.3, nous considérons les principales étapes du processus de construction du profil social suivantes :

- **Etape 1 : Sélection des informations partagées par les membres du réseau égocentrique de l'utilisateur**

  Pour chaque individu *indiv* ∈ *INDIVS*, l'ensemble de ses informations partagées $INFOS_{indiv}$ est extrait.

- **Etape 2 : Extraction et pondération des éléments**

  Pour chaque $info_{indiv,}$ l'ensemble des éléments E au niveau des individus (approche basée sur les individus) ou niveau des communautés (approche basée sur les communautés) est extrait. Puis, les mots clés extraits en utilisant une technique choisie sont pondérés (moyenne, somme, *tf*, *tf-idf*, etc.).

- **Etape 3 : Agrégation de mots-clés et dérivation des mots-clés dans le profil social**

  Il s'agit ici, de combiner des mots-clés obtenus depuis plusieurs individus (inter-individus) /communautés (inter-communautés) en sélectionnant/filtrant et en les pondérant en fonction de leur poids calculé depuis l'étape précédente. Plusieurs techniques peuvent être utilisées pour combiner les mots-clés au niveau inter-individus/inter-communautés (*moyenne*, *combinaison linéaire*, …). Enfin le profil social est dérivé avec ces mots-clés pondérés.

La Figure 4.3 illustre un processus général de construction d'un profil social en appliquant la fonction moyenne comme méthode d'agrégation de poids.

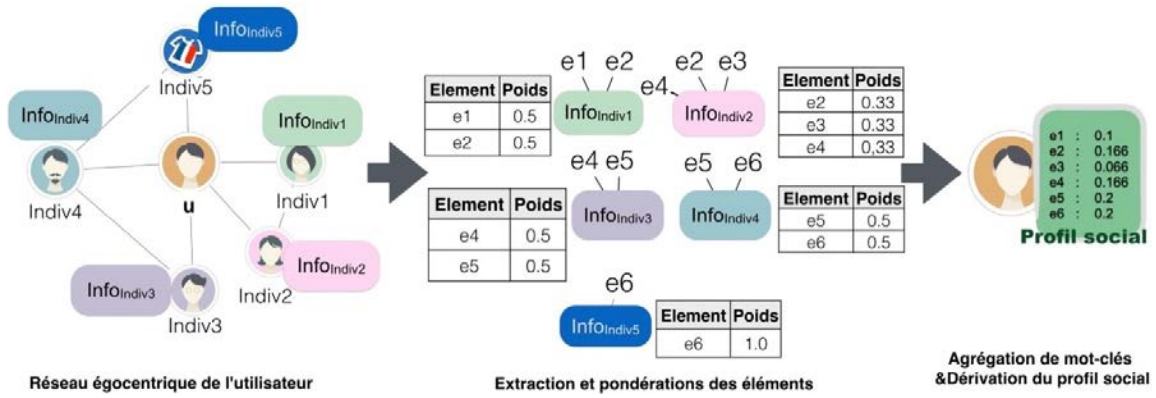


Figure 4.3 Processus général de construction du profil social avec la mesure *tf* et l'agrégation par la fonction moyenne

La section suivante détaille les éléments de notre contribution.

4.3. Construction du profil social en prenant en compte l'évolution du réseau social

Dans cette section, nous présentons les éléments de notre contribution sur la prise en compte de l'évolution du réseau social lors du processus de construction du profil social de l'utilisateur. Pour illustrer notre motivation, l'idée sous-jacente de la méthode ainsi que les calculs utilisés et la façon d'appliquer la méthode proposée au processus existant de construction du profil social, nous allons, tout au long de cette partie, reprendre un cas exemple de « Bob » et de ses relations sociales. Nous donnerons des exemples de calcul ainsi que les détails du processus de construction du profil social de Bob. Pour mettre en évidence le problème de la non prise en compte de l'évolution de son réseau social, nous commencerons tout d'abord par présenter l'étude de cas de Bob pour la construction de son profil social sans prendre en compte cette évolution. Ensuite, nous présenterons la méthode proposée qui tente de pallier ce problème. Enfin, nous présenterons l'application de la méthode proposée sur deux processus existants de construction du profil social : processus basé sur les individus et processus basé sur les communautés.

4.3.1. Etude de cas : profil social de « Bob »

Nous sommes en 2016, Bob est un ingénieur en sécurité informatique dans l'entreprise E-Corp. Il est passionné par la programmation et les nouvelles technologies. Par le passé, il jouait souvent aux jeux vidéo mais actuellement, avec l'âge, il s'éloigne beaucoup de cette activité. Il aime bien voyager, aller au restaurant et regarder des séries. Il supportait l'équipe de foot de France pendant la coupe du monde 2010 mais à cause de leur mauvaise performance, il ne s'intéresse plus au foot après cette compétition. A partir de ces informations, nous en déduisons les intérêts réels de Bob suivants : Geek (nouvelles technologies), Voyages, Restaurant, Séries.

Dans son réseau social Twitter, il est en contact avec :

Carol, qui était une camarade de Bob. Elle le connait depuis 10 ans mais ces derniers temps, ils n'interagissent pas souvent. Carol aime bien jouer aux jeux vidéo et regarder des dessins animés. Voici les informations partagées dans le réseau de Bob par Carol :

- $Info1_{carol}$: What time is it ?! It's Minecraft Time avec Bob #jeuxvideo (2005) *(Bob a aimé ce contenu ♥)*
- $Info2_{carol}$: Totoro j'adore, #dessinanimé (2008)
- $Info3_{carol}$: OdinSphere c'est le top! #jeuxvideo (2010)
- $Info4_{carol}$: Voyage de Shiro, #dessinanimé (2013)
- $Info5_{carol}$: Enfin terminé Ninokuni à 100%, tout seule… #jeuxvideo (2015)

Alice, une autre ancienne camarade de classe qu'il voit encore régulièrement, elle partage des intérêts avec Bob sur les voyages et les restaurants. Par le passé, elle aimait regarder les matchs de tennis. Elle est également amie avec Carol mais elles ne se partagent pas d'intérêts communs. Voici les informations partagées dans le réseau de Bob par Alice :

- $Info1_{alice}$: Allez Nadal !!!! Tennis (2005)
- $Info2_{alice}$: En train de regarder le bon plan pour cet été #voyage (2010)
- $Info3_{alice}$: Organiser un voyage idéal avec Booking.com !!! #voyage (2013) *(Bob a aimé ce contenu ♥)*
- $Info4_{alice}$: Enfin trouvé le meilleur sushi du monde à Tokyo #voyage #restaurant (2016) *(Bob a aimé ce contenu ♥)*

Dave, collègue de travail de Bob est ingénieur de sécurité, il est passionné par l'informatique et la programmation. Dans son temps libre, Dave aime bien regarder des séries. Voici les informations partagées dans le réseau de Bob par Dave:

- $Info1_{dave}$: Comme tous les soirs, je recompile mon kernel #geek (2010) *(Bob a aimé ce contenu ♥)*
- $Info2_{dave}$: Game Of Throne : Team Arya!!!! #series (2012)
- $Info3_{dave}$: Trop de bug tue le bug #geek (2013) *(Bob & Frank ont commenté ce contenu)*
- $Info4_{dave}$: Tranquille devant Game Of Thrones #series (2015) *(Bob a aimé ce contenu ♥)*
- $Info5_{dave}$: XFCE > Gnome2 > KDE > * #geek (2016) *(Bob a commenté ce contenu)*

Frank, un autre collègue de travail de Bob est très passionné par la programmation. Voici les informations partagées dans le réseau de Bob par Frank :

- $Info1_{frank}$: BSD>Linux @Dave #geek (2005) *(Bob a commenté ce contenu)*
- $Info2_{frank}$: Le C c'est la vie !!! #geek (2010)
- $Info3_{frank}$: Le C c'est la vie !!! #geek (2013) *(Bob a commenté ce contenu)*
- $Info4_{frank}$: MacOs est toujours autant buggé @Dave #geek (2015) *(Bob & Dave ont commenté ce contenu)*

Un compte de **l'équipe de France de foot (FFF)** qu'il a initié pendant la coupe du monde 2010 pour suivre les informations sur les joueurs et sur l'équipe pendant cet évènement.

- $Info1_{FFF}$: Coupe du monde 2010, allez les bleus ! #CoupeDuMonde (2010) *(Bob & Frank a aimé ce contenu ♥)*
- $Info2_{FFF}$: Euro2012, allez les bleus ! #Euro (2012)
- $Info4_{FFF}$: Coupe du monde 2014, on va faire mieux que 2010, allez les bleus ! #CoupeDuMonde (2014)
- $Info5_{FFF}$: Euro 2016 en France, allez les bleus #Euro (2016)

Le réseau égocentrique de Bob ainsi que le détail sur les relations et interactions sont illustrés dans la Figure 4.4.

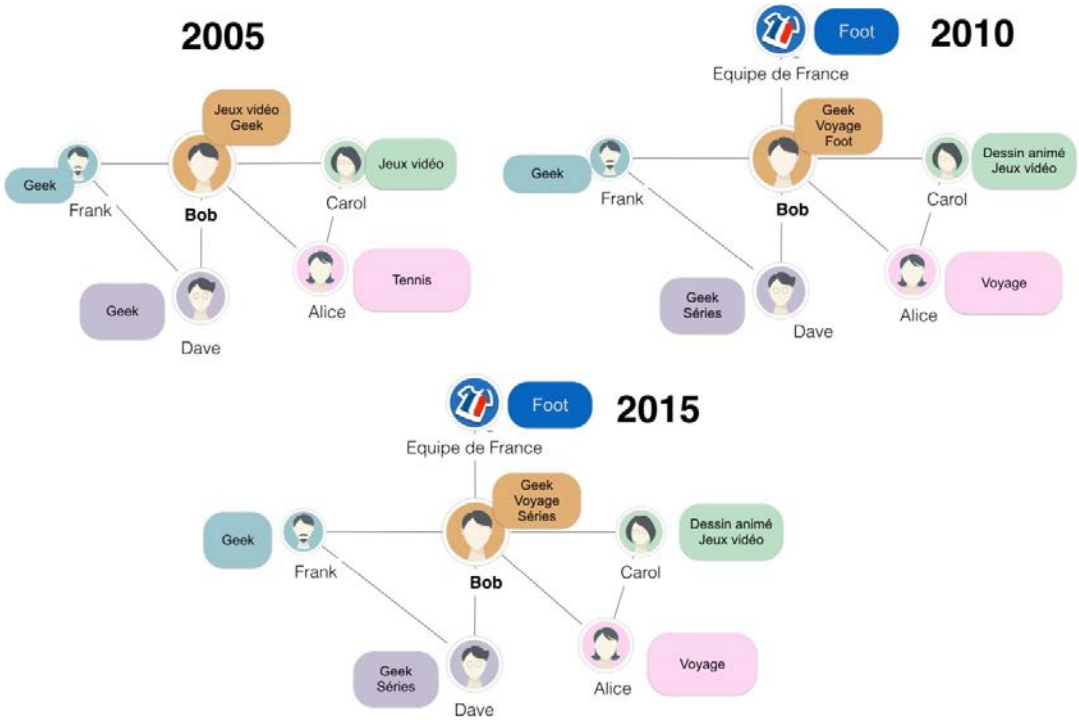


Figure 4.4 Le réseau égocentrique de Bob

Calcul du profil social de Bob sans prendre en compte l'évolution dans son réseau social

Pour calculer, de façon simple, le profil social de Bob sans prendre en compte l'évolution dans son réseau social à l'instant *t*, on prend les informations que les voisins sociaux de Bob partagent comme présenté dans le Tableau 4.1. Notons que pour calculer le poids de chaque intérêt extrait, nous appliquons la fonction moyenne comme fonction d'agrégation des éléments (mots-clés).

indiv	**Extraction et pondération des éléments**		**Combinaison des éléments et dérivation du profil (moyenne)**
	Eléments extraits	**Agrégation des éléments (*tf*)**	
Alice	Tennis Voyage Voyage Voyage Restaurant	Voyage 3/5 = 0.6 Tennis 1/5 = 0.2 Restaurant 1/5 = 0.2	Geek = (0.6+1)/5 = 0.32 Jeux vidéo = (0.6)/5 = 0.12 Voyage = 0.6/5 = 0.12 Euro = (0.5) /5 = 0. 1 Coupe du monde = (0.5) /5 = 0. 1 Dessin animé = 0.4/5 = 0.08 Séries = 0.4/5 = 0.08 Tennis = 0.2/5 = 0.04 Restaurant = 0.2/5 = 0.04
Carol	Jeux vidéo Jeux vidéo Dessin animé Jeux vidéo Dessin animé	Jeux vidéo 3/5 = 0.6 Dessin animé 2/5 = 0.4	
Dave	Geek Séries Geek Séries Geek	Geek 3/5 = 0.6 Séries 2/5 = 0.4	
Frank	Geek Geek Geek Geek	Geek 4/4 = 1	
FFF	Coupe du monde Euro Coupe du monde Euro	Coupe du monde 2/4 = 0.5 Euro 2/4 = 0.5	

Tableau 4.1 Extraction des intérêts sociaux de l'utilisateur Bob sans prise en compte du temps

Avec ce calcul, qui ne prend pas en compte l'évolution du réseau social de Bob, ses intérêts sociaux peuvent ne pas être pertinents pour représenter son profil aujourd'hui :

- L'intérêt **Geek** est en première position car il apparait souvent et régulièrement dans le réseau par le partage des Dave et Frank. Ceci parait pertinent.
- L'intérêt **Jeux vidéo** est à la deuxième position comme l'intérêt **Voyage.** Cela est lié au fait que Carol continue de partager ce genre de contenus alors que Bob ne s'intéresse plus aux jeux vidéo. La position de l'intérêt **Jeux vidéo** parait donc moins pertinente aujourd'hui pour **Bob**.
- L'intérêt **Foot** (Coupe du monde, Euro) est en quatrième position alors qu'il devrait être bien en dessous dans le profil de Bob. Cela est lié au fait que Bob est toujours en relation avec l'équipe de France alors qu'il ne s'intéresse plus au football dans la vie réelle.
- On remarque que l'intérêt **Dessin animé** qui paraît moins pertinent pour Bob est mieux positionné que l'intérêt **Restaurant** alors que ce dernier est un nouvel intérêt de Bob qui vient d'apparaitre récemment dans son réseau social (moins de fréquence de partage sur cet intérêt).

L'exemple ci-dessus montre qu'il est important de prendre en compte l'évolution dans le réseau social de l'utilisateur lors du processus de construction de son profil social. Nous pouvons

remarquer dans un premier temps qu'il est important de prendre en compte la force des liens entre un utilisateur *u* et les individus *INDIV* de son réseau égocentrique. Le but est de sélectionner les informations les plus pertinentes depuis les individus qui ont une influence importante pour l'utilisateur. Le calcul de cette force des liens devrait également prendre en compte un critère temporel entre *u* et chaque *indiv*, en se basant par exemple sur la date de la dernière interaction et/ou le nombre d'interactions pendant la période étudiée. Dans le cas de Bob, cela peut permettre de diminuer considérablement, par exemple, la force du lien entre Bob et l'équipe de France ou la force du lien entre Bob et Carol.

Dans un second temps, il est important de prendre en compte la fraicheur des informations partagées dans le réseau social de l'utilisateur. Les informations qui datent d'un certain temps devraient devenir moins importantes que les informations récentes. Le fait de prendre en compte cette remarque permet d'éviter des informations obsolètes comme l'intérêt **Tennis** qui est issu du profil d'Alice.

Nous avons montré dans cette étude de cas l'importance de la prise en compte de l'évolution du réseau social d'un utilisateur donné dans la construction du profil social de cet utilisateur. Nous allons présenter une synthèse des méthodes et/ou techniques existantes pour prendre en compte ce type d'évolution dans nos travaux.

4.3.2. Synthèse des méthodes/techniques existantes pour la prise en compte de l'évolution du réseau social dans la construction du profil social

Dans les travaux sur l'étude du profil utilisateur classique, il existe des techniques qui permettent de prendre en compte l'évolution des intérêts, comme présenté dans le chapitre 2.3 : soit par l'approche qui prend en compte cette évolution pendant l'étape de construction du profil utilisateur, soit par l'approche de mise à jour du profil utilisateur. Ces approches pourraient être adaptées dans le cas de l'étude du profil social. La différence principale réside dans le fait que dans le cas du profil utilisateur classique, l'évolution des intérêts est reflétée par les changements de comportements de l'utilisateur lui-même alors que dans le cas du profil social, l'évolution des intérêts est reflétée par l'évolution dans son réseau social. Il faut donc prendre en compte le caractère évolutif du réseau social exploité lors du choix de la technique de prise en compte de l'évolution des intérêts.

En se basant sur les techniques d'analyse des réseaux sociaux dynamiques, comme vu précédemment dans la section 3.2.2.2, la prise en compte de l'évolution d'un réseau social peut se faire au travers de deux aspects (Aggarwal et Subbian, 2014).

- Soit par l'analyse de l'évolution du réseau à un instant *t* donné (*Analytical Evolution Analysis*). Avec cette méthode, on essaie de comprendre ce qui s'est passé entre un instant *t-x* et un instant *t*, afin de comprendre le caractère évolutif des éléments du réseau. Cela permet de modéliser l'évolution des informations dans le réseau entre les instants *t-x* et *t*.
- Soit par la méthode de maintenance du modèle au cours du temps (*Maintenance Method*). Cette méthode permet de calculer la pertinence des informations à chaque fois que des changements dans le réseau se produisent. Cela permet de maintenir la fraicheur, et donc la pertinence, des intérêts à chaque fois que le réseau social évolue.

Dans ce travail, nous avons décidé de nous focaliser sur la technique de prise en compte de l'évolution du réseau social durant l'étape de construction du profil plutôt qu'à la technique de mise à jour de ce profil. L'objectif est de construire un profil qui sera pertinent dès sa première construction. Pour ce faire nous nous appuierons sur la méthode *Analytical Evolution Analysis*

présentée précédemment. L'approche de la mise à jour du profil social, correspondant à la méthode de maintenance de modèle *Maintenance Method*, peut être appliquée une fois que le profil sera construit, et ce, pour permettre de maintenir la pertinence du profil malgré l'évolution du réseau. Nous n'aborderons pas ce problème dans nos travaux.

Par rapport aux techniques existantes pour la prise en compte de l'évolution des intérêts pendant l'étape de construction du profil utilisateur classique (section 2.3.1), nous pouvons mentionner deux approches de prise en compte du temps : l'approche par sélection d'instance (*Instance selection*) et l'approche pondérée (*Instance weighting*).

En étudiant les informations et les relations dans un réseau social numérique, certains facteurs peuvent avoir un impact sur le comportement des utilisateurs. Les utilisateurs actifs (*Giant Component*) ont tendance à partager beaucoup d'informations et à créer des liens pendant une période donnée. Il existe également des utilisateurs passifs (*Singleton, Middle Region*) dont les réseaux sociaux évoluent lentement. Le réseau social de l'utilisateur peut également évoluer de manière progressive ou bien soudainement lors d'événements donnant lieu à une accélération des partages.

Dans ce type de réseau, la technique de sélection d'instance (*Instance selection*) présente un risque d'oubli d'information, et ne permet pas la valorisation des informations précieuses. Cette technique peut poser problème dans le cas d'un réseau qui évolue progressivement. De plus, le choix de la période Δt peut également poser des problèmes (on ne sait pas à quel moment il faut couper et oublier des informations).

Nous nous intéressons donc à la technique de pondération temporelle (*Instance weighting*), qui permet de sélectionner et traiter toutes les informations disponibles en prenant en compte certains critères temporels. La méthode que nous proposons est présentée ci-après.

4.3.3. Méthode temporelle proposée

Pour prendre en compte l'évolution du réseau social, nous considérons deux principaux types d'évolution dans le réseau social de l'utilisateur : l'évolution liée à la dynamique des relations entre l'utilisateur et les individus dans son réseau social et l'évolution liée à la dynamique des informations partagées par les individus dans son réseau social (comme montré dans l'étude de cas de Bob).

Il s'avère donc nécessaire d'étudier comment intégrer ces deux types d'évolution du réseau social dans le processus de construction du profil social de l'utilisateur.

Pour intégrer l'évolution liée à la dynamique des relations, il est nécessaire de qualifier les relations les plus pertinentes et à jour pour l'utilisateur. Pour cela, il faut sélectionner les sources d'informations (individus) les plus significatives.

Pour intégrer l'évolution liée à la dynamique des informations, il est nécessaire de ne sélectionner que les informations les plus significatives pour l'utilisateur à partir des informations partagées par les individus sélectionnés.

Nous nous sommes intéressés de manière particulière aux critères temporels selon deux axes liés aux informations temporelles extraites du réseau social :

- L'estampille (*timestamp*) des interactions entre l'utilisateur et les individus de son réseau social pour la prise en compte de l'évolution de leurs relations.
- L'estampille (*timestamp*) de publication des informations pour la prise en compte de l'évolution des informations

Nous proposons d'utiliser la technique de pondération temporelle (*Instance weighting*) par rapport à ces deux axes.

La méthode proposée s'applique principalement à l'étape 2 du processus général de construction du profil social. En effet, à la différence de la technique existante non temporelle, nous appliquons la technique de pondération temporelle (*Instance weighting*) pour calculer le poids temporel des éléments extraits avant de dériver le profil social de l'utilisateur.

La Figure 4.5 représente le modèle d'intégration du poids temporel des éléments extraits dans le processus de construction du profil social. Notons que le profil social présenté dans ce modèle est une partie (profil social) du profil complet montré dans la Figure 4.1.

La sous-section suivante détaille les calculs proposés pour évaluer les poids temporels.

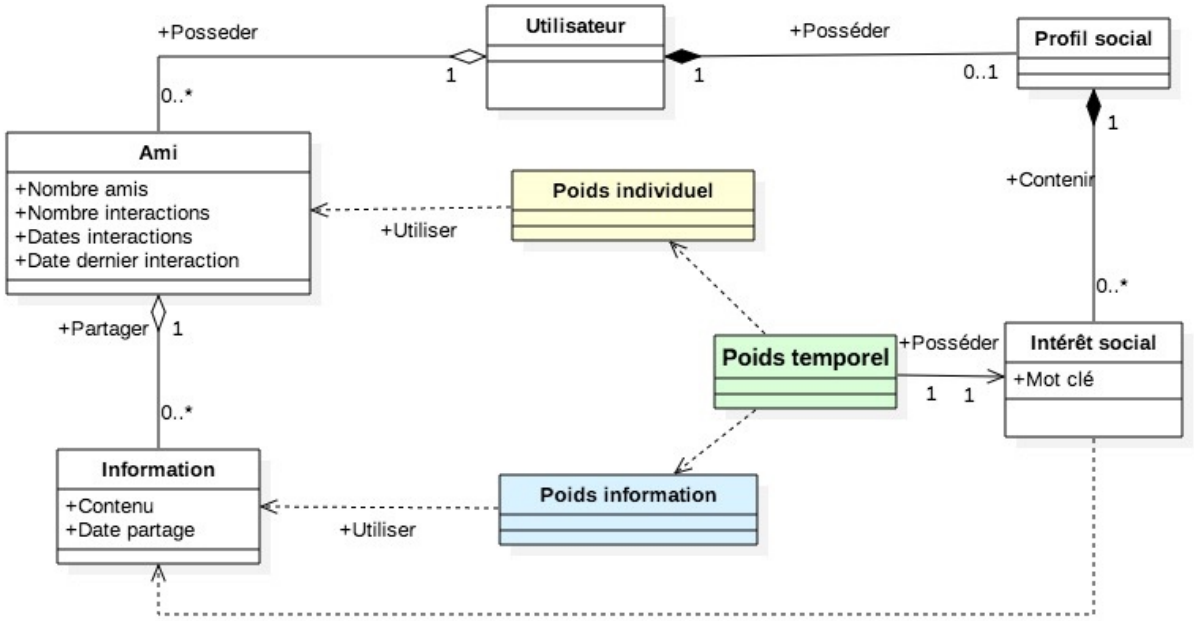


Figure 4.5 Modèle d'intégration du poids temporel dans la construction du profil social

4.3.4. Calcul du poids temporel d'un élément

Le poids temporel de chaque élément est calculé d'une part, à partir de la pertinence temporelle des informations utilisées pour extraire cet élément et d'autre part, à partir de la pertinence temporelle de l'individu (poids temporel de l'individu) qui partage cette information (poids temporel de l'information). Le poids temporel d'une information (resp. d'un individu) représente la pertinence de cette information (resp. de cet individu) en fonction du temps. Dans un réseau social donné, les informations partagées peuvent avoir plus ou moins d'importance que les relations entre les individus. Nous proposons donc une technique de combinaison de ces deux poids pour refléter cette importance.

Nous proposons dans ce qui suit un algorithme générique qui permet de calculer ce poids temporel.

4.3.4.1. Algorithme générique

Notons e_{indiv}, un élément extrait depuis les informations $INFOS_{indiv}$ partagées par l'individu *indiv*, le poids temporel de e_{indiv} est calculé d'une part, à partir de :

1) la pertinence temporelle de *indiv*, que nous appelons **poids temporel de l'individu** $\boldsymbol{P_{temp}^{indiv}}$.

2) la pertinence temporelle de chaque information $info_{indiv} \in INFOS_{indiv}$ qui contient l'élément e_{indiv}, que nous appelons **poids temporel de l'information** $\mathbf{P_{temp}^{info}}$.

Ces deux poids seront combinés pour obtenir le poids temporel final de chaque e_{indiv}.

Le calcul générique du poids temporel des éléments extraits à partir de chaque individu *indiv* dans le réseau égocentrique de l'utilisateur est présenté dans l'algorithme suivant.

Algorithme *calculer_*P_{temp}
Entrée $INFOS_{indiv}$: **Sortie** : Poids temporel des éléments extraits $P_{temp}(E_{indiv})$
Début
1: // Etape 1 : calcul du poids temporel de l'individu $Indiv$ (section 4.3.4.3)
2: $P_{temp}^{indiv}(indiv) = Calculer\_P_{ind}\ (indiv)$;
3: **Pour chaque** $info_{indiv} \in INFOS_{indiv}$ **faire**
4: // Etape 2 : calcul du poids temporel de chaque information $info_{indiv}$ (section 4.3.4.2)
5: $P_{temp}^{info}(info_{indiv}) = Calculer\_P_{inf}(info_{indiv})$
6: **Pour chaque** $e_{indiv} \in info_{indiv}$ **faire**
7: $P_{temp}^{info}(e_{indiv}) = Calculer\_P_{inf}(info_{indiv})$
8: // Etape 3 : combinaison de poids temporel de l'individu et poids temporel de l'information (section 4.3.4.4)
9: $P_{temp}(e_{indiv}) = Combiner(P_{temp}^{indiv}(indiv), P_{temp}^{info}(e_{indiv}))$
10: **Fin pour**
11: **Fin pour**
12:
13: Retourner $\mathrm{P_{temp}}(E_{indiv})$
Fin

Algorithme 1: *calculer_*P_{temp} : calcul générique du poids temporel des éléments

Notons que les techniques utilisées pour calculer le poids temporel de l'information (*Calculer_*P_{inf}) et celles utilisées pour calculer le poids temporel de l'individu (*Calculer_*P_{ind}) ne sont pas fixées a priori. Dans ce travail, nous allons appliquer différentes techniques de calcul de ces deux poids pour étudier l'efficacité de chaque technique et choisir la plus adaptée pour chaque type de réseau social.

Nous présentons en détail, dans la sous-section qui suit, les techniques de calcul de ces deux poids ainsi que la technique de combinaison de ces deux poids.

4.3.4.2. Calcul du poids temporel d'un individu

Pour calculer le poids temporel d'un individu, nous proposons d'appliquer les mesures de calcul de force des liens entre cet individu et l'utilisateur. Comme vu dans le chapitre 3, il existe plusieurs mesures pour calculer la force des liens entre deux individus dans le réseau social. Nous présentons les calculs de poids d'un individu selon différentes fonctions de calcul de force des liens utilisées.

a. Calcul du poids temporel d'un individu avec la mesure de prédiction de liens Adamic/Adar temporelle (Calculer_P_{Ind_AATemp})

Nous proposons dans notre travail d'appliquer les mesures utilisées dans le domaine de la prédiction de liens pour calculer un poids de similarité entre deux nœuds déjà connectés. L'idée est de pondérer le lien entre deux individus par une évaluation de la persistance potentielle du lien entre ces nœuds dans le futur.

Nous nous intéressons à la mesure de prédiction de liens *Adamic/Adar*, introduite dans (Liben-Nowell et Kleinberg, 2003) , qui s'est avérée performante malgré sa simplicité. Cette méthode se base sur le nombre de voisins communs des nœuds *x* et *y*, en considérant que plus les voisins communs possèdent peu de voisins, plus il est probable que x et y se connectent (Adamic et Adar, 2003). Si on note *z* un nœud qui est le voisin commun de *x* et *y*, et Γ(*z*) l'ensemble de ses voisins, la formule pour calculer le score(*x*,*y*) liée à cette méthode est :

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

(4.2)

Pour prendre en compte le critère temporel, nous nous intéressons à la prédiction de liens temporelle qui s'appuie sur les informations structurelles et temporelles lors du calcul du score de similarité entre deux nœuds. (Tylenda, Angelova et Bedathur, 2009) proposent une technique de prédiction temporelle qui applique des facteurs temporels dans le calcul des poids de pertinence des individus.

Nous considérons que les individus qui ont les relations les plus récentes avec l'utilisateur central ont une probabilité plus grande de partager les mêmes intérêts avec cet utilisateur (individus plus significatifs). Nous donnons donc un poids plus important aux informations partagées par ces individus par rapport à celles partagées par des individus ayant des liens moins récents. Il s'agit de calculer la pertinence d'un individu *indiv* par rapport à l'utilisateur central *u* en prenant en compte les informations temporelles de ses liens avec l'utilisateur central (par exemple, la date de création de liens avec l'utilisateur central, la durée de la relation, …).

En se basant sur ce travail, la formule *Adamic/Adar* qui prend en compte les informations temporelles devient :

$$AdamicAdarTemp(u,indiv) = \sum_{z \epsilon \{\Gamma(u) \cap \Gamma(indiv)\}} \frac{w(u,z) \cdot w(z,indiv)}{log|\Gamma(z)|}$$

(4.3)

Rappelons que, quel que soit *x*, un nœud dans le réseau, *Γ(x)* représente l'ensemble des voisins de *x*. Pour deux nœuds *x* et *y*, la fonction w(*x*,*y*) permet de calculer le poids temporel de leur relation.

Pour calculer la fonction *w(x,y)*, nous pouvons appliquer différentes fonctions temporelles pour calculer le poids temporel des relations entre *x* et *y* : fonction linéaire inverse ($f_{lin}(t)$), fonction temporelle exponentielle ($f_{exp}(t)$) ou fonction temporelle polynomiale ($f_{poly}(t)$). Dans ce contexte, *t* représente la fraîcheur de la dernière interaction entre *x* et *y* (*t*=0 étant la plus récente).

$$w(x,y) = f_{func}(timestamp_{actuelle} - timestamp_{dernière\_interaction(x,y)})$$

(4.4)

Où f_{func} représente la fonction temporelle *func* utilisée : f_{lin} , f_{exp} , f_{poly}.

$w(u,z) \cdot w(z,indiv)$ est une fonction de combinaison entre le poids $w(u,z)$ et le poids $w(z,indiv)$. L'opérateur de combinaison peut être paramétré. Nous utilisons ici la moyenne arithmétique des deux poids (formule (4.5)).

$$w(u,z) \cdot w(z,indiv) = \frac{w(u,z) + w(z,indiv)}{2}$$

(4.5)

En se basant sur cette mesure, le poids temporel de chaque *indiv* est calculé par la formule (4.6) ci-dessous. Notons que ce poids est identique pour tous les éléments $e_{info_{indiv}}$ appartenant à chaque $info_{indiv}$.

$$P_{temp}^{indiv}(indiv) = AdamicAdarTemp(u,indiv)$$

(4.6)

Voici un exemple de l'application du calcul du poids temporel des individus sur l'étude de cas de Bob avec la technique de calcul *Calculer_P*$_{Ind\_AATemp}$ en appliquant la fonction temporelle exponentielle f_{exp}.

indiv	**Calcul du poids de l'individu**						
	Voisin commun *(z)*		**Poids temporel de l'individu en fixant $\lambda = 0.1$**				
			w(Bob.z)		***w(z.indiv)***		$P_{temp}^{indiv}(indiv)$
	nom	**degré**	***t***	$f_{exp}(t)$	***t***	$f_{exp}(t)$	
Carol	Alice	5	0	1.0	-	-	$\frac{1.0 + 0.0}{\frac{2}{log(5)}} = 0.71534$
Alice	Carol	3	2005	0.33287	-	-	$\frac{0.33287 + 0.0}{\frac{2}{log(3)}} = 0.34883$
Dave	Frank	4	2015	0.90484	2015	0.90484	$\frac{0.90484 + 0.90484}{\frac{2}{log(4)}} = 1.50290$
Frank	Dave	3	2016	1.0	2015	0.90484	$\frac{1.0 + 0.90484}{\frac{2}{log(3)}} = 1.99618$
FFF	-	-	-	-	-	-	0.0

Exemple 2 : calcul du poids temporel des individus dans l'étude de cas de Bob avec la technique de calcul du poids temporel *Calculer_P*$_{Ind\_AATemp}$ en appliquant la fonction temporelle exponentielle $f_{exp.}$

Nous pouvons remarquer qu'avec ce calcul, les individus qui ont le plus de voisins en commun avec l'utilisateur auront plus d'importance. De plus, avec la fonction temporelle appliquée, pour qu'un individu ait un poids important, il faut que ses voisins communs avec l'utilisateur restent actifs. C'est-à-dire que leurs interactions (entre les voisins communs et l'utilisateur ainssi que celles entre les voisins communs et l'individu) soient récentes (cas de Frank et Dave). Les

individus qui n'ont pas de voisin en commun avec l'utilisateur auront un poids égal à 0 comme le cas de FFF.

b. Calcul du poids temporel d'un individu avec la prédiction de liens en intégrant la date de la dernière interaction (Calculer_$P_{Ind_AAStrTemp}$)

Avec la mesure précédente, la force des liens entre l'utilisateur *u* et un individu *indiv* ne dépend pas des liens directs entre *u* et *indiv* mais des liens de ces derniers avec leurs voisins communs. La force des liens entre *u* et *indiv* ne sera pas mise en valeur s'ils ont peu (ou pas) de voisins en commun, ce qui peut être un critère non discriminant pour certains individus. La nouvelle mesure proposée est presque identique à la précédente mais dépend des liens directs entre *u* et *indiv.*

Nous considérons que les individus qui ont les relations les plus récentes avec l'utilisateur central ont une probabilité plus grande de partager les mêmes intérêts avec cet utilisateur (individus plus significatifs) (cf. mesure de calcul de force de liens présentée dans la section 3.2.1.4). Nous donnons donc un poids plus important aux informations partagées par ces individus par rapport celles partagées par des individus ayant des liens moins récents. Il s'agit de calculer la pertinence d'un individu *indiv* en prenant en compte les informations temporelles de ses liens avec l'utilisateur central *u*. La formule de cette mesure devient :

$$AdamicAdarStrTemp(u, indiv) = w(u, indiv) \cdot \sum_{z \epsilon \{\Gamma(u) \cap \Gamma(indiv)\}} \frac{1}{log|\Gamma(z)|} \quad (4.7)$$

Rappelons que, quel que soit *x*, un nœud dans le réseau, *Γ(x)* représente l'ensemble des voisins de *x*. w(*u*,*indiv*) permet de calculer le poids temporel de relation entre *u* et *indiv*, calculé par la formule (4.4). Au lieu d'appliquer cette fonction pour calculer le poids temporel de la dernière interaction entre l'utilisateur et les voisins commun ou l'individu avec les voisins communs, nous appliquons directement ce poids à la dernière interaction entre l'utilisateur et l'individu en question.

En se basant sur cette mesure, le poids temporel de chaque *indiv* est calculé par la formule (4.8) ci-dessous. Notons que ce poids est identique pour tous les éléments $e_{info_{indiv}}$ appartenant à chaque $info_{indiv}$.

$$P_{temp}^{indiv}(indiv) = AdamicAdarStrTemp(u, indiv) \quad (4.8)$$

Voici un exemple de l'application du calcul du poids temporel des individus sur l'étude de cas de Bob avec la technique de calcul du poids temporel *Calculer_$P_{Ind_AAStrTemp}$* en appliquant la fonction temporelle exponentielle f_{exp}.

indiv	**Calcul du poids de l'individu**							
	Voisin commun (z)				**Poids temporel de l'individu en fixant $\lambda = 0.1$**			
					w(u,indiv)			$P_{temp}^{indiv}(indiv)$
	Nom	**Degré**	**Score**	**Score normalisé**	***t***	$f_{exp}(t)$	**Score normalisé**	
Carol	Alice	5	1.43068	0.68261	2005	0.33287	0.33287	$\frac{0.33287 + 0.68261}{2} = 0.50774$
Alice	Carol	3	2.09590	1.0	2016	1.0	1.0	$\frac{1.0 + 1.0}{2} = 1.0$
Dave	Frank	4	1.66096	0.79248	2016	1.0	1.0	$\frac{1.0 + 0.79248}{2} = 0.89624$
Frank	Dave	3	2.09590	1.0	2015	0.90484	0.90484	$\frac{0.90484 + 1.0}{2} = 0.95242$
FFF	-	-	-	-	2010	0.54881	0.54881	$\frac{0.54881 + 0.0}{2} = 0.27441$

Exemple 3 : calcul du poids temporel des individus dans l'étude de cas de Bob avec la technique de calcul du poids temporel *Calculer_*$P_{Ind\_AAStrTemp}$ en appliquant la fonction temporelle exponentielle $f_{exp.}$

Nous pouvons remarquer qu'avec ce calcul, les individus qui ont le plus de voisins en commun avec l'utilisateur et qui ont interagi plus récemment avec l'utilisateur auront plus d'importance (Alice, Dave, Frank). Le poids de Carol, qui a des voisins en commun avec Bob mais qui n'a pas interagi depuis longtemps avec lui, est diminué par le poids temporel de la date de sa dernière interaction avec Bob. FFF obtient un poids moins important car en plus de ne pas avoir de voisins communs avec Bob, la date de sa dernière interaction avec Bob est plus ancienne par rapport aux autres individus. Ceci montre l'obsolescence des relations entre FFF et Bob.

c. Calcul du poids temporel d'un individu avec la somme temporelle des interactions (Calculer_*P_{Ind_SITmp})*

Nous proposons dans cette section une deuxième technique de calcul du poids temporel d'un individu *indiv* qui se base sur le nombre d'interactions entre l'utilisateur *u* et *indiv*. Plus *u* interagit avec *indiv* plus le poids de *indiv* est élevé.

Pour prendre en compte le critère temporel des interactions, nous proposons de calculer le poids temporel de chaque interaction : plus l'interaction est récente plus elle est considérée importante et significative. Cela permet de donner plus de poids aux individus qui interagissent avec l'utilisateur de manière plus fréquente et plus récente. Le poids des individus qui ont interagi beaucoup avec l'utilisateur dans le passé mais qui n'ont plus d'interactions récentes sera diminué par le fait de l'ancienneté des interactions. Le calcul du poids temporel d'un individu (appelé « somme temporelle ») est présenté dans la formule (4.9)

$$SommeInteractTmp(u, indiv) = \sum_{interact_i \in Interactions} f_{func}(timestamp_{actuelle} - timestamp_{interact_i}) \quad (4.9)$$

Où *Interactions* représente l'ensemble des interactions entre *u* et *indiv*. Comme précédemment, f_{func} représente la fonction temporelle *func* utilisée : f_{lin}, f_{exp}, f_{poly} permettant de calculer le poids temporel de chaque interaction. Ce poids temporel est calculé en se basant sur la fraicheur de l'interaction ($timestamp_{actuelle} - timestamp_{interact_i}$).

Voici un exemple de l'application du calcul du poids temporel des individus sur l'étude de cas de Bob avec la technique de calcul *Calculer_*P_{Ind_SITmp} en appliquant la fonction temporelle exponentielle f_{exp} .

indiv	**Interactions**		**Poids temporel de l'individu en fixant $\lambda = 0.1$**	
	timestamp	***t***	**f_{exp} (t)**	$P_{temp}^{indiv}(indiv)$
Carol	2005	2016 - 2005 = 11	0.33287	0.33287
Alice	2013	2016 - 2013 = 3	0.54881	0.54881 + 1 = 1.54881
	2015	2016 - 2016 = 0	1	
Dave	2010	2016 - 2010 = 6	0.54881	0.54881 + 0.74082+ 0.90484 + 1 = 3.19447
	2013	2016 - 2013 = 3	0.74082	
	2015	2016 - 2016 = 1	0.90484	
	2016	2016-2016 = 0	1.0	
Frank	2005	2016 - 2005 = 11	0.33287	0.33287 + 0.74082+ 0.90484 = 1.97853
	2013	2016 - 2013 = 3	0.74082	
	2015	2016 - 2015 = 1	0.90484	
FFF	2010	2016 - 2010 = 6	0.54881	0.54881

Exemple 4 : calcul du poids temporel des individus dans l'étude de cas de Bob avec la technique de calcul *Calculer_*P_{Ind_SITmp} en appliquant la fonction temporelle exponentielle f_{exp}

Nous pouvons remarquer qu'avec ce calcul, les individus qui interagissent le plus souvent avec l'utilisateur auront plus d'importance (Dave, dans le cas présent). Cependant, le poids dépend de la fraicheur de chaque interaction : Carol et FFF ont eu une interaction avec Bob mais FFF a plus d'importance car la date de l'interaction entre FFF et Bob est plus récente. Le poids d'Alice est moins important que celui de Frank et de Dave car l'écart entre la première et la dernière interaction entre Alice et Bob est plus petit que celui entre Bob et Frank ou entre Bob et Dave. Cette technique permet donc d'une part, de diminuer le poids de la relation entre Bob et FFF ou celui entre Bob et Carol (ces relations paraissent actuellement moins significatives pour Bob) et d'autre part, de renforcer le poids des relations persistances et récurrentes comme celles de Bob avec Frank ou Dave.

Après le poids temporel des individus, nous présentons dans ce qui suit le calcul du poids temporel des informations.

4.3.4.3. Calcul du poids temporel des informations contenant un élément

A partir des *INFOS* partagées par les *INDIVS* du réseau égocentrique de l'utilisateur, nous proposons de pondérer leur poids avec une fonction temporelle qui va permettre d'augmenter/diminuer la pertinence des informations selon leur fraîcheur. Notons que les fonctions temporelles sont interchangeables et nous proposons d'appliquer les différentes fonctions temporelles présentées dans la section 2.3.1.2 pour calculer le poids temporel des informations : fonction linéaire inverse ($f_{lin}(t)$), fonction temporelle exponentielle ($f_{exp}(t)$) ou fonction temporelle polynomiale ($f_{poly}(t)$).

Comme ces trois fonctions temporelles prennent en compte différemment la façon de faire varier le poids des informations selon leur fraicheur, nous proposons d'appliquer ces trois fonctions temporelles afin de déterminer la fonction la plus appropriée.

Dans notre contexte, par rapport aux formules de calcul de poids temporel indiqués ci-dessus, nous calculons le poids temporel d'une information selon différentes manières de calculer la fraicheur *t* de l'information :

- ***Calculer_P*$_{Inf\_timestamp}$** : calcul du poids temporel d'une information $info_{indiv}$ avec la valeur de *t* basée sur la durée entre le *timestamp* de la publication de $info_{indiv}$ et le *timestamp* actuel (*timestamp* au moment du calcul), appelé t_{actuel}. Le calcul de t_{actuel} est présenté dans la formule (4.10).

$$t_{actuel}(info_{indiv}) = \left| timestamp_{actuelle} - timestamp_{info_{indiv}} \right| \quad (4.10)$$

  Par exemple, si on est en 2016, la valeur de t_{actuel} des informations partagées en 2014 vaudra 2016 - 2014 = 2, celle des informations partagées en 2015 vaudra 2016 - 2015 = 1 (cf. Figure 4.6)

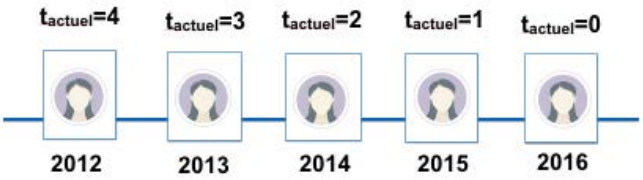


Figure 4.6 Fraicheur de chaque *Info*$_{indiv}$ par rapport à sa date de la publication

  A partir de la fraicheur obtenue, nous pouvons calculer le poids temporel d'une information avec la formule (4.11):

$$P_{temp}^{info\_actuel}(info_{indiv}) = f_{func}(t_{actuel}(info_{indiv}) \quad (4.11)$$

  Avec ce calcul, plus une information est récente plus elle est importante.

- ***Calculer_P*$_{Inf\_lastInteract}$** : calcul de poids temporel d'une information en se basant sur la durée entre le *timestamp* de publication de $info_{indiv}$ et le *timestamp* de la dernière interaction entre l'individu $indiv$ et l'utilisateur *u*, appelée $t_{interact}$. L'idée sous-

jacente d'utiliser $t_{interact}$ est que nous supposons que les informations qui sont partagées dans la même période que la dernière interaction entre *u* et *indiv* ont une probabilité plus importante d'être liées aux informations récentes qu'ils s'échangent (sujets en communs).

Le calcul de $t_{intract}$ est présenté dans la formule (4.12) .

$$t_{interact}(info_{indiv}) = \left|timestamp_{dernierinteract\ (indiv,u)} - timestamp_{\ info_{indiv}}\right| \quad (4.12)$$

Avec ce calcul, le poids de chaque $info_{indiv}$ dépendra de la durée entre sa date de publication et celle de la dernière interaction entre *u* et *indiv*. Par exemple, si la date de la dernière interaction entre *indiv* et *u* remonte à 2014, la valeur de $t_{interaction}$ des informations publiées en 2015 et 2013 vaut 1, la valeur de $t_{interaction}$ des informations publiées en 2016 et 2012 vaut 2 et ainsi de suite (cf. Figure 4.7).

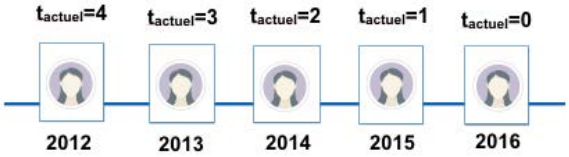


Figure 4.7 La fraicheur de chaque $Info_{indiv}$ partagée par un *indiv* par rapport à la date de la dernière interaction entre *indiv* et u

A partir de la fraicheur obtenue, nous pouvons calculer le poids temporel de l'information avec la formule (4.13) :

$$P_{temp}^{info\_interact}(info_{indiv}) = f_{func}(t_{actuel}(info_{indiv})) \quad (4.13)$$

Où f_{func} représente la fonction temporelle f_{unc} utilisée : f_{lin} , f_{exp} , f_{poly}.

***Calculer_P**$_{Inf\_combin}$* : calcul du poids temporel d'une information en combinant les deux poids $P_{temp}^{info\_actuel}$ et $P_{temp}^{info\_interact}$ (formule (4.11) et (4.13)). Le poids temporel de l'information $info_{indiv}$ est calculé par la moyenne arithmétique de ces deux poids comme présenté dans la formule (4.14).

$$P_{temp}^{info}(info_{indiv}) = \frac{P_{temp}^{info\_actuel}(info_{indiv}) + P_{temp}^{info\_interact}(info_{indiv}))}{2} \quad (4.14)$$

Nous présentons ci-dessus l'exemple de calcul de poids temporel des informations des voisins sociaux de Bob en appliquant la fonction temporelle exponentielle $f_{exp.}$

indiv	**Extraction et pondération du mot clés**							
	info	***e***	***tmp***	**Poids temporel de l'information en fixant $\lambda = 0.1$**				
				$f(t_{actuel})$		$f(t_{interaction})$		$P_{temp}^{info}(info_{indiv})) = \frac{f(t_{actuel}) + f(t_{interact})}{2}$
Carol	$Info1_{carol}$	Jeux vidéo	2005	2016-2005 = 11	0.33287	|2005-2005| = 0	1.0	0.66644
	$Info2_{carol}$	Dessin animé	2008	2016-2018 = 8	0.44933	|2005-2008| = 3	0.74082	0.59507
	$Info3_{carol}$	Jeux vidéo	2010	2016-2010 = 6	0.54881	|2005-2010| = 5	0.60653	0.57767
	$Info4_{carol}$	Dessin animé	2013	2016-2013 = 3	0.81873	|2005-2013| = 8	0.44933	0.59507
	$Info5_{carol}$	Jeux vidéo	2015	2016-2015 = 1	0.90484	|2005-2015| = 10	0.36788	0.63636
Alice	$Info1_{Alice}$	Tennis	2005	2016-2005 = 11	0.33287	|2016-2005| = 11	0.33287	0.33287
	$Info2_{Alice}$	Voyage	2010	2016-2010 = 6	0.54881	|2016-2010| = 6	0.54881	0.54881
	$Info3_{Alice}$	Voyage	2013	2016-2013 = 3	0.74082	|2016-2013| = 3	0.74082	0.74082
	$Info4_{Alice}$	Voyage	2016	2016-2016 = 0	1.0	|2016-2016 = 0	1.0	1.0
Dave	$Info1_{Dave}$	Geek	2010	2016-2010 = 6	0.54881	|2016-2010| = 6	0.54881	0.54881
	$Info2_{Dave}$	Séries	2012	2016-2012 = 4	0.67032	|2016-2012| = 4	0.67032	0.67032
	$Info3_{Dave}$	Geek	2013	2016-2013 = 3	0.74082	|2016-2013| = 3	0.74082	0.74082
	$Info4_{Dave}$	Séries	2015	2016-2015 = 1	0.90484	|2016-2015| = 1	0.90484	0.90484
	$Info5_{Dave}$	Geek	2016	2016-2016 = 0	1.0	|2016-2016| = 0	1.0	1.0
Frank	$Info1_{Frank}$	Geek	2005	2016-2005 = 11	0.33287	|2015-2005| = 10	0.36788	0.35038
	$Info2_{Frank}$	Geek	2010	2016-2010 = 6	0.54881	|2015-2010| = 5	0.60653	0.57767
	$Info3_{Frank}$	Geek	2013	2016-2013 = 3	0.74082	|2015-2013| = 2	0.81873	0.77977
	$Info4_{Frank}$	Geek	2015	2016-2015 = 1	0.90484	|2015-2015| = 0	1.0	0.95242
FFF	$Info1_{FFF}$	Coupe Du Monde	2010	2016-2010 = 6	0.54881	|2010-2010| = 0	1.0	0.77441
	$Info2_{FFF}$	Euro	2012	2016-2012 = 4	0.67032	|2010-2012| = 2	0.81873	0.74453
	$Info3_{FFF}$	Coupe Du Monde	2014	2016-2014 = 2	0.81873	|2010-2014| = 4	0.67032	0.74453
	$Info4_{FFF}$	Euro	2016	2016-2016 = 0	1.0	|2010-2016| = 6	0.54881	0.77441

Exemple 5 : Calcul du poids temporel des informations *INFOS* de tous les voisins sociaux de Bob avec la technique *Calculer_*P_{Inf_combin} en appliquant la fonction exponentielle f_{exp}

Après avoir calculé le poids temporel de chaque information $info_{indiv} \in INFOS_{indiv}$, nous extrayons les éléments *e* appartenant à chaque information. Notons que le poids temporel d'un élément *e* extrait d'une information *info*$_{indiv}$ est égal au poids temporel de l'information dont il est extrait ($P_{temp}^{info}(e) = P_{temp}^{info}(info_{indiv})$).

Comme un *indiv* peut posséder plusieurs informations et que, à partir de chaque information, on peut extraire plusieurs éléments, nous agrégeons d'abord les éléments appartenant aux informations de *indiv*. Comme présenté dans l'algorithme1, l'agrégation se calcule en faisant la somme des poids temporels de l'information de chaque élément, normalisé par la somme totale de tous les éléments trouvés dans l'ensemble des éléments extraits depuis $INFOS_{indiv}$.

$$P_{temp}^{info}(e_{indiv}) = \frac{\sum_{e_{indiv} \in info_{indiv,e}} P_{temp}^{info}(info_{indiv,e})}{\sum_{f \in E_{indiv}} P_{temp}^{info}(f)} \quad (4.15)$$

E_{indiv} représente l'ensemble de tous les éléments extraits à partir de $INFOS_{indiv}$.

Nous présentons ci-dessus l'exemple de calcul de poids temporel des éléments extraits à partir des informations des voisins sociaux de Bob à partir des poids temporels (en se basant sur les poids temporels des informations).

indiv	e_{indiv}	$P_{temp}^{info}(e_{indiv})$
Carol	Jeux vidéo	$\frac{1.88047}{3.07061} = 0.61241$
	Dessin animé	$\frac{1.19015}{3.07061} = 0.38759$
Alice	Tennis	$\frac{0.33287}{3.6225} = 0.09189$
	Voyage	$\frac{2.28963}{3.6225} = 0.63206$
	Restaurant	$\frac{1.0}{3.6225} = 0.27605$
Dave	Geek	$\frac{2.28963}{3.86479} = 0.59243$
	Séries	$\frac{1.57516}{3.86479} = 0.40757$
Frank	Geek	$\frac{2.66024}{2.66024} = 1.0$
FFF	Coupe du monde	$\frac{1.51893}{3.03786} = 0.5$
	Euro	$\frac{1.51893}{3.03786} = 0.5$

Exemple 6 : Calcul du poids temporel des éléments de tous les voisins sociaux de Bob

Nous pouvons remarquer que le poids des informations partagées par les individus dans le réseau égocentrique de Bob dépend de leur fraicheur par rapport à la date actuelle et par rapport à la date de leur dernière interaction avec Bob. Par exemple, les informations sur « Jeux vidéo » que partage Carol ont des poids moins importants que ceux des informations sur « Voyage » partagées par Alice car la date de la dernière interaction entre Bob et Alice est plus récente. En effet, les informations que partage Alice ont un poids plus important parce qu'elles sont récentes mais aussi parce qu'elles sont partagées dans la même période que la dernière interaction entre

Alice et Bob. Elles sont donc considérées plus pertinentes par rapport aux intérêts récents de Bob.

Nous avons présenté les différentes techniques permettant de calculer le poids temporel des individus (cf. section 4.3.4.2) et des informations (cf. section 4.3.4.3). Nous présentons dans la section suivante, le calcul du poids temporel final de chaque élément extrait à partir des informations partagées par les individus du réseau égocentrique de l'utilisateur.

4.3.4.4. Calcul du poids temporel final d'un élément

Pour calculer le poids temporel final de chaque élément e_{indiv} extrait à partir des informations partagées par les individus du réseau égocentrique de l'utilisateur, nous proposons de combiner le poids temporel de l'individu *indiv* ($P_{temp}^{indiv}(indiv)$ et le poids temporel de cet élément $P_{temp}^{info}(e_{indiv})$.

Pour ce calcul, il nous parait cohérent d'appliquer la même fonction temporelle pour le calcul du poids temporel des individus et le calcul du poids temporel des informations.

De plus, nous observons que, pour les techniques de calcul proposées, le poids temporel des individus est souvent supérieur à 1 alors que le poids temporel des informations est compris entre 0 et 1. La combinaison pourrait par conséquent ne pas être cohérente. Nous appliquons donc, avant de combiner les deux poids, une normalisation en divisant par la valeur max (formule (4.16)) pour ramener les valeurs de ces deux poids entre 0 et 1.

$$score_{normalized} = \frac{score}{score\_max} \quad (4.16)$$

Nous combinons ensuite le poids temporel $P_{temp}^{indiv}(indiv)$ avec le poids $P_{temp}^{info}(e_{indiv})$. Ces deux poids sont calculés en se basant sur deux types d'informations différentes (informations partagées ou relations). Dans un réseau social donné, les informations partagées peuvent être plus importantes que les relations entre les individus ou vice-versa. Nous proposons la formule de combinaison (4.1) qui permet de faire varier l'influence de ces deux poids sur le poids temporel combiné. Le poids temporel de e_{indiv} est calculé par la formule suivante :

$$P_{temp}(e_{indiv}) = combinaision\ (P_{temp}^{indiv}(indiv), P_{temp}^{info}(e_{indiv}), \gamma) \quad (4.17)$$

La valeur γ permet de varier la proportion entre le poids P_{temp}^{indiv} et le poids P_{temp}^{info} et sera déterminée expérimentalement.

Nous présentons l'exemple du calcul du poids temporel final des éléments extraits à partir des informations partagées dans le réseau social de Bob dans l'exemple 4 ci-dessous. Dans cet exemple, le poids temporel d'un individu et le poids temporel d'une information sont calculés en appliquant la fonction temporelle exponentielle f_{exp}.

indiv	Poids temporels de l'individu $P_{temp}^{indiv}(indiv)$		Eléments (e_{indiv})	Poids temporels de l'information $P_{temp}^{info}(e_{indiv})$		Poids temporel final avec $\gamma = 0.5$
	Poids	Poids normalisé		Poids	Poids normalisé	
Carol	0.33287	0.10420	Jeux vidéo	0.61241	0.61241	0.5*(0.10420) + (1-0.5)*0.61241 = 0.35830
			Dessin animé	0.38759	0.38759	0.5*(0.10420) + (1-0.5)*0.38759 = 0.24590
Alice	1.74082	0.54495	Tennis	0.09189	0.09189	0.5*(0.54495) + (1-0.5)*0.09189 = 0.31842
			Voyage	0.63206	0.63206	0.5*(0.54495) + (1-0.5)*0.63206 = 0.58850
			Restaurant	0.27605	0.27605	0.5*(0.54495) + (1-0.5)*0.27605 = 0.41050
Dave	3.19447	1.0	Geek	0.59243	0.59243	0.5*(1.0) + (1-0.5)*0.59243 = 0.79622
			Séries	0.40757	0.40757	0.5*(1.0) + (1-0.5)*0.40757 = 0.70378
Frank	1.97853	0.61936	Geek	1.0	1.0	0.5*(0.61936) + (1-0.5)*1.0 = 0.80968
FFF	0.54881	0.17180	Coupe du monde	0.5	0.5	0.5*(0.17180) + (1-0.5)*0.5 = 0.33590
			Euro	0.5	0.5	0.5*(0.17180) + (1-0.5)*0.5 = 0.33590

Exemple 7 : calcul du poids temporel final des éléments extraits à partir des informations partagées par les individus dans le réseau égocentrique de Bob

Nous pouvons remarquer qu'avec ce calcul, le poids de chaque élément varie en fonction du poids temporel de l'information dont il est extrait et du poids temporel de l'individu qui partage cette information. Par exemple, le poids de l'élément « Jeux vidéo » partagé par Carol est diminué quand on fait la combinaison entre son poids temporel de l'information avec son poids temporel de l'individu (Carol possède le poids minimum par rapport à tous les individus dans le réseau). Le poids temporel de l'élément « Euro » partagé par FFF est également diminué à cause du poids temporel de l'individu de FFF qui est très faible. Le poids temporel de l'élément « Séries » partagé par Dave devient important parce que le poids temporel de Dave est élevé. Le poids temporel de l'élément « Restaurant » est mis en valeur grâce au poids temporel d'Alice qui est plus important que celui de FFF et de Carol.

Les poids dans cet exemple sont calculés en fixant des paramètres $\gamma = 0.5$ et $\lambda = 0.1$. Dans les expérimentations, nous allons faire varier les valeurs de ces paramètres pour trouver celles qui sont optimales. Ainsi, nous allons également tester différentes techniques de calcul de poids temporel d'un individu et le poids temporel d'une information présentées précédemment. Nous allons également tester différentes fonctions temporelles. Les poids temporels des éléments calculés peuvent donc varier selon les techniques utilisées et les valeurs de paramètres fixées.

Pour conclure, nous listons dans la Figure 4.8 ci-dessous, les différentes techniques qui permettent de calculer le poids temporel d'un élément.

Technique de calcul du poids temporel de l'individu

Technique	Description	Calcul
Calculer_P_{Ind_AATemp}	Calcul de poids temporel de de l'individu en appliquant la technique de prédiction de lien AdamicAdar temporelle	$P_{temp}^{indiv}(u,indiv) = \sum_{z \in \{\Gamma(u) \cap \Gamma(indiv)\}} \frac{w(u,z) \cdot w(z,indiv)}{\log\lvert\Gamma(z)\rvert}$
Calculer_P_{Ind_AAStrTemp}	Calcul de poids temporel de de l'individu en appliquant la mesure de prédiction de liens AdamicAdar et la date de sa dernière interaction avec l'utilisateur principal	$P_{temp}^{indiv}(u,indiv) = w(u,indiv) \cdot \sum_{z \in \{\Gamma(u) \cap \Gamma(indiv)\}} \frac{1}{\log\lvert\Gamma(z)\rvert}$
Calculer_P_{Ind_SITemp}	Calcul de poids temporel de de l'individu en appliquant la somme de poids temporel de ses interactions avec l'utilisateur principal	$P_{temp}^{indiv}(u,indiv) = \sum_{\text{int}eract_i \in \text{int}eractions} f_{func}(timestamp_{actuel} - timestamp_{\text{int}eract_i})$

(x)

Fonction temporelle

f_{fun}		Calcul
f_{Exp}	Exponentielle	$f_{\exp}(t) = e^{-\lambda t}$
f_{Poli}	Polinomial	$f_{poli}(t) = (t+1)^{-\lambda}$
f_{Lin}	Linèaire	$f_{poli}(t) = \frac{1}{t+1}$

(x)

Technique de calcul du poids temporel de l'information

Technique	Description	Calcul
Calculer_P_{Inf}	Calcul de poids temporel de l'information en se basant sur la fraîcheur de son timestamp et la fraîcheur de la dernière interaction entre l'individu et l'utilisateur principal	$P_{temp}^{info}(\text{inf}o_{indiv}) = \frac{f_{func}(t_{actuel} - (\text{inf}o_{indiv})) + f_{func}(t_{\text{int}eract} - (\text{inf}o_{indiv}))}{2}$

Figure 4.8 Liste des techniques et fonction temporelles permettant de faire la combinaison de technique pour calculer le poids temporel

La sous-section suivante détaille comment notre méthode de pondération temporelle peut être facilement intégrée dans les processus de construction du profil social.

4.3.5. Application de la méthode temporelle aux processus existants de construction du profil social

La méthode temporelle vue précédemment dans la section 4.3.4 sera appliquée aux deux processus existants qui sont 1) l'approche basée sur les individus et 2) l'approche basée sur les communautés.

L'objectif de l'application de la méthode temporelle proposée est de tenter d'améliorer la pertinence des résultats obtenus dans les processus existants en prenant en compte le facteur temporel.

Pour l'approche basée sur les communautés, la méthode temporelle proposée sera appliquée sur le processus (CoBSP) de (Tchuente, 2013). Pour l'approche basée sur les individus, la méthode temporelle proposée sera appliquée sur le processus basé sur les individus (IBSP) définie dans la partie évaluation du travail de (Tchuente, 2013). En effet, comme cette approche ne possède pas de processus déjà défini, (Tchuente, 2013) s'est appuyé sur les travaux de filtrage social de (Cabanac, 2011) dans le but d'avoir une valeur comparative à son approche basée sur les communautés. Nous intégrons donc à ces deux approches, la prise en compte de l'aspect temporel.

Notons que dans les travaux de (Tchuente, 2013), les 2 processus existants possèdent 4 étapes (cf. section 3.3.3). Dans ce travail, nous présenterons les deux processus adaptés en seulement 3 étapes, afin de correspondre au processus générique de construction de profil social décrit précédemment dans la section 4.2.4.

4.3.5.1. L'approche basée sur les individus

En s'appuyant sur l'approche de construction du profil social basé sur les individus (IBSP), nous représentons le processus temporel de construction du profil social basé sur les individus (IBSPT[63]) pour un utilisateur *u* donné. Nous précisons que la méthode temporelle proposée est intégrée principalement à l'étape 2 du processus.

Etape 1 : Extraire le réseau égocentrique de *u* contenant ses voisins sociaux *INDIV*

Etape 2 : Calculer le profil de chaque $indiv \in INDIV$.

- ***Etape 2.1*** *Extraction de mots-clés et calcul des poids temporels :*

  Pour chaque $info_{indiv} \in INFOS_{indiv}$, nous extrayons un ensemble d'éléments $E_{info_{indiv}}$. Pour chaque élément $e_{info_{indiv}} \in E_{info_{indiv}}$, nous calculons son poids temporel en appliquant la méthode temporelle P_{temp} présentée dans l'algorithme 1.

$$P_{temp}(e_{indiv}) = calculer\_P_{temp}(info_{indiv}) \quad (4.18)$$

  Notons que dans le processus non-temporel existant (IBSP), ce calcul se base uniquement sur la fréquence de termes *tf* de chaque élément dans la communauté et sans normalisation.

[63] T pour « Temporel »

- ***Etape 2.2*** *: Calcul du poids semantico-structurel*

Notons d'abord que cette étape est une étape existante dans le travail de (Tchuente, 2013). Nous reprenons cette étape en remplaçant le poids calculé avec la mesure *tf* par le poids temporel $P_{temp}(e_{indiv})$ calculé de l'étape précédente. Nous définissons des couples $(e_{indiv}, P_{temp}(e_{indiv}))$ pour désigner les éléments avec le poids temporel (ex. ($e1_{indiv}$,0.5), ($e2_{indiv}$, 0.4), …). Pour chaque couple $(e_{indiv}, P_{temp}(e_{indiv}))$, on calcule son poids semantico-structurel par la combinaison du poids structurel de $indiv$ et le poids sémantique de e_{indiv}.

Le poids structurel, noté P_{struct}, de $indiv$ est le degré de centralité de $indiv$ par rapport à tous les individus *INDIVS* dans le réseau égocentrique de l'utilisateur ***u*** (calculé par la formule (3.1)).

$$P_{struct}(indiv) = \ centralitédegre(indiv, INDIVS)$$

(4.19)

Le degré de centralité permet de caractériser les individus en se basant sur leur caractéristique structurelle. Le fait qu'un individu soit complètement isolé ou central dans le réseau égocentrique de l'utilisateur u peut également être porteur d'une information vis-à-vis de u.

- Le poids sémantique, noté P_{sem} , de chaque élément e_{indiv} est calculé en se basant sur le score $P_{temp}(e_{indiv})$.

$$P_{sem}(e_{indiv}) = P_{temp}(e_{indiv})$$

(4.20)

Bien qu'on ne puisse pas a priori indiquer la qualité ou l'impact des poids structurel et sémantique sur la pertinence du poids d'un élément *e*, nous combinons ces deux poids par une fonction de combinaison avec le paramètre α qui désigne la proportion entre les deux poids et qui sera déterminé pendant les expérimentations. Nous obtenons donc le poids semantico-structurel de l'élément e_{indiv}, noté $P_{semstruct}$ avec la formule ci-dessous.

$$P_{semstruct}\ (e_{indiv}) = \ combinaison(P_{struct}(indiv), P_{sem}(e_{indiv}), \alpha)$$

(4.21)

En reprenant l'étude de cas de Bob, nous donnons ci-dessous l'exemple de calcul du poids semantico-structurel $P_{semstruct}$ des éléments extraits à partir des informations partagées par Carol en se basant sur leur poids structurel et leur poids sémantique.

indiv	**Poids structurel** $P_{struct}(indiv)$	**Eléments** (e_{indiv})	**Poids sémantique** $P_{sem}(e_{indiv})$,	$P_{semstruct}$ **avec** $\alpha = 0.1$
Carol	0.25	Jeux vidéo	0.35830	0.1* (0.25) + (1 - 0.1)* (0.35830) =0.34747
		Dessin animé	0.24590	0.1* (0.25) + (1 - 0.1)* (0.24590) =0.24631
Alice	0.25	Voyage	0.58850	0.1* (0.25) + (1 - 0.1)* (0.58850) =0.55465
		Restaurant	0.41050	0.1* (0.25) + (1 - 0.1)* (0.41050) =0.39445
		Tennis	0.31842	0.1* (0.25) + (1 - 0.1)* (0.31842) =0.31158
Dave	0.25	Geek	0.79622	0.1* (0.25) + (1 - 0.1)* (0.79622) =0.74160
		Séries	0.70378	0.1* (0.25) + (1 - 0.1)* (0.70378) =0.65840
Frank	0.25	Geek	0.80968	0.1* (0.25) + (1 - 0.1)* (0.80968) =0.75371
FFF	0	Euro	0.33590	0.1* (0.0) + (1 - 0.1)* (0.33590) =0.30231
		Coupe du monde	0.33590	0.1* (0.0) + (1 - 0.1)* (0.33590) =0.30231

Exemple 8 : calcul du poids semantico-structurel $P_{semstruct}$ des éléments extraits à partir des informations partagées par Carol.

Etape 3 : Dans cette étape, nous combinons et dérivons les éléments du profil de tous les individus *INDIVS* pour construire le profil social de l'utilisateur *u*. A la fin de l'étape 2 chaque élément peut avoir différents poids **semantico-structurel** dans le profil des différents individus. Pour calculer le poids final de cet élément, nous appliquons la fonction *Lin_CombinMNZ* proposé par (Hubert, Loiseau et Mothe, 2007) pour combiner les différents poids depuis les différents individus.

La fonction *Lin_CombMNZ* est utilisée en recherche d'information pour 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. Dans notre contexte, la formule associée à la *Lin_CombMNZ* est la suivante (4.22) :

$$Lin\_CombMNZ(e) = \sum_{i=1}^{INDIVS\,|P_{semstruct}(e_{indiv_{i-1}}) < P_{semstruct}(e_{indiv_i})} \left(P_{semstruct}(e_{indiv_i}) * i\right) \quad (4.22)$$

i est le rang d'un individu ($indiv_1$ représente l'individu qui est au dernier rang), $P_{semstruct}\left(e_{indiv_i}\right)$ est le poids calculé pour l'intérêt *e* pour l'individu $indiv_i$, $|INDIVS|$ est le nombre total d'individus. Cette fonction agit en deux temps pour trouver le score de fusion de chaque élément. Dans un premier temps, les individus sont classés par ordre croissant en fonction du poids de l'élément de son profil $P_{semstruct}(e_{indiv_{i-1}}) < P_{semstruct}(e_{indiv_i})$. Dans un second temps (la combinaison linéaire), le poids attribué par chaque individu est multiplié

par le rang *i* des individus dans la classification ordonnée précédente. Ainsi, Si on dispose de 5 individus, le score de l'individu qui dispose du score le plus important pour l'élément *e* est multiplié par 5, le score du second individu attribuant le score le plus important pour l'élément *e* est multiplié par 4, ..., le score de l'individu attribuant le score le moins important pour l'élément *e* est multiplié par 1.

Le poids final, noté P_{social} de chaque élément est calculé avec la formule (4.23) suivante :

$$P_{social}(e) = Lin\_CombinMNZ\ (e\ )$$

(4.23)

Finalement, nous dérivons chaque élément calculé pour obtenir un vecteur pondéré d'éléments représentant les intérêts sociaux de l'utilisateur (profil social de l'utilisateur).

En reprenant l'étude de cas de Bob, nous donnons ci-dessous l'exemple de calcul du poids final des éléments extraits à partir des informations partagées par les individus dans son réseau égocentrique, en appliquant la fonction de combinaison *Lin_CombMNZ*. Notons que les poids $P_{semstruct}$ donnés dans l'exemple sont calculés en utilisant les poids temporels calculés dans l'exemple 4 et en fixant la valeur de $\alpha = 0.01$.

e / *indiv*	Jeux vidéo	Dessin animé	Tennis	Voyage	Restaurant	Geek	Séries	Coupe du monde	Euro
Carol	0.34747	0.24631							
Alice			0.31158	0.55465	0.39445				
Dave						0.74160	0.6584		
Frank						0.75371			
FFF								0.30231	0.30231
Classement	1 : 0.34747	1: 0.24631	1 : 0. 31158	1 : 0. 55465	1 : 0. 39445	1 : 0. 74160 2 : 0. 75371	1 : 0.6584	1 : 0. 30231	1 : 0. 30231
Poids *Lin_CombMNZ*	0. 34747* 5 = 1.73735	0. 24631*5 =1.23154	0. 31158*5 =1.55788	0. 55465*5 =2.77326	0. 39445*5 =1.97225	0. 74160*5 + 0. 75371*4 =6.73494	0.6584*5 =3.29202	0. 30231 *5 =1.51155	0. 30231*5 =1.51155
Ordre final	Geek > Séries > Voyage >Restaurant>Jeux vidéo > Tennis > (Coupe du monde =Euro) > Dessin animé								

Exemple 9 : calcul du poids final des éléments à partir de tous les individus dans le réseau égocentrique de Bob.

Nous pouvons constater qu'avec l'application de la méthode temporelle proposée sur le processus de construction du profil social existant, les intérêts qui sont pertinents et à jour (« Voyage », « Séries », « Restaurant ») sont bien mieux classés que les intérêts qui deviennent obsolètes pour l'utilisateur ou qui n'ont pas de lien avec l'utilisateur (« Jeux vidéo », « Coupe du monde », « Euro », « Dessin animé »). L'intérêt récurrent comme « Geek » reste encore renforcé.

Nous présentons ci-dessous l'algorithme du processus temporel de construction du profil social IBSPT.

Entrée : u **Sortie** : Profil Social de u

DEBUT

ProfilSocial$_u$ = {} //profil social de u
//Etape1 extraction du réseau égocentrique de l'utilisateur
$G' = Réseau\ égocentrique(u)$;
$INDIVS = liste\ des\ noeuds\ de\ G' - u$; //individu dans le réseau égocentrique G'
$E_{INDIVS} = \{\}$ //L'ensemble des listes d'élément de tous les INDIVS
// Etape 2 : extraction et pondération des mots-clés
Pour chaque $indiv \in INDIVS$ Faire
  $INFOS_{indiv} = extraireInformations(indiv)$ //liste d'informations de indiv
  $E_{indiv} = \{\}$ //liste des éléments et leur poids temporel extraits à partir de $INFOS_{indiv}$

  //2.1 pondération temporelle des éléments
  Pour chaque $info_{indiv} \in INFO_{indiv}$ Faire
    //Extraction de l'ensemble des éléments à partir de $info_{indiv}$
    $E_{info_{indiv}}$ = ExtraireEléments($info_{indiv}$)
    Pour chaque $e_{info_{indiv}} \in E_{info_{indiv}}$ Faire
      //calcul de poids temporel de l'élément e avec la méthode temporelle (Algorithme 1.1)
      $P_{temp}(e_{indiv}) = calculerP_{temp}(info_{indiv})$
      //ajouter l'élément et son poids dans E_{indiv}
      $E_{indiv}.ajouter(e_{info_{indiv}}, P_{temp}(e_{info_{indiv}}))$
    Fin pour
  Fin pour
  //2.3 Calcul du poids semantico-structurel
  $E'_{indiv} = \{\}$ //liste des éléments et leur poids semantico-structurel
  $P_{struct}(indiv) = centralitédegre(indiv_i, INDIVS)$
  Pour chaque $(e_{indiv}, P_{temp}(e_{indiv})) \in E_{indiv}$ faire
    $P_{sem}(e_{indiv}) = \frac{P_{temp}(e_{indiv})}{\sum_{i \in INDIVs} P_{temp}(e_i)}$
    $P_{semstruct}(e_{indiv}) = combinaison(P_{struct}(indiv), P_{sem}(e_{indiv}), \alpha)$
    //Ajouter le poids e_{indiv} au E'_{indiv} avec son poids semantico-structurel
    $E'_{indiv}.ajouter(e_{indiv}, P_{semstruct}(e_{indiv}))$
  Fin pour
  //Ajouter la liste d'élément pondéré de *indiv* au E_{INDIVS}
  $E_{INDIVS}.ajouter(E'_{indiv})$
Fin pour

//Etape 3 : Combinaison des éléments des tous les individus INDIVS pour dériver le profil social de u
E'' = Ensemble des éléments apparaissant dans au moins un ensemble $E'_{indiv} \in E_{INDIVS}$
Pour chaque $e \in E''$ Faire
  // Combiner les différents poids $P_{SemStruct}$ associes à e pour différents *indiv* :
  $P_{social}(e) = Lin\_CombinMNZ(e, E_{INDIVS})$ //
  $ProfilSocial_u.ajouter(e, P_{social}(e))$
Fin pour

RETOURNER $ProfilSocial_u$

FIN

Algorithme 2: processus IBSP**T**

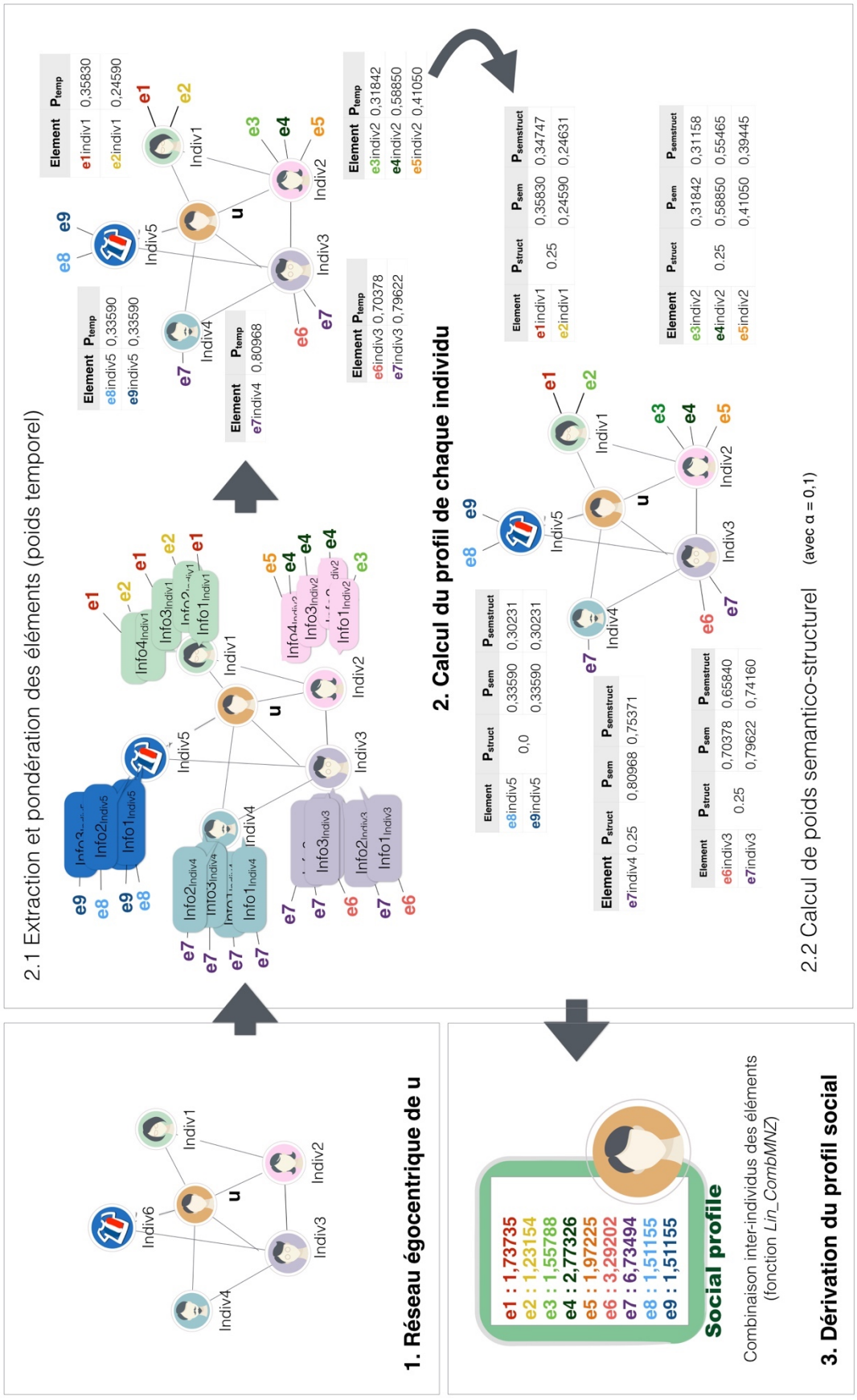


Figure 4.9 Illustration du processus IBSP

4.3.5.2. L'approche basée sur les communautés

Nous présentons ici le processus temporel de construction du profil social basé sur les communautés (CoBSPT) pour un utilisateur *u* donné. Comme dans l'approche IBSPT, la méthode temporelle proposée est intégrée principalement à l'étape 2 du processus de construction du profil social basé sur les communauté (CoBSP) proposé par (Tchuente, 2013). Notons que cette approche est similaire à l'approche IBSPT précédente mais s'applique au niveau des communautés et non au niveau des individus.

Etape 1 : Extraire le réseau égocentrique de *u* contenant ses voisins sociaux *INDIVS et* extraire l'ensemble des communautés *C* à partir du réseau égocentrique en utilisant l'algorithme iLCD (cf. étape 1 du processus CoBSP de (Tchuente, 2013) section 3.3.3).

Etape 2 : Calculer le profil de chaque communauté $c \in C$.

Le profil de chaque communauté est calculé par agrégation des éléments extraits à partir de tous les individus de la communauté.

- ***Etape 2.1*** *Extraction et calcul du poids temporel des éléments pour chaque individu*
  Cette étape consiste à extraire, pour chaque $indiv \in c$, les éléments à partir des informations qu'il partage et à calculer leur poids temporel. Pour chaque $indiv \in c$, pour chaque $info_{indiv} \in INFOS_{indiv}$, nous extrayons un ensemble d'éléments $E_{info_{indiv}}$. Pour chaque élément $e_{info_{indiv}} \in E_{info_{indiv}}$, nous calculons son poids temporel en appliquant la méthode temporelle P_{temp} présentée dans l'algorithme 1.

$$P_{temp}(e_{indiv}) = calculer\_P_{temp}(info_{indiv}) \quad (4.24)$$

- ***Etape 2.2*** *Agrégation d'éléments au niveau de la communauté*

Une fois que le poids temporel de chaque élément à partir de toutes les $INFOS_{indiv}$ de tous les $indiv \in c$ est calculé, nous agrégeons ces éléments en faisant la somme des poids de chaque élément, normalisée par le nombre d'individus dans la communauté c. Notons $P'_{temp}(e_c)$, le poids agrégé d'un élément *e* trouvé dans la communauté *c*.

$$P'_{temp}(e_c) = \frac{\sum_{indiv \in c} P_{temp}(e_{indiv})}{n}$$

n représente le nombre d'individus dans *c* (4.25)

- ***Etape 2.3*** *: Calcul du poids semantico-structurel de la communauté*

Cette étape est une étape existante dans le travail de (Tchuente, 2013). Comme dans le processus IBSPT, nous reprenons cette étape en remplaçant le poids calculé avec la mesure *tf* par le poids temporel $P'_{temp}(e_c)$ calculé de l'étape précédente. Nous définissons des couples $(e_c, P'_{temp}(e_c))$ pour désigner les éléments avec leur poids temporel (ex. ($e1_c$,0.5), ($e2_c$, 0.4), …). Pour chaque couple $(e_c, P'_{temp}(e_c))$, nous calculons son poids semantico-structurel par la combinaison du poids structurel de *c* et du poids sémantique de e_c.

Le poids structurel, noté P_{struct}, de c est le degré de centralité de c par rapport à l'ensemble C de toutes les communautés dans le réseau égocentrique de l'utilisateur u (calculé par la formule (3.5)).

$$P_{struct}(c) = \ centralitédegre(c, C)$$

(4.26)

Le poids sémantique, noté P_{sem} , de chaque élément e_c est calculé en se basant sur le poids $P_{temp}(e_c)$

$$P_{sem}(e_c) = \ P_{temp}(e_c)$$

(4.27)

Nous combinons ensuite le poids structurel et le poids sémantique de chaque e par une fonction de combinaison avec le paramètre α qui désigne la proportion entre les deux poids et qui sera déterminé lors des expérimentations. Nous obtenons donc le poids semantico-structurel de l'élément e_c, noté $P_{semstruct}$ avec la formule ci-dessus.

$$P_{semstruct}\ (e_c) = \ combinaison(P_{struct}(c), P_{sem}(e_c), \alpha)$$

(4.28)

Etape 3 : Dans cette étape, nous combinons et dérivons les éléments du profil de toutes les communautés C dans le profil social de l'utilisateur u. A la fin de l'étape 2, chaque élément peut avoir différents poids **semantico-structurel** dans le profil des différentes communautés. Pour calculer le poids final de cet élément, nous appliquons, comme dans le cas de l'approche basée sur les individus, la fonction *Lin_CombinMNZ* (formule (4.22)) pour combiner les différents poids depuis les différentes communautés. Le poids final, noté P_{social} de chaque élément est calculé avec la formule suivante.

$$P_{social}(e) = \ CombinMNZ\ (e\ )$$

(4.29)

Finalement, nous dérivons chaque élément calculé pour obtenir un vecteur pondéré d'éléments représentant les intérêts sociaux de l'utilisateur (profil social de l'utilisateur).

Nous présentons ci-dessous l'algorithme du processus temporel de construction du profil social CoBSPT.

Entrée : u **Sortie** : Profil Social de u
DEBUT
1: ProfilSocial$_u$ $= \{\}$ //profil social de u
2: //Etape1 extraction du réseau égocentrique de l'utilisateur
3: $G' = Réseau\ égocentrique(u)$;
4: $INDIVS = liste\ des\ noeuds\ de\ G' - u$; //individu dans le réseau égocentrique G'
5: $C=$ iLCD(G') // extraction des communautés par l'algorithme iLCD
6: // Etape 2 : Profilage de communautés

7: **Pour chaque** $c \in C$
8: $E_c = \{\}$ //L'ensemble de liste des éléments de tous les $INDIVS \in c$
9: **Pour chaque** $indiv \in c$ Faire
10: $INFOS_{indiv} = extraireInformations(indiv)$ //liste d'informations de indiv

11: //2.1 pondération temporelle de mot clé
12: **Pour chaque** $info_{indiv} \in INFO_{indiv}$ Faire
13: $E_{info_{indiv}}$ = ExtraireEléments($info_{indiv}$)
14: **Pour chaque** $e_{info_{indiv}} \in E_{info_{indiv}}$ Faire
15: //calcul de poids temporel de l'élément e avec la méthode temporelle Algorithme 1
16: $P_{temp}(e_{info_{indiv}}) = calculerP_{temp}(info_{indiv})$
17: //ajouter l'élément et son poids dans E_{indiv}
18: $E_c.ajouter(e_{info_{indiv}}, \mathrm{P_{temp}}(e_{info_{indiv}}))$
19: **Fin pour**
20: **Fin pour**
21: **Fin pour**
22:
23: //2.2 Agrégation de mot clés au niveau de la communauté
24: E_{temp} = Ensemble des éléments apparaissant au moins une fois dans E_c
25: $E'_c = \{\}$
26: **Pour chaque** élément $\mathrm{e_c} \in E_{temp}$ Faire
27: $\mathrm{P'_{temp}(e_c)} = \frac{\sum \mathrm{P_{temp}(e_c)}}{\sum_{\mathrm{f} \in \mathrm{E_c}} \mathrm{P_{temp}(f)}}$ //agrégation des poids pour chaque $\mathrm{e_{indiv}}$
28: // Ajouter le couple $\mathrm{e_c}$, $\mathrm{P'_{temp}(e_c)}$ à E'_c
29: $E'_c.ajouter(\mathrm{e_c}, \mathrm{P'_{temp}(e_c)})$
30: **Fin pour**
31:
32: //2.3 Calcul de poids semantico-structurel
33: $P_{struct}(c) = centralitédegre(c, C)$
34: **Pour chaque** $(\mathrm{e_c}, \mathrm{P'_{temp}(e_c)}) \in E'_c$ faire
35: $P_{sem}(e_c) = \frac{P\prime_{temp}(e_c)}{\sum_{i \in C} P\prime_{temp}(e_i)}$
36: $P_{semstruct}(\mathrm{e_c}) = \mathrm{combinaison}(\mathrm{P_{struct}}(c), \mathrm{P_{sem}}(\mathrm{e_c}), \alpha)$
37: //Mettre à jour le poids de e_c avec son poids semantico-structurel
38: $E'_{indiv}.mettreAJourPoids(e_c, P_{semstruct}(e_c))$
39: **Fin pour**
40: //Ajouter la liste d'éléments pondérés trouvés dans c au E_C
41: $E_C.ajouter(E'_c)$
42: **Fin pour**
43:
44: //Etape 3 : Combinaison des éléments des tous les individus INDIVS pour dériver le profil social de u
45: E'' = Ensemble des éléments apparaissant dans au moins un ensemble $E'_c \in E_C$
46: **Pour chaque** $e \in E''$ Faire
47: // Combiner les différents poids $P_{SemStruct}$ associes à e pour différents $indiv$:
48: $P_{social}(e) = Lin\_CombinMNZ(e, E_C)$ //
49: $ProfilSocial_u.ajouter(e, P_{social}(e))$
50: **Fin pour**
51:
52: **RETOURNER** $ProfilSocial_u$
FIN

Algorithme 3: processus CoBSP**T**

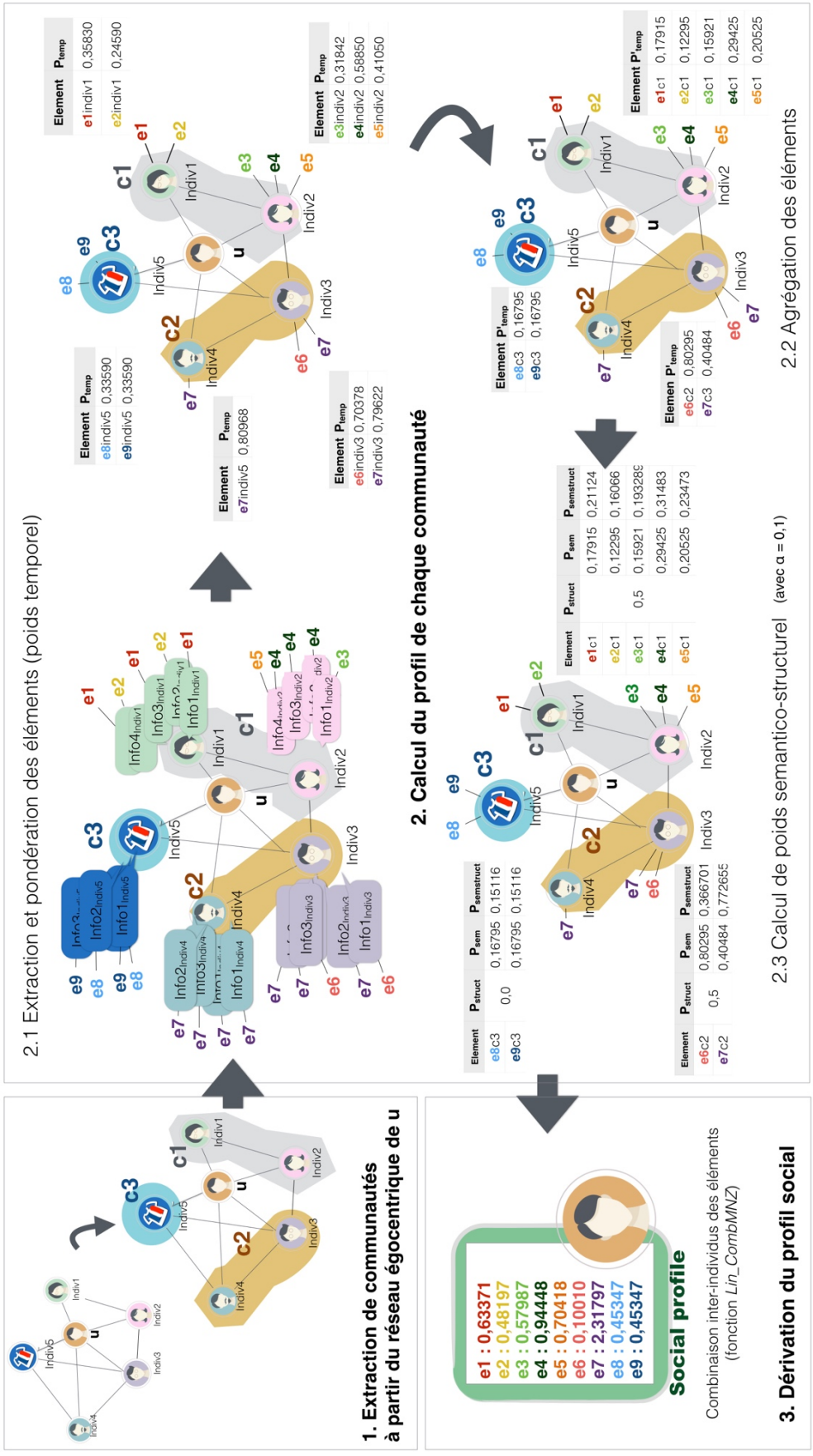


Figure 4.10 Illustration du processus CoBSPT

Les algorithmes vus précédemment utilisent différents paramètres (λ, γ, α). Dans ce qui suit, nous présentons les objectifs et les principes généraux concernant l'étude des différents paramètres en fonction des types et des propriétés des réseaux sociaux.

4.4. Etude paramétrique suivant les types et les propriétés des réseaux sociaux

4.4.1. Etude paramétrique

Dans les formules utilisées dans le processus de construction du profil social, plusieurs paramètres (λ, γ, α) interviennent quelle que soit l'approche utilisée (basée sur les individus ou basée sur les communautés). De ce fait, pour ces deux approches, nous avons réalisé une étude paramétrique pour trouver les valeurs les plus appropriées pour construire un profil social le plus pertinent possible. Dans cette étude, les valeurs de chaque paramètre permettent de prendre en compte de différentes manières les caractéristiques du réseau :

- **γ** (formule(4.17) section 4.3.4) représente la proportion entre le poids temporel des informations et le poids temporel des individus,. Cette valeur nous permet de faire varier, selon le type de réseau, l'importance du poids temporel des individus par rapport à celle du poids temporel des informations (permet de donner plus d'importance aux informations ou aux relations).

- **λ** représente le taux de dépréciation des informations ou des relations (*time decay rate*) pour la fonction temporelle exponentielle (formule(2.7) section 2.3.1.2**)** et pour la fonction temporelle polynomiale (formule (2.8), section 2.3.1.2). Les taux de dépréciation de l'information et celui des relations peuvent être différents. Deux valeurs de **λ** (selon le calcul réalisé, information ou relation) peuvent montrer plus précisément, selon le type de réseau, le taux de dépréciation de l'information et celui des relations entre l'utilisateur et les individus dans son réseau social. Notons que la fonction linéaire inverse, on n'utilise pas la valeur de λ.

- **α** est défini dans l'étape du calcul du poids sémantico-structurel des intérêts (formules (4.21) et (4.28) section 4.3.5). Ce paramètre représente la proportion du poids structurel par rapport à celle du poids sémantique d'un intérêt et a été étudié dans le travail de (Tchuente, 2013).

L'étude paramétrique nous permettra de trouver la meilleure (ou une des meilleures) combinaison des paramètres, afin d'obtenir le profil social le plus pertinent. Celle-ci permet également d'étudier l'impact de la dynamique des relations et celle des informations selon le réseau social étudié.

L'étude paramétrique que nous proposons, consiste à combiner les différentes valeurs des trois paramètres de la façon suivante :

Soit $\gimel$ l'ensemble des valeurs possibles pour **γ,** $\mathcal{L}$ l'ensemble des valeurs possibles pour λ et A l'ensemble des valeurs possibles pour α, la combinaison des différentes valeurs des trois paramètres donnera lieu à la construction de $|\gimel * \mathcal{L} * A|$ profils sociaux.

La section suivante présente les deux axes sur lesquels l'étude paramétrique sera menée.

4.4.2. Analyse des résultats de l'étude paramétrique suivant le type et les propriétés du réseau social

Etant donnée la nature hétérogène des réseaux sociaux (vitesse d'évolution des informations, des relations, nature des liens dans le réseau…), les valeurs optimales trouvées pour les paramètres peuvent être différentes d'un type de réseau à un autre. Pour un même type de réseau social, différentes propriétés du réseau social pourraient également donner différents résultats.

Nous proposons donc d'étudier les résultats du processus de construction du profil social par rapport aux types et aux propriétés du réseau égocentrique de l'utilisateur.

4.4.2.1. Etude selon le type de réseau social

La méthode proposée est mise en œuvre sur deux types de réseaux sociaux : DBLP qui est un réseau de publications scientifiques et Twitter qui est un réseau de micro-blogs. Ces réseaux sociaux possèdent différentes caractéristiques en termes d'objectifs d'utilisation, de types d'informations partagées, de types de relations et d'interactions entre les individus dans le réseau mais aussi dans leurs aspects évolutifs. Ces expérimentations permettront d'étudier la pertinence des processus de construction du profil social proposés par rapport aux deux types de réseau sociaux étudiées.

4.4.2.2. Etude selon les propriétés du réseau égocentrique de l'utilisateur

La taille du réseau égocentrique est le nombre d'individus dans le réseau égocentrique donc le nombre de voisins sociaux de l'utilisateur. Autrement dit, la taille du réseau est son degré dans le réseau social entier (cf. section 3.2.2.1.a). Nous supposons que la taille du réseau peut influencer le résultat du processus de construction du profil social. En effet, si la taille du réseau égocentrique de l'utilisateur est importante, il y a certainement un nombre important d'informations partagées dans ce réseau. De ce fait, il pourrait y avoir plus de bruit dans les informations traitées et éventuellement une évolution plus importante des informations. On pourrait avoir l'effet inverse dans le cas d'un réseau ayant moins de voisins sociaux.

La densité du réseau égocentrique est le nombre de liens sur le nombre de liens possibles dans le réseau égocentrique. Nous supposons que la densité du réseau peut également influencer le résultat du processus de construction du profil social. La formule de calcul de la densité du réseau social (3.8) est donnée dans la section 3.2.2.1.b,

Nous étudierons donc, dans la partie expérimentations, l'impact de la taille et de la densité du réseau égocentrique de l'utilisateur sur la pertinence de la méthode temporelle proposée.

4.5. Conclusion

Ce chapitre explique comment prendre en compte la caractéristique évolutive des réseaux sociaux dans le processus de construction du profil social de l'utilisateur. Nous avons proposé une méthode temporelle qui s'applique principalement dans l'étape d'extraction et de pondération des éléments (mots-clés) de ce processus. Cette méthode permet de pondérer les intérêts en se basant sur le poids temporel des informations extraites (fraicheur) et sur le poids temporel des individus (force des liens des individus avec l'utilisateur).

Cette méthode a été intégrée dans les processus existants de construction du profil social selon l'approche basée sur les individus et l'approche basée sur les communautés.

Nous avons fait en sorte que cette méthode temporelle soit la plus générique possible pour pallier l'hétérogénéité des réseaux sociaux en effet :

- d'une part, elle permet de paramétrer, selon le type de réseau étudié, les techniques utilisées dans le calcul des poids temporels des individus et celui des informations. Ces techniques seront évaluées lors des expérimentations pour retenir celle qui est la plus appropriée pour chaque type de réseau.
- d'autre part, dans le calcul des poids temporels des individus et des informations, les fonctions temporelles exploitées sont également paramétrables et seront elles aussi évaluées lors des expérimentations.

Une étude paramétrique permettra de trouver la valeur optimale des paramètres utilisés dans le calcul des poids temporels :

- L'importance (proportion) du poids temporel des individus par rapport au poids temporel des informations dans le calcul du poids temporel final sera étudiée pour trouver la combinaison optimale.
- Le taux de dépréciation (*Time Decay Rate*) des informations et des relations sera également étudié pour trouver la valeur optimale.

Dans les expérimentations menées, nous avons appliqué la méthode et les techniques de calcul proposés et nous avons étudié les paramètres dans deux types de réseaux sociaux. L'objectif de ces expérimentations est d'évaluer l'efficacité de notre proposition et faire émerger, en fonction des résultats obtenus, les facteurs importants qui permettent de construire un profil social pertinent selon le type du réseau étudié.

Les expérimentations et évaluations de notre contribution sont détaillées dans le chapitre suivant.

5. EXPERIMENTATIONS

Dans cette section, nous présentons l'évaluation de la méthode et des techniques de calcul proposées au chapitre 4. Nous commençons par donner la synthèse sur les stratégies d'évaluation qui pourraient être utilisées dans notre contexte. Puis, nous présentons le protocole d'évaluation proposé. Ensuite, nous présentons les expérimentations menées pour chaque réseau social étudié (DBLP, Twitter) ainsi que les résultats obtenus que nous discutons dans la dernière section.

5.1. Synthèse sur les stratégies d'évaluation de la proposition

Dans la littérature, nous trouvons trois stratégies d'évaluation principales qui pourraient être utilisées pour évaluer la méthode et les techniques de calcul présentées dans le chapitre précédent : l'évaluation par confrontation à la perception humaine, l'évaluation automatisée par filtrage social, l'évaluation automatisée par comparaison entre profils sociaux. Pour chacune de ces stratégies, nous détaillons ses avantages et ses inconvénients pour, au final, en retenir une qui est automatisable et induit le moins de distorsions entre le phénomène observé et les conclusions tirées sur la pertinence des contributions.

5.1.1. Evaluation par confrontation à la perception humaine

Cette stratégie d'évaluation permet de n'évaluer que les algorithmes de la phase de dérivation du profil social. Les profils sociaux sont construits et proposés aux utilisateurs pour qu'ils en jugent la pertinence, c'est-à-dire dans quelle mesure le profil social construit (mots-clefs) correspond à ou traduit leurs intérêts personnels réels.

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. L'inconvénient de cette stratégie est qu'il peut être difficile de la mettre en œuvre dans le cas de jeux de données importants. En effet, il serait difficile de recueillir 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 et minimiser les biais induits par cette présentation.

5.1.2. Evaluation automatisée par filtrage social

L'évaluation automatisée par filtrage social est la stratégie d'évaluation couramment utilisée dans les travaux de la littérature. Les travaux de filtrage social, 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 ceux 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, il s'agira d'évaluer l'impact du profil social obtenu 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.

L'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. De plus, une évaluation automatique est envisageable en n'exploitant que le suivi ou non de la recommandation par l'utilisateur, sans jugement de pertinence sur l'ensemble des recommandations. De ce fait, ce type d'évaluation automatique peut être appliqué facilement dans les cas de jeux de données importants.

Cependant, cette stratégie de validation peut avoir deux inconvénients majeurs. Le premier est lié à la confusion sur l'objet de validation. Dans la modélisation de l'utilisateur, on dispose des algorithmes de construction du profil social et du profil utilisateur individuel. Dans les mécanismes de filtrage, on peut également disposer des algorithmes d'exploitation de ce profil social en le combinant avec le profil utilisateur individuel dans la plupart des cas. Dans ce cas, la question qui se pose lors de l'évaluation des mécanismes de filtrage est de savoir s'il s'agit d'évaluer les techniques de filtrage ou les techniques de construction du profil social.

Le second inconvénient repose sur l'importance des ressources à mobiliser. Ceci est un corollaire du premier inconvénient. En effet, si on évalue à la fois les techniques de filtrage et les techniques de construction du profil, l'évaluation s'avère beaucoup plus complexe. Si l'on dispose de n techniques de construction du profil et de m techniques d'usage des profils sociaux dans les mécanismes de filtrage, il faudra évaluer $n*m$ possibilités de filtrage d'information. Ceci devient très lourd à mettre en œuvre, surtout dans le cas d'une évaluation semi-automatisée qui implique des jugements des utilisateurs sur chacune des possibilités.

5.1.3. Evaluation automatisée et comparative entre profil social et profil utilisateur individuel

Cette stratégie d'évaluation consiste à évaluer uniquement les techniques de dérivation du profil social en comparant les profils engendrés avec le profil de l'utilisateur (son profil individuel explicite) lorsqu'il existe. Cette stratégie d'évaluation repose sur l'hypothèse que le profil utilisateur individuel contient les intérêts pertinents de l'utilisateur. Pour valider la pertinence

des algorithmes de construction du profil social, on les compare, en recherchant celui qui produit le profil social le plus proche du profil utilisateur. Dans cette stratégie d'évaluation, il faudrait donc considérer les profils utilisateurs individuels qui se rapprochent le plus des intérêts réels des utilisateurs : profils renseignés explicitement par les utilisateurs eux-mêmes, ou ne considérer que les utilisateurs ayant un volume d'activité très important permettant de construire un profil utilisateur individuel pertinent par rapport aux intérêts réels de l'utilisateur. Dans ce cadre, les mesures telles que la précision, le rappel ou le cosinus entre les vecteurs des profils sociaux et le vecteur de profil utilisateur peuvent être utilisées si les profils sont présentés par des vecteurs de termes pondérés comme dans notre travail.

L'avantage de cette stratégie est qu'elle permet l'évaluation des algorithmes de dérivation du profil social de façon directe et indépendamment des mécanismes de filtrage d'information. De plus, cette évaluation qui est automatique peut permettre des évaluations sur des jeux de données très importants.

Cette stratégie peut avoir un inconvénient dans le cas où le profil utilisateur individuel construit n'est pas représentatif de l'utilisateur réel ; des résultats positifs de comparaison ne donneraient aucune validité au profil construit. Autrement dit, il faut que le profil individuel de l'utilisateur soit pertinent lors de l'évaluation du profil social. Notons que dans certains cas, il se peut que l'on ne dispose pas du profil explicite de l'utilisateur. Le profil individuel de l'utilisateur doit être construit à partir des informations que l'on possède de lui. Il faut donc que le processus utilisé pour construire ce profil donne un résultat pertinent pour avoir un terrain de vérité significatif. Néanmoins, la littérature fournit plusieurs méthodes de construction du profil individuel d'un utilisateur

La section suivante détaille le protocole d'évaluation mis en œuvre dans nos travaux.

5.2. Protocole d'évaluation

5.2.1. Stratégie d'évaluation utilisée

Dans notre contexte, l'évaluation automatisée et comparative entre profil social et profil utilisateur parait la plus appropriée. En effet, nous avons besoin d'un volume assez important d'échantillons de tests pour renforcer et prouver la pertinence de la méthode proposée et pour laquelle plusieurs combinaisons de paramètres et de méthodes de calcul sont à évaluer. La première évaluation par confrontation à la perception humaine est lourde à mettre en place et la seconde par filtrage social nous paraît impliquer trop de biais. Enfin, ces deux solutions deviennent inapplicables pour évaluer de nombreuses combinaisons de paramètres et de méthodes de calcul.

Comme expliqué dans la section 4.3.5, la méthode temporelle proposée est appliquée aux deux processus de construction du profil social existants qui ne prennent pas en compte l'évolution du réseau social de l'utilisateur et qui sont définis dans (Tchuente, 2013). Nous obtenons donc deux processus de construction du profil social (que nous appellerons « profil social temporel ») suivants :

- Processus IBSPT (*Individual Based Social Profiling – Temporal)* étendu du processus IBSP, dérivé de l'approche basée sur les individus,
- Processus CoBSPT (*Community Based Social Profiling – Temporal)* étendu du processus CoBSP, dérivé de l'approche basée sur les communautés.

Dans le travail de (Tchuente, 2013), les deux processus IBSP et CoBSP ont été comparés entre eux dans le contexte du réseau de publications scientifiques DBLP et du réseau social Facebook. Dans (Tchuente, 2013), le processus IBSP déduit à partir des travaux sur le filtrage social est défini comme une référence (« *baseline* ») pour évaluer le processus CoBSP ; c'est en effet une méthode classiquement utilisée pour calculer un profil utilisateur individuel. Dans le cas de DBLP, le processus CoBSP produit de meilleurs résultats que le processus IBSP pour les utilisateurs possédant au moins 50 voisins sociaux. Les évaluations relatives à la densité du réseau égocentrique permettent d'observer que plus la densité du réseau 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 est épars (peu dense), moins les communautés extraites par l'algorithme de détection de communautés sont réellement significatives pour l'égo. Alors que plus ce réseau est dense, plus l'algorithme de détection de communautés est capable d'extraire des communautés significatives et représentatives pour l'égo.

Dans ce travail, nous comparons donc les processus étendus qui intègrent la méthode temporelle proposée avec les processus existants du travail de (Tchuente, 2013) (IBSPT avec IBSP et CoBPT avec CoBSP). Nous allons également comparer IBSPT avec CoBSPT.

Dans la méthode temporelle proposée, différentes techniques de calcul de poids temporel d'informations et d'individus ainsi que différentes fonctions temporelles ont été présentées et vont être paramétrées, combinées et exploitées. Nous évaluons également les profils construits à partir de ces différentes techniques et fonctions temporelles pour trouver la ou les technique(s)/fonction(s) temporelle(s) appropriée(s). Notons que nous ne présenterons ici que les profils sociaux construits avec les combinaisons qui donnent des résultats significatifs et/ou intéressants. Compte tenu du nombre important de calculs réalisés (par combinaison des fonctions et paramètres), les résultats ne peuvent être présentés intégralement dans le volume de ce mémoire. Tous les calculs énoncés ont été réalisés, les résultats significatifs sont présentés dans ce chapitre et des résultats complémentaires sont présentés en annexe.

La stratégie d'évaluation retenue est d'appliquer les algorithmes proposés, d'en comparer les résultats et de valider celui qui permet de construire le profil social le plus proche du profil explicite de l'utilisateur en utilisant les mesures de similarité suivantes : la précision, le rappel et la F-mesure (cf. Figure 5.1). Ces mesures classiques en RI nous permettrons de valider les contenus des profils sociaux construits : avec une précision forte, les intérêts construits sont représentatifs, avec un rappel élevé, les intérêts construits permettent de couvrir un nombre important d'intérêts de l'utilisateur.

Dans un premier temps, notre objectif est d'évaluer la pertinence de la méthode temporelle proposée par rapport aux processus qui ne prennent pas en compte l'aspect temporel. Dans un deuxième temps, il s'agit de comparer avec l'application de la méthode temporelle, les techniques de calcul et fonctions temporelles entre elles.

Le protocole d'évaluation pour chaque dataset restera identique. Il n'y a que la technique d'acquisition de données pour construire le profil individuel utilisateur et le profil social qui varie.

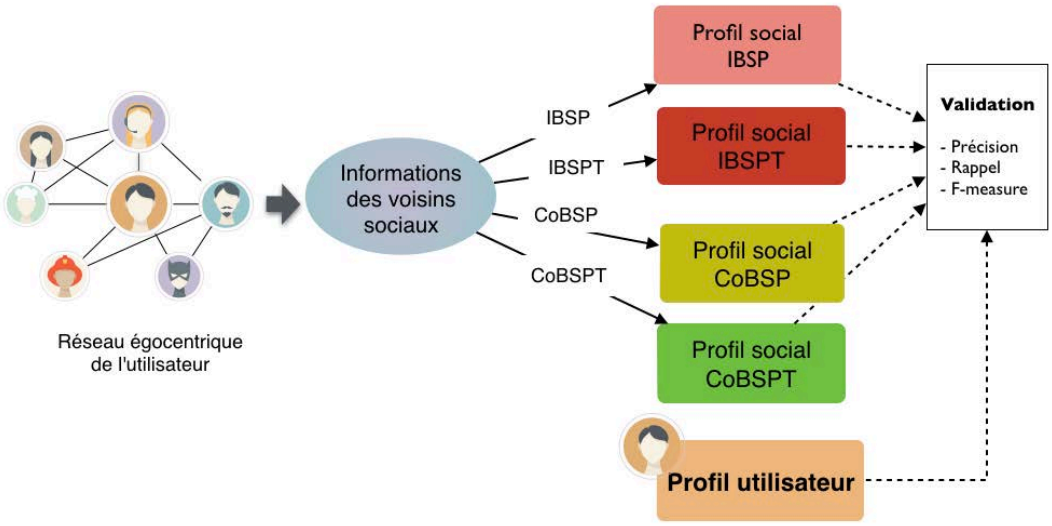


Figure 5.1 Protocole d'évaluation

De plus, comme indiqué dans la section 4.4.2.2, nous évaluerons les résultats selon :

- la taille du réseau égocentrique,
- la densité du réseau égocentrique.

La sous-section suivante présente les mesures retenues pour les évaluations.

5.2.2. Evaluation

Après l'étape de construction des profils, nous obtenons différents profils sociaux construits par les algorithmes *IBSP*, *IBSPT*, *CoBSP* et *CoBSPT*. Les profils sociaux construits seront comparés au profil individuel de l'utilisateur via les mesures de précision, de rappel et de F-mesure.

Dans notre contexte d'évaluation, la précision d'un algorithme de dérivation du profil social est évaluée par le nombre d'intérêts calculés ou « prédits » (profil social) qui sont présents aussi dans le profil utilisateur individuel par rapport au nombre total des intérêt calculés dans le profil social (formule (5.1)).

$$Précision = \frac{nb(intérêts\ dans\ le\ profil\ social\ \cap\ intérêts\ dans\ le\ profil\ utilisateur)}{nb(intérêts\ dans\ le\ profil\ social)} \quad (5.1)$$

Le rappel d'un algorithme de dérivation du profil social quant à lui, est évalué par le nombre d'intérêts prédits (profil social) qui sont présents aussi dans le profil utilisateur individuel par rapport au nombre d'intérêts du profil utilisateur individuel (formule (5.2)).

$$Rappel = \frac{nb(intérêts\ dans\ le\ profil\ social\ \cap\ intérêts\ dans\ le\ profil\ utilisateur)}{nb(intérêts\ dansle\ profil\ utilisateur)} \quad (5.2)$$

La F-mesure (*F-measure* ou *F-score* en anglais) est une mesure populaire qui combine la précision et le rappel par leur moyenne harmonique. Elle est également connue sous le nom de mesure F1 du fait qu'avec cette mesure, la précision et le rappel sont pondérés de façon égale. Le calcul du F-mesure est présenté dans la formule (5.3).

$$F - mesure = 2 * \frac{précision * rappel}{précision + rappel}$$

(5.3)

Pour calculer la précision, le rappel et la F-mesure, nous nous intéressons uniquement aux intérêts les plus pertinents renvoyés par chaque algorithme de dérivation du profil social. En effet, les profils sociaux construits sont de grande taille et ne seraient pas exploitables en l'état. Cela aurait pour effet potentiel d'écrouler les valeurs de précision de façon artificielle. Et dans le même temps, un profil constitué de 1000 intérêts par exemple est peu exploitable dans la pratique. Si le profil utilisateur est constitué de *n* intérêts (vecteur de taille *n*), la précision et le rappel de chaque algorithme de dérivation du profil social seront calculés à partir du top *n+m* (avec m>0) premiers intérêts (vecteur de taille *n+m*) présents dans le profil social. Le facteur m est ici utilisé pour prendre en compte la variation introduite par les algorithmes de calculs et permettre d'évaluer leur capacité à retrouver des éléments pertinents (rappel).

La sous-section suivante détaille l'étude paramétrique qui sera menée pour chaque expérimentation.

5.2.3. Etudes paramétriques

Nous avons effectué une étude paramétrique pour trouver la meilleure combinaison des valeurs γ, λ et α présentées dans le processus de construction du profil social temporel mais aussi pour trouver l'influence de chaque paramètre dans le processus de construction de profils sociaux. Les principes et objectifs de cette étude sont détaillés en section 4.4.

Nous rappelons que :

- γ, présenté dans la formule (4.17) section 4.3.4, représente la proportion entre le poids temporel des individus et le poids temporel des informations dans la phase de calcul du poids temporel d'un élément (intérêt).
- λ représente le taux de dépréciation (Time Decay Rate) des relations et des informations présenté dans la formule (2.7) section 2.3.1.2.
- α représente la proportion entre le poids structurel et le poids sémantique présenté dans la formule (4.21) et dans la formule (4.28) de la section 4.3.5.

Pour ce faire nous avons fait varier les paramètres γ, λ et α de la façon suivante :

- Les valeurs de γ et α sont fixées entre 0 et 1, (γ, α) ∈{0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95, 1.0} pour tous les profils sociaux temporels.
- Les valeurs de λ sont fixées entre 0 et 1, λ ∈ {0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95, 1.0} pour les profils sociaux temporels construits avec les techniques dont le poids temporel est calculé à partir des fonctions temporelles exponentielle et polynomiale.

Ce protocole permet de construire non seulement les profils sociaux pour l'approche proposée avec différentes combinaisons de paramètres mais aussi pour l'approche existante en fixant la

valeur de $\gamma = 0.0$ (poids d'information seul) et la valeur de $\lambda = 0.0$ avec les fonctions exponentielles et polynomiales (effet du facteur temps neutralisé).

Les expérimentations décrites dans la section suivante mettent en œuvre ce protocole dévaluation dans deux réseaux sociaux différents : DBLP puis Twitter.

5.3. Expérimentations

Pour chacun des deux réseaux sociaux étudiés, nous présentons ses caractéristiques, la technique utilisée pour acquérir les données puis la construction du profil utilisateur et du profil social ainsi que les résultats obtenus.

5.3.1. Expérimentations sur DBLP

Le premier domaine d'expérimentation choisi concerne les réseaux d'auteurs de publications scientifiques pour lesquels on peut calculer les intérêts à partir des titres de leurs publications. La nature du lien de co-auteurs entre deux auteurs sous-entend que ces deux derniers se connaissent à priori dans la vie réelle.

Par rapport à la classification des réseaux sociaux donnée dans l'état de l'art (cf. section 3.1), un réseau de co-auteurs ne présente pas, a priori, les caractéristiques des réseaux sociaux numériques comme les autres réseaux qui sont majoritairement issus des médias sociaux (Facebook, LinkedIn, Twitter, ...). Nous considérons plutôt ce type de réseau comme un réseau social traditionnel dans lequel les liens entre les nœuds sont basés sur leurs connaissances réelles. Les auteurs publient ensemble parce qu'ils se connaissent, partagent et échangent des informations sur le même sujet de recherche. On peut donc tout de même rapprocher les réseaux de co-auteurs des sites de réseautage social puisqu'ils rassemblent et relient des personnes qui se connaissent dans la vie réelle.

Il est absolument nécessaire de ne pas construire les profils sociaux et individuels à partir de données identiques (les publications d'un auteur peuvent également être des publications de ses co-auteurs) car cela induirait un biais dans l'expérimentation réalisée. Construire les deux profils à partir du dataset DBLP seul n'est donc pas envisageable. Nous avons choisi d'utiliser deux sources distinctes pour construire les deux profils : DBLP[64] pour le profil social et Mendeley[65] pour le profil utilisateur.

DBLP : nous utilisons les données de DBLP, une librairie digitale qui référence de nombreux articles scientifiques publiés dans le domaine informatique dans le monde. Cette librairie perme d'interroger la liste des publications d'un auteur. Nous trouvons dans la Figure 5.2 l'exemple d'interface web DBLP qui présente la liste des publications d'un auteur. Cette librairie présente aussi le réseau de co-auteurs des publications scientifiques pour un auteur par la liste de ses co-auteurs. Nous exploitons ce réseau pour construire le réseau égocentrique de chaque auteur (contenant ses co-auteurs) et dériver son profil social.

[64] dblp.uni-trier.de

[65] www. mendeley.com

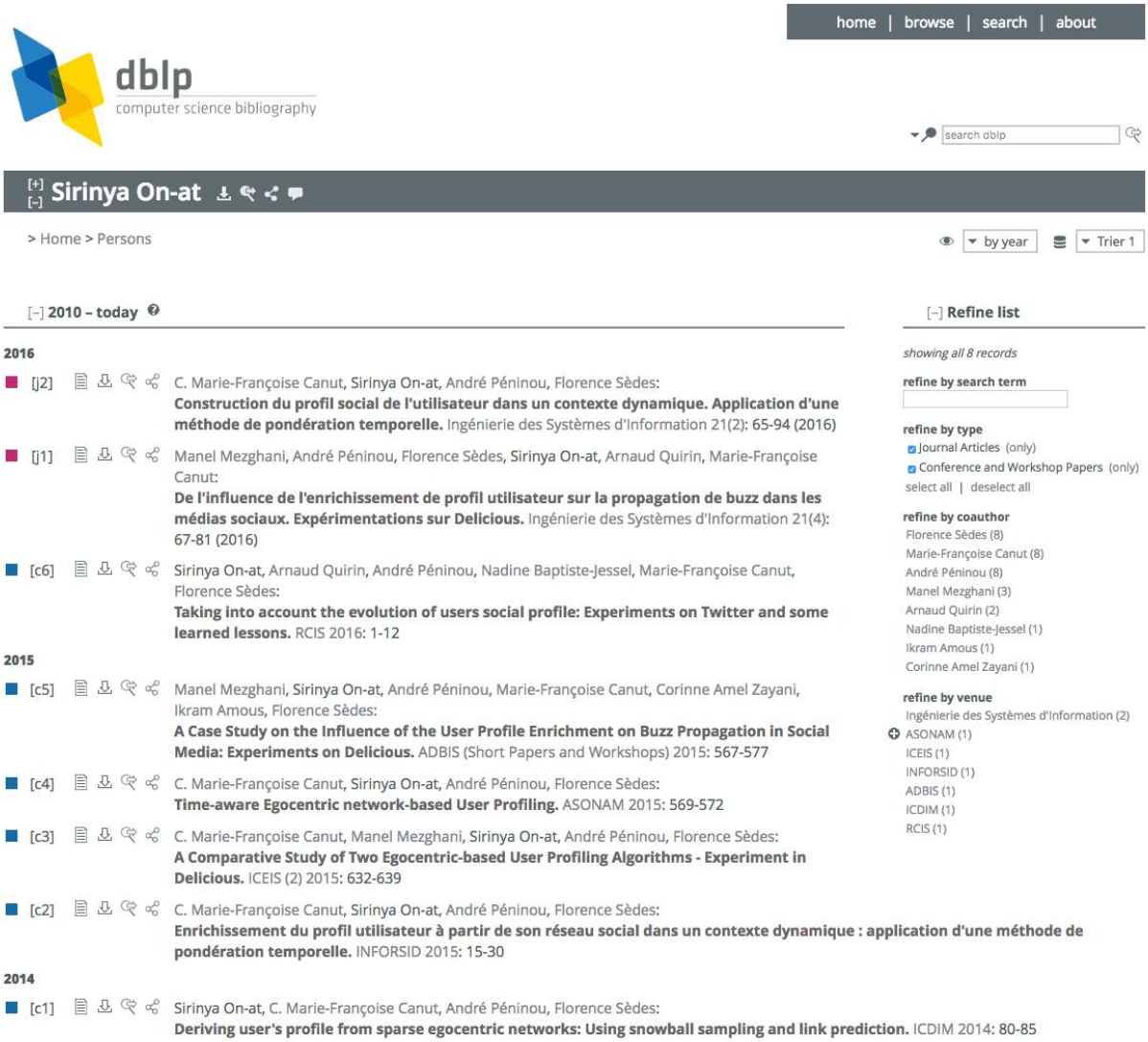


Figure 5.2 Page web DBLP présentant la liste des publications de l'auteur Sirinya ON-AT

MENDELEY : c'est un réseau social d'auteurs d'articles scientifiques de plus en plus utilisé. Nous nous intéressons à ce réseau car les auteurs peuvent indiquer explicitement la liste de leurs intérêts dans leur profil. Ces informations représentent donc de façon pertinente leur profil réel (et peuvent être utilisées pour construire leur profil utilisateur). Nous trouvons dans Figure 5.3 l'exemple d'un profil Mendeley d'un auteur scientifique.

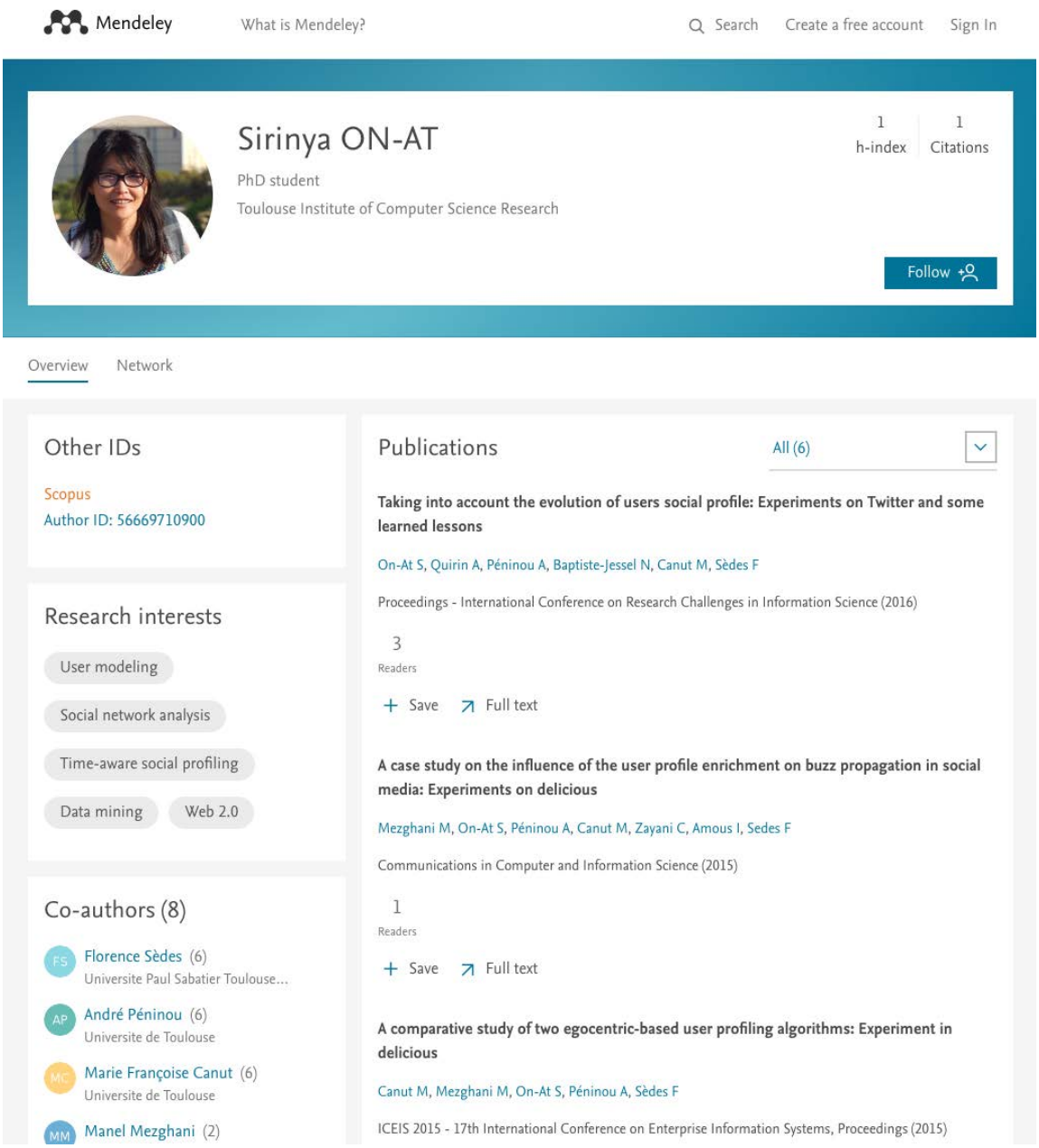


Figure 5.3 Exemple du profil sur le site Mendeley.com de l'auteur Sirinya ON-AT

5.3.1.1. Présentation du réseau social DBLP

Le réseau DBLP est donc un réseau de publications scientifiques. En se basant sur les caractéristiques des réseaux sociaux présentées dans la section 3.1.2, nous considérons le réseau DBLP comme un réseau biparti (les nœuds sont les auteurs et les publications) où les individus auront une connexion entre eux s'ils ont publié ensemble au moins une fois. Le réseau égocentrique de chaque utilisateur (auteur) est extrait en transformant le réseau biparti en réseau uniparti (cf. la Figure 5.4) en prenant en compte leurs co-auteurs et les titres de leurs publications ainsi que leur date de publication.

Nœuds : auteurs de publication(s) scientifique(s).

Relations : deux nœuds (auteurs) sont connectés entre eux s'ils ont publié au moins une fois ensemble.

Informations : titres des publications des auteurs dont on extrait les mots-clés.

Interactions : fait de publier ensemble. La date d'interaction est donc la date de publication.

Granularité de temps : année.

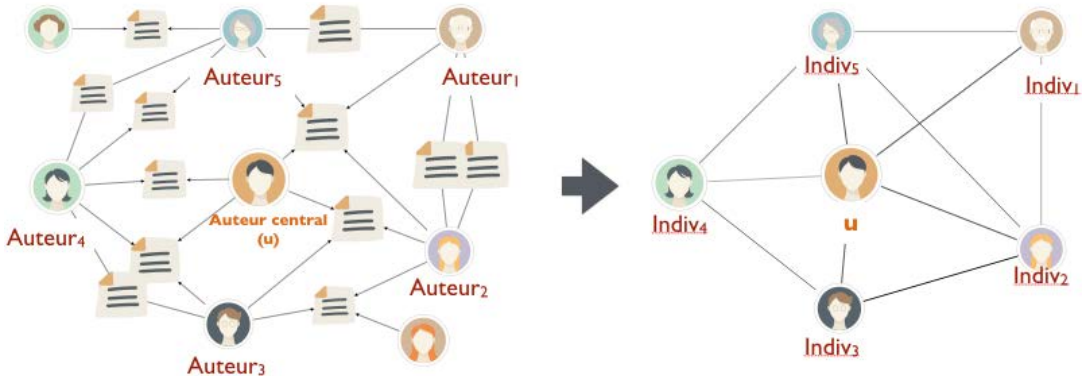


Figure 5.4 Présentation de réseau DBLP et extraction du réseau égocentrique de l'utilisateur

5.3.1.2. Accès aux donnés et présentation du dataset

Pour obtenir des profils sociaux et profils utilisateurs qui soient comparables, l'échantillon de test doit comporter les auteurs de la base DBLP qui possèdent en même temps un profil dans la base Mendeley. La question qui se pose à cette étape est : comment retrouvons-nous un même auteur dans les bases DBLP et Mendeley ?

Mendeley et DBLP disposent toutes les deux d'une API permettant d'accéder aux données : dans DBLP on peut récupérer les auteurs, les articles publiés ainsi que la liste de leurs co-auteurs, dans Mendeley on peut récupérer la liste des auteurs scientifiques inscrits, la liste de leurs publications (si elles sont renseignées), la liste de leurs contacts sur le réseau, les informations liées à leur domaine de recherche ainsi que la liste de leurs intérêts (si elle a été renseignée). Nous utilisons le champ commun de ces deux bases qui est le nom de l'auteur, pour croiser et extraire les données identiques de ces deux bases.

a. Accès aux données dans Mendeley

Nous avons dans un premier temps, récupéré les auteurs dans le site Mendeley qui ont des critères correspondant à notre cas d'étude. En effet, les auteurs doivent exister dans la base DBLP et posséder dans Mendeley, un nombre important d'intérêts pour qu'on puisse extraire des informations assez significatives afin d'avoir un profil utilisateur assez riche. Nous avons choisi dans cette évaluation des auteurs de Mendeley qui possèdent au moins 5 intérêts indiqués explicitement dans leur profil. Nous trouvons dans la Figure 5.3, l'exemple des intérêts indiqués explicitement par l'utilisateur Sirinya ON-AT dans son profil.

b. Accès aux données dans DBLP

Les informations de la base DBLP telles que la liste de co-auteurs, la liste de publications ainsi que les détails de ces dernières sont accessibles sous forme de fichier XML et peuvent être extraites par l'API DBLP (Ley, 2009). Cette API permet d'extraire uniquement certaines parties des données. Dans notre cas, nous nous intéressons à la liste des publications et à la liste de co-auteurs, comme présenté sur la Figure 5.5. Ensuite, le fichier XML renvoyé par l'API peut être traité avec un parseur tel que SAX pour extraire les données (Ley, 2009). En utilisant cette

l'API, nous partons des auteurs extraits depuis Mendeley pour extraire leur réseau égocentrique depuis leurs données dans DBLP.

(a)

```
▼<coauthors author="Sirinya On-at" urlpt="o/On=at:Sirinya">
   <author urlpt="a/Amous:Ikram" count="1">Ikram Amous</author>
   <author urlpt="b/Baptiste=Jessel:Nadine" count="1">Nadine Baptiste-Jessel</author>
   <author urlpt="c/Canut:Marie=Fran=ccedil=oise" count="8">Marie-Françoise Canut</author>
   <author urlpt="m/Mezghani:Manel" count="3">Manel Mezghani</author>
   <author urlpt="p/P=eacute=ninou:Andr=eacute=" count="8">André Péninou</author>
   <author urlpt="q/Quirin:Arnaud" count="2">Arnaud Quirin</author>
   <author urlpt="s/S=egrave=des:Florence" count="8">Florence Sèdes</author>
   <author urlpt="z/Zayani:Corinne_Amel" count="1">Corinne Amel Zayani</author>
 </coauthors>
```

(b)

```
▼<dblp>
 ▼<inproceedings key="conf/asunam/CanutOPS15" mdate="2015-11-21">
    <author>C. Marie-Françoise Canut</author>
    <author>Sirinya On-at</author>
    <author>André Péninou</author>
    <author>Florence Sèdes</author>
   ▼<title>
      Time-aware Egocentric network-based User Profiling.
    </title>
    <pages>569-572</pages>
    <year>2015</year>
    <booktitle>ASONAM</booktitle>
    <ee>http://doi.acm.org/10.1145/2808797.2809415</ee>
    <crossref>conf/asunam/2015</crossref>
    <url>db/conf/asunam/asonam2015.html#CanutOPS15</url>
  </inproceedings>
</dblp>
```

Figure 5.5 (a) Liste de co-auteurs de l'auteur Sirinya ON-AT. (b) Exemple de description d'un article publié par un co-auteur de Sirinya ON-AT.

Notre échantillon de test contient 236 utilisateurs ayant entre 5 et 495 co-auteurs. La moyenne du nombre de co-auteurs de tous les utilisateurs (taille du réseau) est de 70. Au total, nous avons récupéré 10105 auteurs et 522132 publications datées entre 1959 et 2017.

5.3.1.3. Evaluation

La Figure 5.6 présente le protocole d'évaluation (cf. Figure 5.1) instancié dans le contexte de DBLP.

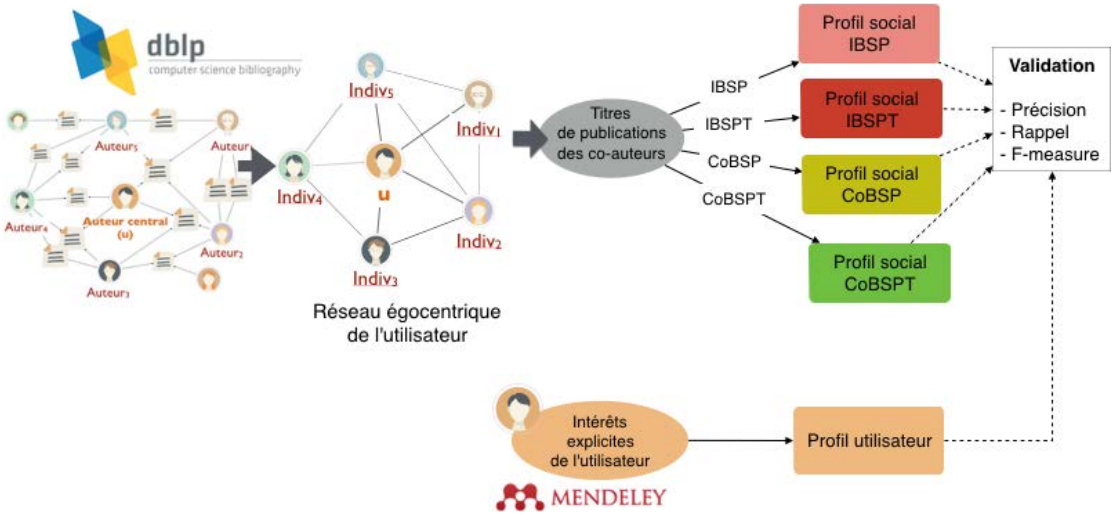


Figure 5.6 Protocole d'évaluation sur le réseau DBLP

Nous présentons dans ce qui suit, la construction du profil utilisateur et la construction du profil social dans cette évaluation.

a. Vérité de terrain : extraction du profil explicite de l'utilisateur à partir de Mendeley

Pour construire le profil utilisateur, nous utilisons les intérêts indiqués explicitement par l'utilisateur dans son profil Mendeley. En appliquant le même traitement que celui de la construction du profil social, nous utilisons les dictionnaires et les filtres pour extraire des termes significatifs qui seront considérés comme les intérêts dans le profil utilisateur individuel. A la différence du profil social, la pondération de chaque mot est définie à 1 car généralement les termes indiqués explicitement dans les profils Mendeley ne sont ni ordonnés ni pondérés. Néanmoins, comme la totalité des mots clefs du profil individuel sera utilisée, cela ne perturbera pas les calculs de précision et rappel.

b. Construction du profil social : étude paramétrique

Les profils sociaux seront construits à partir des différents processus proposés dans notre contribution. Toutefois, comme indiqué dans la méthodologie de construction du profil utilisateur classique (section 2.2.2.1), il est nécessaire d'appliquer différents traitements sur les données extraites de DBLP. Nous récupérons d'abord les co-auteurs de l'utilisateur. Nous stockons ainsi la liste des publications de tous les auteurs qui font partie de son réseau égocentrique. La deuxième étape consiste à extraire les communautés à partir de ce réseau. Nous analysons dans la troisième étape, les titres des publications pour en extraire les termes les plus significatifs en utilisant un dictionnaire de synonymes et de filtres. Ces synonymes sont ensuite considérés comme des occurrences de ce mot. Seuls les mots retenus seront considérés comme des intérêts du-co-auteur. Enfin, nous représentons également le profil social par un vecteur de termes (intérêts) pondérés. Nous avons effectué une étude paramétrique pour trouver la meilleure combinaison des valeurs α, γ et λ comme indiqué dans la section 5.2.3.

L'expérimentation a été programmée en Java en utilisant les bibliothèques SAX[66], JUNG[67] et LibInterface (pour l'algorithme iLCD)[68]. L'ensemble des traitements représente 19 000 lignes de codes. Les expérimentations ont été menées sur la plateforme OSIRIM[69], mise en place au sein de notre laboratoire, en utilisant 64 cœurs, 128 GO de RAM. Les données correspondent à 265 MO d'espace disque.

5.3.1.4. Résultats

Dans cette section nous présentons les résultats de nos évaluations. Notons que pour calculer la précision et le rappel, nous nous intéressons uniquement aux intérêts les plus pertinents renvoyés par chaque algorithme de dérivation du profil social (top n premiers intérêts ordonnés en fonction de leur poids). Les résultats présentés dans cette section sont calculés en considérant le top 5 des intérêts. Notons qu'avec le top10 (et plus) des intérêts, les résultats en termes de précision peuvent être moins bons à cause de l'augmentation de la taille du profil social. Pour les utilisateurs qui ont moins de 10 intérêts dans leur profil explicite (Mendeley), les résultats en termes de précision peuvent ne pas être significatifs en considérant le top 10 (et plus) des intérêts. En effet, la taille du profil social augmente mais le nombre maximum des intérêts

[66] sax.sourceforge.net – parseur de données XML

[67] jung.sourceforge.net – manipulation de données sous forme de graphes

[68] cazabetremy.fr/iLCD

[69] osirim.irit.fr

pertinents trouvés reste toujours égal au nombre des intérêts trouvés dans le profil social. Les résultats pour les top 10 et 15 sont présentés respectivement en Annexe 1-a et Annexe 1-b.

a. Notation

Afin de désigner les profils construits à partir, d'une part, des différents techniques de calcul du poids temporel des intérêts et, d'autre part, des différentes fonctions temporelles, nous considérons les notations suivantes avec : $\boldsymbol{Processus}_{\boldsymbol{FoncTemporelle,\ TechniqueCalculPoidsTemporel,}}$ où :

- *Processus* désigne le processus de construction utilisé selon les deux grandes approches décrites précédemment :
    - *Approche basée sur les individus*
        - IBSP désigne le processus de construction du profil basé sur les individus sans prise en compte du temps,
        - IBSPT désigne le processus de construction du profil basé sur les individus avec prise en compte de la méthode temporelle proposée.
    - *Approche basée sur les communautés*
        - CoBSP désigne le processus de construction du profil basé sur les communautés sans prise en compte du temps,
        - CoBSPT désigne le processus de construction du profil basé sur les communautés avec prise en compte de la méthode temporelle proposée.
- *FoncTemporelle* désigne la fonction temporelle appliquée (cf. section 2.3.1.2)
    - *Exp* pour la fonction temporelle exponentielle,
    - *Poly* pour la fonction temporelle polynomiale,
    - *Lin* pour la fonction temporelle linéaire inverse.
- *TechniqueCalculPoidsTemporel* désigne la technique de calcul du poids temporel des intérêts appliquée
    - *AATemp* désigne la technique appliquant la mesure de la prédiction de Adamic/Adar Temporelle (formule (4.6)) pour calculer le poids temporel des individus et celui des informations (formule (4.14)).
    - *AAStrTemp* désigne la technique appliquant la mesure de la prédiction de liens AdamicAdar en intégrant la date de la dernière interaction pour calculer le poids temporel des individus (formule (4.8)) et celui des informations (formule (4.14)).
    - *SITemp* désigne la technique appliquant la somme temporelle des interactions pour calculer le poids temporel des individus (formule (4.9)) et le calcul du poids temporel des informations (formule (4.14)).

Par exemple, $IBSPT_{Exp,AATemp}$ représente les profils construits à partir du processus basé sur les individus en appliquant la fonction temporelle exponentielle sur la technique de calcul du poids temporel. Cette dernière se base sur la mesure *AdamicAdar Temporelle* (formule (4.6)) pour calculer le poids temporel des individus et en appliquant le calcul du poids temporel des informations (formule (4.14)).

b. Résultats globaux

Nous présentons d'abord les résultats globaux de l'étude paramétrique pour les 236 utilisateurs en termes de précision, rappel et F-mesure. Les résultats sont calculés pour chaque combinaison de paramètres (α, λ et γ) et sont présentés par la précision moyenne (resp. rappel moyen et F-mesure moyenne) pour l'échantillon c'est-à-dire par la moyenne de précision (resp. rappel, F-mesure) de tous les utilisateurs dans l'échantillon (donc pour une combinaison des paramètres (α, λ et γ) donnée).

- Résultats sans la prise en compte du poids de centralité (en fixant α = 0)

Pour les 3 paramètres γ, λ et α il s'agit d'étudier leurs valeurs optimales dans le processus de construction du profil social.

Pour étudier l'impact de la prise en compte du temps sans prendre en compte les autres facteurs, nous avons décidé d'étudier, dans un premier temps, les résultats des processus de construction du profil social en fixant la valeur α = 0.0 ; c'est-à-dire sans prendre en compte le poids structurel (centralité de degré des communautés ou individus dans le cas des processus basés sur les communautés ou sur les individus) (cf. section 4.3.5.1 et 4.3.5.2). Sachant que α est le paramètre déjà défini dans le processus existant qui ne prend pas en compte le temps (cf. section 4.4.1).

La Figure 5.7 ci-après présente la comparaison des résultats, en termes de précision, rappel et F-mesure, des profils sociaux construits par les différentes techniques de calcul du poids temporel pour l'approche de construction de profil social basée sur les individus (processus IBSP et IBSPT) et pour celle basée sur les communautés (processus CoBSP et CoBSPT). Les résultats de chaque technique sont présentés par la précision moyenne (resp. rappel moyen et F-mesure moyenne) correspondant à la meilleure combinaison de paramètres (γ et λ) ; la combinaison qui a obtenu la meilleure moyenne.

Dans un premier temps, dans les processus existants, la normalisation des poids des intérêts n'a pas été effectuée de la même façon que dans les processus appliquant la méthode temporelle proposée (cf. section 4.3.4.4). La comparaison des résultats issus de ces différents processus peut donc être biaisée : on ne sait pas si l'amélioration ou la détérioration des résultats est due à la prise en compte des facteurs temporels ou purement à la normalisation. Nous avons donc appliqué la même technique de normalisation des poids aux processus existants. Ceci rend les résultats des processus existants (IBSP et CoBSP) meilleurs que dans le cas où on n'effectue pas la normalisation ; notons que cela était le cas (pas de normalisation) dans les travaux antérieurs de (Tchuente, 2013). Dans la Figure 5.7 nous présentons donc, dans un premier temps, les résultats de l'approche existante avec et sans technique de normalisation pour montrer l'amélioration obtenue par la normalisation dans les processus existants. Dans les sections suivantes, nous présenterons uniquement les résultats des processus proposés par rapport à ceux des processus existants « normalisés » afin de bien mettre en évidence l'impact de la prise en compte du temps dans les processus par rapport aux processus existants.

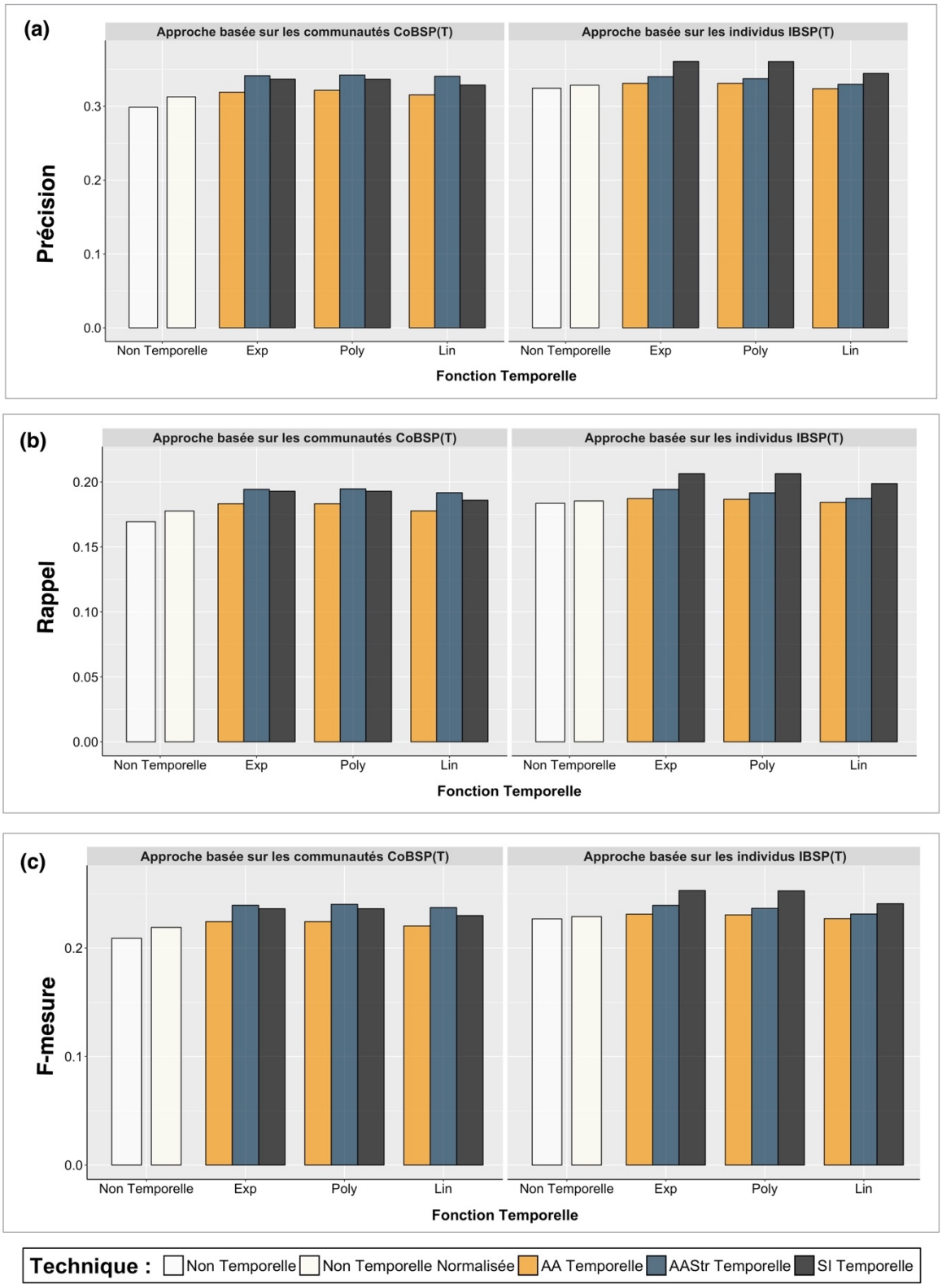


Figure 5.7 Comparaison de la meilleure précision moyenne (a), meilleur rappel moyen (b) et meilleure F-mesure moyenne (c), des résultats des processus de construction du profil social en appliquant les différentes méthodes temporelles aux processus (IBSPT et CoBSPT) avec ceux des processus qui ne prennent pas en compte les critères temporel (IBSP et CoBSP)

Pour **l'approche basée sur les communautés**, les meilleurs résultats en termes de précision, rappel et F-mesure obtenus par le processus CoBSP *(*non prise en compte du temps) sont respectivement 0.29764, 0.16714 et 0.20683. Pour le processus CoBSPT *(prise en compte du temps)*, nous comparons les résultats par rapport à la fonction temporelle utilisée et pour chacune selon la technique de calcul du poids temporel des intérêts.

- Pour les techniques appliquant la fonction temporelle exponentielle (Exp) :
    - la technique ExpAATemp produit la meilleure précision (0.31816) lorsque $\lambda = 0.01$ et $\gamma = 0.95$ avec 6.895 % d'amélioration par rapport au processus CoBSP. Le meilleur rappel (0.17819) est obtenu lorsque $\lambda = 0$ et $\gamma = 0.95$ avec 6.613 % d'amélioration par rapport au processus CoBSP. La meilleure F-mesure (0.22055) est obtenue lorsque $\lambda = 0.01$ et $\gamma = 0.95$ avec 6.629 % d'amélioration par rapport au processus CoBSP.
    - la technique ExpAAStrTemp produit la meilleure précision (0.33632) lorsque $\lambda = 0.1$ et $\gamma = 0.95$ avec 12.996 % d'amélioration par rapport au processus CoBSP. Le meilleur rappel (0.1913) est obtenu lorsque $\lambda = 0$ et $\gamma = 0.95$ avec 14.455 % d'amélioration par rapport au processus CoBSP. La meilleure F-mesure (0.23484) est obtenue lorsque $\lambda = 0.1$ et $\gamma = 0.95$ avec 13.541 % d'amélioration par rapport au processus CoBSP.
    - la technique ExpSITemp produit la meilleure précision (0.33585) lorsque $\lambda = 0.05$ et $\gamma = 0.95$ avec 12.838 % d'amélioration par rapport au processus CoBSP. Le meilleur rappel (0.1885) est obtenu lorsque $\lambda = 0$ et $\gamma = 0.95$ avec 12.783 % d'amélioration par rapport au processus CoBSP. La meilleure F-mesure (0.23341) est obtenue lorsque $\lambda = 0.05$ et $\gamma = 0.95$ avec 12.849 % d'amélioration par rapport au processus CoBSP.
- Pour les techniques appliquant la fonction temporelle polynomiale (Poly) :
    - la technique PolyAATemp produit la meilleure précision (0.31788) lorsque $\lambda \in \{0.25, 0.5\}$ et $\gamma = 0.95$ avec 6.8 % d'amélioration par rapport au processus CoBSP. Le meilleur rappel (0.1782) est obtenu lorsque $\lambda = 0.25$ et $\gamma = 0.95$ avec 6.617 % d'amélioration par rapport au processus CoBSP. La meilleure F-mesure (0.22087) est obtenue lorsque $\lambda = 0.25$ et $\gamma = 0.95$ avec 6.786 % d'amélioration par rapport au processus CoBSP.
    - la technique PolyAAStrTemp produit la meilleure précision (0.33758) lorsque $\lambda = 0.5$ et $\gamma = 0.95$ avec 13.419 % d'amélioration par rapport au processus CoBSP. Le meilleur rappel (0.1913) est obtenu lorsque $\lambda = 0$ et $\gamma = 0.95$ avec 14.455 % d'amélioration par rapport au processus CoBSP. La meilleure F-mesure (0.23563) est obtenue lorsque $\lambda = 0.5$ et $\gamma = 0.95$ avec 13.924 % d'amélioration par rapport au processus CoBSP.
    - la technique PolySITemp produit la meilleure précision (0.33341) lorsque $\lambda \in \{0.001, 0.01\}$ et $\gamma = 0.95$ avec 12.02 % d'amélioration par rapport au processus CoBSP. Le meilleur rappel (0.1885) est obtenu lorsque $\lambda = 0$ et $\gamma = 0.95$ avec 12.783 % d'amélioration par rapport au processus CoBSP. La meilleure F-mesure (0.23228) est obtenue lorsque $\lambda = 0$ et $\gamma = 0.95$ avec 12.302 % d'amélioration par rapport au processus CoBSP.
- Pour les techniques appliquant la fonction temporelle Linéaire (Lin) :
    - la technique LinAATemp produit la meilleure précision (0.3032) lorsque $\gamma = 0.95$ avec 1.867 % d'amélioration par rapport au processus CoBSP. Le

meilleur rappel (0.16787) est obtenu lorsque γ = 0.95 avec 0.436 % d'amélioration par rapport au processus CoBSP. La meilleure F-mesure (0.20893) est obtenu lorsque γ = 0.95 avec 1.014 % d'amélioration par rapport au processus CoBSP.

- la technique LinAAStrTemp produit la meilleure précision (0.34054) lorsque γ = 0.95 avec 14.413 % d'amélioration par rapport au processus CoBSP. Le meilleur rappel (0.19179) est obtenu lorsque γ = 0.95 avec 14.746 % d'amélioration par rapport au processus CoBSP. La meilleure F-mesure (0.23728) est obtenue lorsque γ = 0.95 avec 14.721 % d'amélioration par rapport au processus CoBSP.
- la technique LinSITemp produit la meilleure précision (0.31988) lorsque γ = 0.95 avec 7.473 % d'amélioration par rapport au processus CoBSP. Le meilleur rappel (0.17996) est obtenu lorsque γ = 0.95 avec 7.671 % d'amélioration par rapport au processus CoBSP. La meilleure F-mesure (0.22285) est obtenue lorsque γ = 0.95 avec 7.745 % d'amélioration par rapport au processus CoBSP.

Pour l'**approche basée sur les individus**, les meilleurs résultats en termes de précision, rappel et F-mesure obtenus par le processus IBSP (non prise en compte du temps), sont respectivement 0.32587, 0.18406 et 0.22722. Pour le processus IBSPT (prise en compte du temps), nous comparons les résultats par rapport à la fonction temporelle utilisée et pour chacune selon la technique de calcul du poids temporel des intérêts.

- Pour les techniques appliquant la fonction temporelle Exponentielle (Exp) :
  - la technique ExpAATemp produit la meilleure précision (0.33054) lorsque λ = 0 et γ = 0.05 avec 1.43 % d'amélioration par rapport au processus IBSP. Le meilleur rappel (0.18645) est obtenu lorsque λ = 0.05 et γ = 0.1 avec 1.301 % d'amélioration par rapport au processus IBSP. La meilleure F-mesure (0.23017) est obtenue lorsque λ = 0.05 et γ = 0.1 avec 1.297 % d'amélioration par rapport au processus IBSP.
  - la technique ExpAAStrTemp produit la meilleure précision (0.34001) lorsque λ = 0.1 et γ = 0.75 avec 4.338 % d'amélioration par rapport au processus IBSP. Le meilleur rappel (0.19422) est obtenu lorsque λ = 0.1 et γ = 0.75 avec 5.519 % d'amélioration par rapport au processus IBSP. La meilleure F-mesure (0.2392) est obtenue lorsque λ = 0.1 et γ = 0.75 avec 5.271 % d'amélioration par rapport au processus IBSP.
  - la technique ExpSITemp produit la meilleure précision (0.35993) lorsque λ = 0.01 et γ = 0.95 avec 10.45 % d'amélioration par rapport au processus IBSP. Le meilleur rappel (0.20421) est obtenu lorsque λ = 0.01 et γ = 0.95 avec 10.95 % d'amélioration par rapport au processus IBSP. La meilleure f-mesure (0.25238) est obtenu lorsque λ = 0.01 et γ = 0.95 avec 11.071 % d'amélioration par rapport au processus IBSP.
- Pour les techniques appliquant la fonction temporelle polynomiale (Poly) :
  - la technique PolyAATemp produit la meilleure précision (0.33096) lorsque λ = 0.001 et γ = 0.05 avec 1.56 % d'amélioration par rapport au processus IBSP. Le meilleur rappel (0.18602) est obtenu lorsque $\lambda \in \{0.0, 0.001\}$ et γ = 0.05 avec 1.067 % d'amélioration par rapport au processus IBSP. La

meilleure F-mesure (0.23012) est obtenue lorsque $\lambda = 0.001$ et $\gamma = 0.05$ avec 1.277 % d'amélioration par rapport au processus IBSP.

- la technique PolyAAStrTemp produit la meilleure précision (0.33733) lorsque $\lambda \in \{0.75, 0.95\}$ et $\gamma = 0.75$ avec 3.515 % d'amélioration par rapport au processus IBSP. Le meilleur rappel (0.19169) est obtenu lorsque $\lambda \in \{0.75, 0.95\}$ et $\gamma = 0.75$ avec 4.148 % d'amélioration par rapport au processus IBSP. La meilleure F-mesure (0.23661) est obtenue lorsque $\lambda = 0.95$ et $\gamma = 0.75$ avec 4.131 % d'amélioration par rapport au processus IBSP.
- la technique PolySITemp produit la meilleure précision (0.35964) lorsque $\lambda = 0.05$ et $\gamma = 0.95$ avec 10.363 % d'amélioration par rapport au processus IBSP. Le meilleur rappel (0.20403) est obtenu lorsque $\lambda = 0.1$ et $\gamma = 0.95$ avec 10.851 % d'amélioration par rapport au processus IBSP. La meilleure F-mesure (0.25211) est obtenu lorsque $\lambda = 0.1$ et $\gamma = 0.95$ avec 10.954 % d'amélioration par rapport au processus IBSP.

- Pour les techniques appliquant la fonction temporelle linéaire inverse (Lin) :
  - la technique LinAATemp produit la meilleure précision (0.31954) lorsque $\gamma = 0.75$ avec une perde de 1.943 % par rapport au processus IBSP. Le meilleur rappel (0.18284) est obtenu lorsque $\gamma = 0.75$ avec une perte de 0.666 % par rapport au processus IBSP. La meilleure F-mesure (0.22529) est obtenue lorsque $\gamma = 0.75$ avec une perte -0.85 % par rapport au processus IBSP.
  - la technique LinAAStrTemp produit la meilleure précision (0.32975) lorsque $\gamma = 0.75$ avec 1.189 % d'amélioration par rapport au processus IBSP. Le meilleur rappel (0.18742) est obtenu lorsque $\gamma = 0.75$ avec 1.828 % d'amélioration par rapport au processus IBSP. La meilleure F-mesure (0.23136) est obtenue lorsque $\gamma = 0.75$ avec 1.822 % d'amélioration par rapport au processus IBSP.
  - la technique LinSITemp produit la meilleure précision (0.34439) lorsque $\gamma = 0.95$ avec 5.682 % d'amélioration par rapport au processus IBSP. Le meilleur rappel (0.19487) est obtenu lorsque $\gamma = 0.95$ avec 5.875 % d'amélioration par rapport au processus IBSP. La meilleure F-mesure (0.2409) est obtenu lorsque $\gamma = 0.95$ avec 6.023 % d'amélioration par rapport au processus IBSP.

Nous remarquons que pour les deux approches de construction du profil social (basée sur les individus ou sur les communautés), en appliquant les fonctions temporelles exponentielle et polynomiale, la méthode temporelle proposée produit de meilleurs résultats que les processus qui ne prennent pas en compte des critères temporels (approche existante), quelle que soit la technique de calcul du poids temporel des intérêts utilisée. Avec la fonction temporelle linéaire inverse, la méthode temporelle produit de meilleurs résultats sauf dans le cas de la technique AdamicAdar Temporelle (AATemp) pour l'approche basée sur les individus, où on obtient une perte de 0.85%.

❖ Résultats en fonction de γ et λ pour l'approche basée individus

Pour chacune des fonctions temporelles utilisées (exponentielle, polynomiale et linéaire), la Figure 5.8 présente les résultats obtenus en fonction de γ (axe horizontal) en termes de précision (axe vertical) pour l'approche de construction du profil social basée sur les individus IBSPT. Les différentes courbes représentent différentes valeurs de λ.

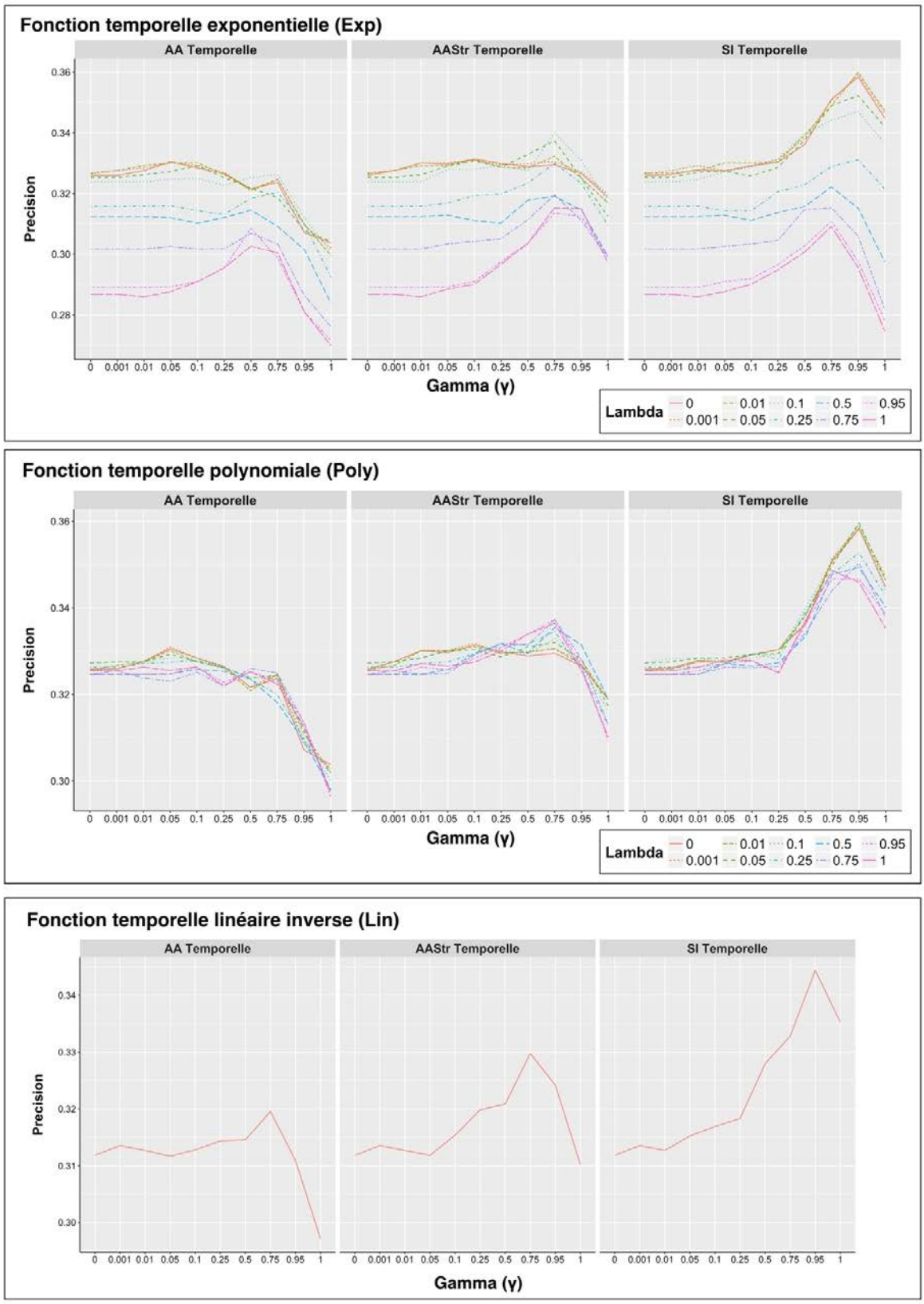


Figure 5.8 Comparaison en termes de précision des résultats en fonction de γ en axe horizontal et λ (différentes courbes) du processus basé sur les individus (IBSPT), pour les techniques utilisant les trois fonctions temporelles exponentielle, polynomiale et linéaire.

Nous remarquons qu'avec la fonction linéaire inverse qui ne dépend pas de la valeur de λ, les résultats ont tendance à monter quand γ devient de plus en plus important. Les meilleurs résultats sont obtenus quand γ est fixé assez haut (0.75 pour la technique AATemporelle et AAStrTemporelle et 0.95 pour la technique SITemp). En deçà, les résultats sont moins bons voire très mauvais.

Pour les fonctions exponentielle et polynomiale, les résultats varient de façon différente selon la technique appliquée et la valeur de λ.

Pour les techniques appliquant la fonction temporelle exponentielle, nous observons que les résultats changent en fonction de γ et λ. La valeur optimale de λ change en fonction de la technique appliquée et de la valeur de γ fixée et vice-versa. Nous observons globalement pour les trois techniques qu'avec la valeur de λ fixée très bas (entre 0 et 0.1), les résultats sont clairement et globalement meilleurs par rapport à ceux obtenus en fixant λ plus haut (près de 1 par exemple). Les meilleures valeurs sont trouvées quand λ est fixé entre 0 et 0.1.

- Pour la technique ExpAATemp, les meilleurs résultats pour λ fixé très bas (entre 0 et 0.05) sont observés quand γ est très bas. Les meilleurs résultats pour λ fixé entre 0.01 et 0.25 sont trouvés quand γ est fixé à 0.75. Pour les courbes de λ fixé entre 0.5 et 1, les meilleures valeurs sont retrouvées quand γ = 0.5. Nous observons que pour les courbes de λ fixé très haut (entre 0.95 et 1), les résultats ont tendance à monter quand γ devient de plus en plus important et rebaissent quand γ dépasse 0.5. En comparant les résultats des courbes obtenues avec λ fixée entre 0 et 0.1, avec lesquelles on obtient de meilleurs résultats, nous remarquons qu'il n'y a pas beaucoup d'écart entre ces résultats (λ = 0, 0.01, 0.1).
- Pour la technique ExpAAStrTemp, pour λ fixé très bas (entre 0 et 0.01), nous ne voyons beaucoup de changement de résultats en fonction de γ. Au-delà, les résultats ont tendance à monter quand γ monte. Les meilleures valeurs sont trouvées quand γ = 0.75 et λ = 0.1. Pour les courbes avec λ fixé très haut (entre 0.95 et 1), les résultats ont tendance à monter quand γ devient de plus en plus important jusqu'à 0.75 et diminuent au-delà.
- Pour la technique ExpSITemp, quelle que soit la valeur de λ, les résultats ont tendance à monter quand γ devient important jusqu'à 0.75, qui est la valeur optimale, et diminuent au-delà.

Pour les techniques appliquant la fonction temporelle polynomiale, les résultats changent également en fonction de valeur de λ mais pas de la même façon qu'avec la fonction temporelle exponentielle. En effet, en appliquant cette fonction temporelle, nous ne voyons pas clairement une grande différence des résultats en fonction de valeur de λ.

- Pour la technique PolyAATemp, les meilleurs résultats sont trouvés quand λ est fixé bas et avec γ fixé très bas. Les résultats ont tendance à baisser quand γ devient plus important.
- Pour la technique PolyAAStrTemp, les meilleurs résultats sont obtenus quand λ est bas et γ fixé haut. Quand λ est fixé très bas (entre 0 et 0.01), nous ne voyons pas beaucoup de changement des résultats en fonction de γ. Au-delà, les résultats ont tendance à monter quand γ devient plus important jusqu'à la valeur optimale 0.75, puis diminuent.
- Pour la technique PolySITemp, les meilleurs résultats sont trouvés quand λ est fixé bas avec γ fixé très haut. Les résultats ont tendance à monter quand γ devient important jusqu'à 0.75, qui est la valeur optimale, et diminuent au-delà.

Nous observons globalement que, quand γ = 1 les résultats diminuent dramatiquement quelle que soit la technique appliquée et la valeur de λ.

- Résultats en fonction de γ et λ pour l'approche basée communautés

Pour chacune des fonctions temporelles utilisées (exponentielle, polynomiale et linéaire), la Figure 5.9 présente les résultats obtenus en fonction de γ (axe horizontal) en termes de précision (axe vertical) pour l'approche de construction du profil social basée sur les communautés CoBPT. Les différentes courbes représentent différentes valeurs de λ.

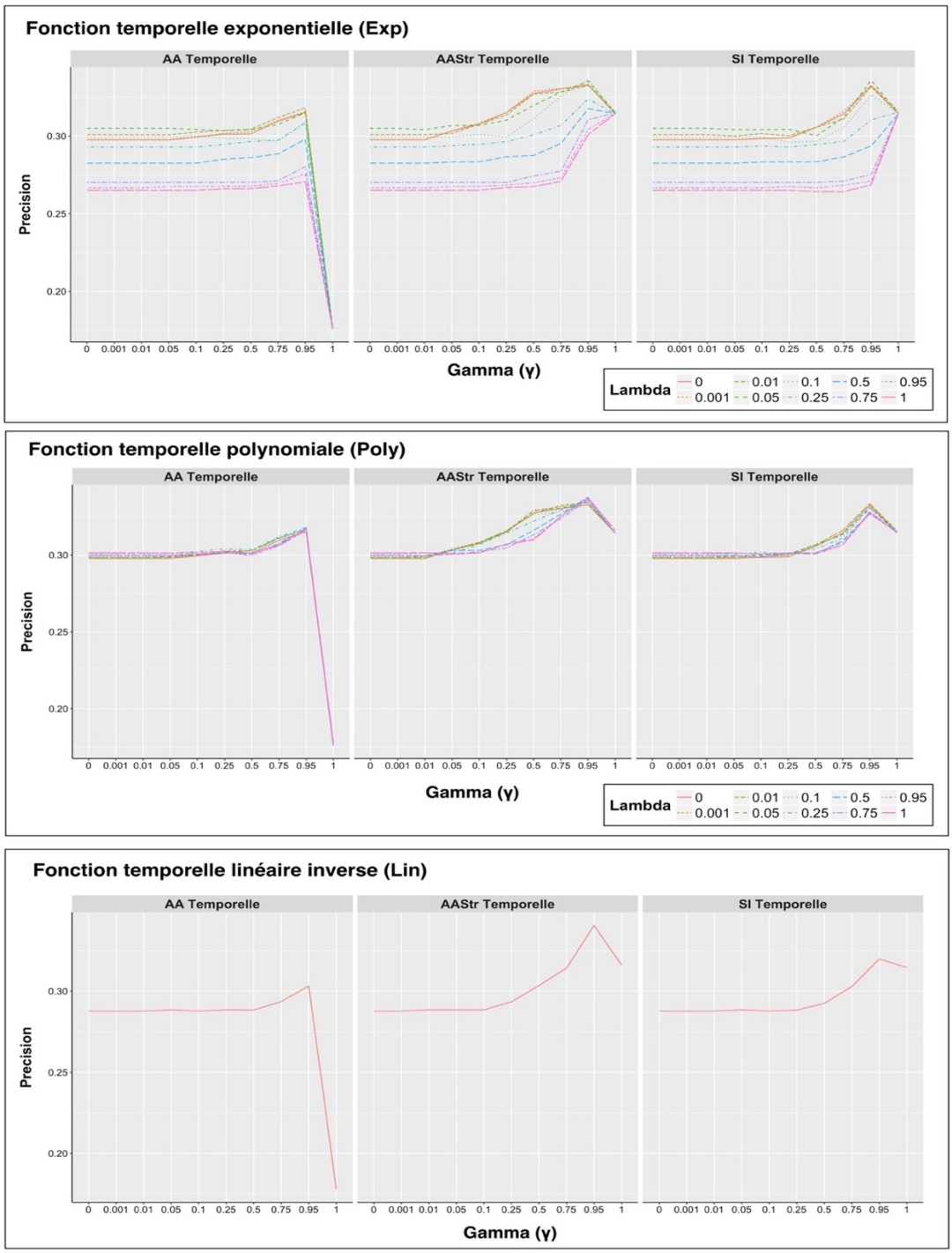


Figure 5.9 Comparaison en termes de précision des résultats en fonction de γ en axe horizontal et λ (différentes courbes) des processus basés sur les communauté (CoBSPT), pour les techniques appliquant les trois fonctions temporelles exponentielle, polynomiale et linéaire.

Nous remarquons qu'avec la fonction linéaire inverse qui ne dépend pas de la valeur de λ, les résultats ont tendance à monter quand γ devient de plus en plus important. Les meilleurs résultats sont trouvés quand γ est fixé haut (0.95 pour les trois techniques). Au-delà, les résultats diminuent.

Pour les fonctions exponentielle et polynomiale, les résultats varient de façon différente selon la technique appliquée et la valeur de λ. Globalement, les résultats varient de la même manière que dans l'approche basés sur les individus.

Pour les techniques appliquant la fonction temporelle exponentielle, en fixant λ très bas (entre 0 et 0.1) nous obtenons généralement de meilleurs résultats par rapport à des valeurs de λ plus hautes (entre 0.25 et 1). Les résultats ont tendance à monter quand γ devient de plus en plus important jusqu'à la valeur optimale trouvée pour 0.95 et diminuent au-delà.

Pour les techniques appliquant la fonction temporelle polynomiale nous ne voyons pas clairement une grande différence des résultats en fonction de la valeur de λ. Les résultats ont tendance à monter quand γ devient de plus en plus important jusqu'à la valeur optimale trouvée pour 0.95 et diminuent au-delà. La valeur optimale de λ est différente selon les techniques appliquées. Pour la technique PolyAATemp, quelle que soit la valeur de γ fixée, on n'observe pas beaucoup d'écart entre les résultats lors-qu'on fait varier λ. Pour la technique PolyAAStrTemp, les résultats restent très proches en fixant γ très bas. Au-delà, les résultats avec λ bas ont tendance à devenir meilleurs que ceux avec λ fixé plus haut. Cependant, quand γ = 0.95, qui est la valeur optimale, nous trouvons de bons résultats quand la valeur de λ est haute. Néanmoins les écarts de résultats ne sont pas très significatifs. Pour la technique PolySITemp, les résultats restent très proches quelle que soit la valeur de λ en fixant γ très bas. Pour γ plus haut, les résultats avec λ bas ont tendance à devenir meilleurs que ceux avec λ plus haut jusqu'à γ = 0.95, qui est la valeur optimale.

Comme dans le cas des approches basées sur les individus, les résultats chutent quand γ =1.

D'après l'observation des résultats en fonction de γ et λ, nous pouvons tirer les conclusions suivantes. Les valeurs optimales de γ qui montrent l'importance du poids temporel des individus par rapport au poids temporel des informations, sont généralement hautes (entre 0.75 et 0.95). Ceci est clairement visible dans les techniques de calcul qui appliquent la fonction temporelle linaire (où les résultats ne dépendent pas de λ). La valeur haute de γ montre qu'il faut prendre en compte avec une proportion assez importante le poids temporel des individus par rapport au poids temporel des informations. Quand γ est à 1, les résultats chutent dramatiquement. Ceci montre que, même si la proportion du poids temporel des individus est importante, il est nécessaire de prendre en compte aussi le poids temporel des informations. La combinaison des deux poids temporels dans la méthode proposée est donc justifiée ici. Le résultat peut paraître contre intuitif mais il montre que les utilisateurs sont plus « influencés » par les individus de leur réseau que par les informations qui circulent entre eux. Si les raisons des connexions entre personnes dans le réseau sont, dans le cas d'espèce, l'homophilie (réseau de co-auteurs), il n'en demeure pas moins que le réseau social est plus « représentatif » d'un utilisateur que les publications elles-mêmes (en termes de mots clefs).

Nous pouvons aussi constater que les résultats varient en fonction de λ et que la valeur optimale dépend de la technique utilisée mais aussi de la valeur de γ. Les valeurs optimales de λ généralement différentes de 0 montrent que la prise en compte du temps est importante pour améliorer les résultats des processus basés sur les individus même si le taux de dépréciation (λ) est assez bas dans certains cas. Ici encore, la nature du réseau et l'évolution plutôt lente des thèmes de recherche des scientifiques peuvent justifier ce résultat.

❖ Résultats en fonction de α

Après avoir montré l'efficacité de la méthode temporelle proposée sans tenir compte la valeur de α, nous avons étudié les résultats en prenant en compte le poids de structure de centralité (qui consiste à faire varier α).

La Figure 5.10 représente les meilleurs résultats en termes de précision, correspondant à la meilleure combinaison de paramètres (γ, λ), des techniques utilisées pour l'approche de construction du profil social basée sur les communautés (processus CoBSP et CoBSPT) et pour l'approche de construction du profil social basée sur les individus (processus IBSP et IBSPT) en fonction de la valeur de α.

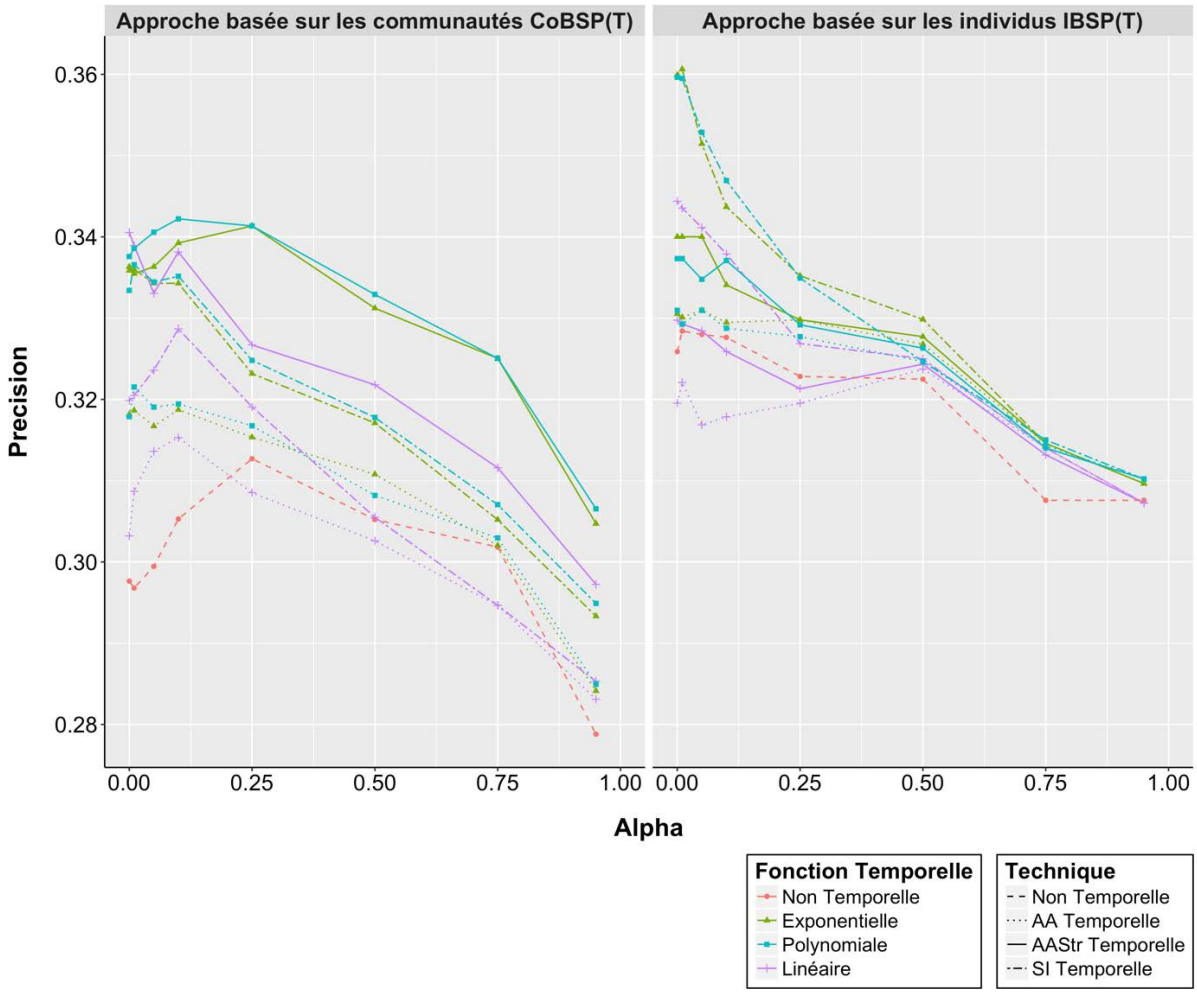


Figure 5.10 Comparaison des meilleurs résultats en termes de précision avec l'application des différentes techniques pour l'approche basée sur les communautés (processus CoBSP et CoBSPT) et pour l'approche basée sur les individus (processus IBSP et IBSPT) en fonction des valeurs de α.

Pour l'approche basée sur les communautés, les meilleurs résultats en termes de précision, rappel et F-mesure obtenus par le processus CoBSP (non prise en compte du temps) sont respectivement 0.31271, 0.17767 et 0.21907.

Pour le processus CoBSPT, nous présentons les résultats par rapport à la technique utilisée : nous comparons les résultats selon la méthode temporelle appliquée et pour chacune, selon la technique de calcul du poids temporel des intérêts.

- Pour les techniques appliquant la fonction temporelle exponentielle (Exp) :

- La technique ExpAATemp produit la meilleure précision (0.31901) lorsque $\alpha = 0.001$, $\lambda = 0.01$ et $\gamma = 0.95$ avec 2.016 % d'amélioration par rapport au processus CoBSP. Le meilleur rappel (0.18324) est obtenu lorsque $\alpha = 0.05$, $\lambda = 0$ et $\gamma = 0.95$ avec 3.133 % d'amélioration par rapport au processus CoBSP. La meilleure F-mesure (0.22433) est obtenue lorsque $\alpha = 0.1$, $\lambda = 0$ et $\gamma = 0.75$ avec 2.402 % d'amélioration par rapport au processus CoBSP.
- La technique ExpAAStrTemp produit la meilleure précision (0.34134) lorsque $\alpha = 0.25$, $\lambda = 0$ et $\gamma = 0.75$ avec 9.158 % d'amélioration par rapport au processus CoBSP. Le meilleur rappel (0.19439) est obtenu lorsque $\alpha = 0.25$, $\lambda = 0$ et $\gamma = 0.75$ avec 9.406 % d'amélioration par rapport au processus CoBSP. La meilleure F-mesure (0.23941) est obtenue lorsque $\alpha = 0.25$, $\lambda = 0$ et $\gamma = 0.75$ avec 9.282 % d'amélioration par rapport au processus CoBSP.
- La technique ExpSITemp produit la meilleure précision (0.3367) lorsque $\alpha = 0.001$, $\lambda = 0.05$ et $\gamma = 0.95$ avec 7.673 % d'amélioration par rapport au processus CoBSP. Le meilleur rappel (0.19289) est obtenu lorsque $\alpha = 0.1$, $\lambda = 0$ et $\gamma = 0.95$ avec 8.564 % d'amélioration par rapport au processus CoBSP. La meilleure F-mesure (0.23628) est obtenue lorsque $\alpha = 0.1$, $\lambda = 0$ et $\gamma = 0.95$ avec 7.856 % d'amélioration par rapport au processus CoBSP.

- Pour les techniques appliquant la fonction temporelle Polynomiale (Poly) :
  - La technique PolyAATemp produit la meilleure précision (0.32153) lorsque $\alpha = 0.01$, $\lambda = 0.95$ et $\gamma = 0.95$ avec 2.822 % d'amélioration par rapport au processus CoBSP. Le meilleur rappel (0.18324) est obtenu lorsque $\alpha = 0.05$, $\lambda = 0$ et $\gamma = 0.95$ avec 3.133 % d'amélioration par rapport au processus CoBSP. La meilleure F-mesure (0.22433) est obtenue lorsque $\alpha = 0.1$, $\lambda = 0$ et $\gamma = 0.75$ avec 2.402 % d'amélioration par rapport au processus CoBSP.
  - La technique PolyAAStrTemp produit la meilleure précision (0.34222) lorsque $\alpha = 0.1$, $\lambda = 1$ et $\gamma = 0.95$ avec 9.438 % d'amélioration par rapport au processus CoBSP. Le meilleur rappel (0.19475) est obtenu lorsque $\alpha = 0.1$, $\lambda = 1$ et $\gamma = 0.95$ avec 9.61 % d'amélioration par rapport au processus CoBSP. La meilleure F-mesure (0.24031) est obtenue lorsque $\alpha = 0.1$, $\lambda = 1$ et $\gamma = 0.95$ avec 9.694 % d'amélioration par rapport au processus CoBSP.
  - La technique PolySITemp produit la meilleure précision (0.33657) lorsque $\alpha = 0.01$, $\lambda = 0.25$ et $\gamma = 0.95$ avec 7.631 % d'amélioration par rapport au processus CoBSP. Le meilleur rappel (0.19289) est obtenu lorsque $\alpha = 0.1$, $\lambda = 0$ et $\gamma = 0.95$ avec 8.564 % d'amélioration par rapport au processus CoBSP. La meilleure F-mesure (0.23628) est obtenue lorsque $\alpha = 0.1$, $\lambda = 0$ et $\gamma = 0.95$ avec 7.856 % d'amélioration par rapport au processus CoBSP.
- Pour les techniques appliquant la fonction temporelle Linéaire (Lin) :
  - la technique LinAATemp produit la meilleure précision (0.31531) lorsque $\alpha = 0.1$ et $\gamma = 0.95$ avec 0.832 % d'amélioration par rapport au processus CoBSP. Le meilleur rappel (0.17777) est obtenu lorsque $\alpha = 0.1$ et $\gamma = 0.95$ avec 0.053 % d'amélioration par rapport au processus CoBSP. La meilleure F-mesure (0.22033) est obtenue lorsque $\alpha = 0.1$ et $\gamma = 0.95$ avec 0.575 % d'amélioration par rapport au processus CoBSP.
  - la technique LinAAStrTemp produit la meilleure précision (0.34054) lorsque $\alpha \in \{0.0, 0.001\}$ et $\gamma = 0.95$ avec 8.901 % d'amélioration par rapport

au processus CoBSP. Le meilleur rappel (0.19179) est obtenu lorsque $\alpha \in \{0.0, 0.001\}$ et $\gamma = 0.95$ avec 7.942 % d'amélioration par rapport au processus CoBSP. La meilleure F-mesure (0.23728) est obtenue lorsque $\alpha \in \{0.0, 0.001\}$ et $\gamma = 0.95$ avec 8.312 % d'amélioration par rapport au processus CoBSP.

- la technique LinSITemp produit la meilleure précision (0.32868) lorsque $\alpha = 0.1$ et $\gamma = 0.95$ avec 5.108 % d'amélioration par rapport au processus CoBSP. Le meilleur rappel (0.186) est obtenu lorsque $\alpha = 0.1$ et $\gamma = 0.95$ avec 4.688 % d'amélioration par rapport au processus CoBSP. La meilleure F-mesure (0.22994) est obtenue lorsque $\alpha = 0.1$ et $\gamma = 0.95$ avec 4.961 % d'amélioration par rapport au processus CoBSP.

Pour l'approche basée sur les individus, les meilleurs résultats en termes de précision, rappel et F-mesure obtenus par le processus IBSP (non prise en compte du temps) sont respectivement 0.32842, 0.18542 et 0.22896.

Pour le processus IBSPT, nous présentons les résultats par rapport à la technique utilisée : nous comparons les résultats selon la méthode temporelle appliquée et pour chacune, selon la technique de calcul du poids temporel des intérêts.

- Pour les techniques appliquant la fonction temporelle exponentielle (Exp) :
  - La technique ExpAATemp produit la meilleure précision (0.33096) lorsque $\alpha = 0.05$, $\lambda \in \{0.0, 0.001\}$ et $\gamma = 0.01$ avec 0.774 % d'amélioration par rapport au processus IBSP. Le meilleur rappel (0.1873) est obtenu lorsque $\alpha = 0.05$, $\lambda = 0.05$ et $\gamma = 0.05$ avec 1.012 % d'amélioration par rapport au processus IBSP. La meilleure F-mesure (0.23124) est obtenue lorsque $\alpha = 0.05$, $\lambda = 0.05$ et $\gamma = 0.05$ avec 0.999 % d'amélioration par rapport au processus IBSP.
  - La technique ExpAAStrTemp produit la meilleure précision (0.34001) lorsque $\alpha \in \{0.0, 0.001, 0.01, 0.05\}$, $\lambda = 0.1$ et $\gamma = 0.75$ avec 3.531 % d'amélioration par rapport au processus IBSP. Le meilleur rappel (0.19435) est obtenu lorsque $\alpha = 0.05$, $\lambda = 0.1$ et $\gamma = 0.75$ avec 4.815 % d'amélioration par rapport au processus IBSP. La meilleure F-mesure (0.23933) est obtenue lorsque $\alpha = 0.05$, $\lambda = 0.1$ et $\gamma = 0.75$ avec 4.533 % d'amélioration par rapport au processus IBSP.
  - La technique ExpSITemp produit la meilleure précision (0.36063) lorsque $\alpha = 0.01$, $\lambda = 0.01$ et $\gamma = 0.95$ avec 9.81 % d'amélioration par rapport au processus IBSP. Le meilleur rappel (0.20638) est obtenu lorsque $\alpha = 0$, $\lambda = 0$ et $\gamma = 1$ avec 11.305 % d'amélioration par rapport au processus IBSP. La meilleure F-mesure (0.25301) est obtenue lorsque $\alpha = 0.01$, $\lambda = 0.01$ et $\gamma = 0.95$ avec 10.508 % d'amélioration par rapport au processus IBSP.
- Pour les techniques appliquant la fonction temporelle polynomiale (Poly) :
  - La technique PolyAATemp produit la meilleure précision (0.33096) lorsque $\alpha \in \{0.0, 0.05\}$, $\lambda \in \{0.0, 0.001\}$ et $\gamma \in \{0.01, 0.05\}$ avec 0.774 % d'amélioration par rapport au processus IBSP. Le meilleur rappel (0.1867) est obtenu lorsque $\alpha = 0.05$, $\lambda = 0$ et $\gamma = 0.01$ avec 0.69 % d'amélioration par rapport au processus IBSP. La meilleure F-mesure (0.23059) est obtenue lorsque $\alpha = 0.05$, $\lambda = 0$ et $\gamma = 0.01$ avec 0.713 % d'amélioration par rapport au processus IBSP.

- La technique PolyAAStrTemp produit la meilleure précision (0.33733) lorsque $\alpha \in \{0.0, 0.001, 0.01\}$, $\lambda \in \{0.75, 0.95\}$ et $\gamma = 0.75$ avec 2.714 % d'amélioration par rapport au processus IBSP. Le meilleur rappel (0.19169) est obtenu lorsque $\alpha \in \{0.0, 0.001, 0.01\}$, $\lambda \in \{0.75, 0.95\}$ et $\gamma = 0.75$ avec 3.385 % d'amélioration par rapport au processus IBSP. La meilleure F-mesure (0.23661) est obtenue lorsque $\alpha \in \{0.0, 0.001, 0.01\}$, $\lambda = 0.95$ et $\gamma = 0.75$ avec 3.341 % d'amélioration par rapport au processus IBSP.
- La technique PolySITemp produit la meilleure précision (0.36049) lorsque $\alpha = 0.001$, $\lambda = 0.05$ et $\gamma = 0.95$ avec 9.767 % d'amélioration par rapport au processus IBSP. Le meilleur rappel (0.20638) est obtenu lorsque $\alpha = 0$, $\lambda = 0$ et $\gamma = 1$ avec 11.305 % d'amélioration par rapport au processus IBSP. La meilleure F-mesure (0.2527) est obtenue lorsque $\alpha = 0.001$, $\lambda = 0.05$ et $\gamma = 0.95$ avec 10.373 % d'amélioration par rapport au processus IBSP.

- Pour les techniques appliquant la fonction temporelle linéaire (Lin) :
  - la technique LinAATemp produit la meilleure précision (0.32372) lorsque $\alpha = 0.5$ et $\gamma = 0.25$ avec une perte de -1.429 % par rapport au processus IBSP. Le meilleur rappel (0.18432) est obtenu lorsque $\alpha = 0.01$ et $\gamma = 0.75$ avec une perte de - 0.592 % par rapport au processus IBSP. La meilleure F-mesure (0.2271) est obtenue lorsque $\alpha = 0.01$ et $\gamma = 0.75$ avec une perte de -0.812 % d'amélioration par rapport au processus IBSP.
  - la technique LinAAStrTemp produit la meilleure précision (0.32975) lorsque $\alpha \in \{0.0, 0.001\}$ et $\gamma = 0.75$ avec 0.405 % d'amélioration par rapport au processus IBSP. Le meilleur rappel (0.18742) est obtenu lorsque $\alpha \in \{0.0, 0.001\}$ et $\gamma = 0.75$ avec 1.082 % d'amélioration par rapport au processus IBSP. La meilleure F-mesure (0.23136) est obtenue lorsque $\alpha \in \{0.0, 0.001\}$ et $\gamma = 0.75$ avec 1.05 % d'amélioration par rapport au processus IBSP.
  - la technique LinSITemp produit la meilleure précision (0.34439) lorsque $\alpha \in \{0.0, 0.001\}$ et $\gamma = 0.95$ avec 4.864 % d'amélioration par rapport au processus IBSP. Le meilleur rappel (0.19878) est obtenu lorsque $\alpha = 0.05$ et $\gamma = 1$ avec 7.208 % d'amélioration par rapport au processus IBSP. La meilleure F-mesure (0.2409) est obtenue lorsque $\alpha \in \{0.0, 0.001\}$ et $\gamma = 0.95$ avec 5.219 % d'amélioration par rapport au processus IBSP.

Nous remarquons que, dans les deux approches de construction du profil social (basée sur les individus ou sur les communautés), pour les techniques appliquant les fonctions temporelles exponentielles et polynomiales, la méthode temporelle proposée produit généralement de meilleurs résultats quelle que soit la technique de calcul du poids des intérêts et pour toutes les valeurs de α. Pour les techniques appliquant la fonction temporelle linéaire, la méthode temporelle proposée produit avec la technique AATemp soit des résultats moins bons soit très peu d'amélioration par rapport aux résultats des processus existants. Les résultats ont tendance à baisser quand α diminue.

En comparant les techniques appliquant les fonctions temporelles exponentielle et polynomiale, nous ne voyons pas d'écart important entre les résultats de ces deux approches. Comme montré ci-après, la fonction temporelle « gagnante » est différente selon l'approche de construction du profil social choisie et la technique appliquée.

Dans l'approche basée sur les communautés (processus CoBSPT), la fonction temporelle polynomiale parait donner de meilleurs résultats que la fonction exponentielle. Pour l'approche basée sur les individus (processus IBSPT), la fonction temporelle exponentielle donne clairement de meilleurs résultats avec la technique AAStrTemp.

En comparant les techniques de calcul de poids, nous constatons que la technique AATemp donne généralement de moins bons résultats la technique AAStrTemp et SITemp pour les deux processus CoBSPT ou IBSPT. En comparant la technique AAStrTemp et SITemp, la technique « gagnante » est différente en fonction du processus choisi. Avec le processus CoBSPT, la technique AAStrTemp donne généralement de meilleurs résultats. Avec le processus IBSPT, les meilleurs résultats sont obtenus en appliquant la technique SITemp.

En comparant les résultats des processus des deux approches (IBSPT vs CoBSPT), les résultats du processus IBSPT produisent de meilleurs résultats pour les 236 auteurs. Ceci parait contradictoire par rapport aux travaux de (Tchuente, 2013) qui ont montré que le processus basé sur les communautés produit de meilleurs résultats par rapport au processus basé sur les individus. Ceci vient du fait que, dans le travail de (Tchuente, 2013), pour trouver les meilleurs résultats, les auteurs pris en compte dans les calculs ont au minimum 70 co-auteurs et ont un réseau entre 10% et 30% de densité. Dans notre échantillon de test les utilisateurs possèdent un nombre de co-auteurs plus varié. Dans l'ensemble des 236 utilisateurs étudiés, il existe des auteurs ayant peu de voisins sociaux (co-auteurs) dans leur réseau égocentrique (réseau peu dense). Cela peut amener l'algorithme de détection de communautés à détecter des communautés qui ne sont pas pertinentes (une seule grosse communauté par exemple) et produire des résultats non pertinents et ainsi diminuer la précision totale de cette approche. Comme déjà précisé dès l'introduction, nous supposons que la taille du réseau et la densité jouent un rôle sur les résultats obtenus. Ces points sont étudiés dans la section suivante.

c. Résultats selon la taille et la densité du réseau

Nous avons étudié les résultats en fonction de la taille du réseau ainsi que de sa densité pour étudier l'impact de ces deux facteurs sur les résultats obtenus. Pour cela, l'échantillon de test est découpé en plusieurs sous-échantillons en fonction de la taille et la densité du réseau.

L'étude paramétrique comme présentée dans la section précédente est appliquée à chaque sous-échantillon. Pour chaque échantillon, les résultats pour chaque technique de calcul du poids temporel sont présentés par la précision moyenne (resp. rappel moyen et F-mesure moyenne) pour l'échantillon, correspondant à la meilleure combinaison de paramètres (γ, λ, α); la combinaison qui a obtenu la meilleure moyenne.

Les résultats les plus importants sur chaque sous-échantillon se trouvent en annexe. Dans l'Annexe 1-c sont présentés les résultats en fonction de taille du réseau et dans l'Annexe 1-d ceux en fonction de la densité du réseau.

En ce qui concerne la taille du réseau, dans le cas des auteurs ayant un nombre peu important de co-auteurs (moins de 50), les processus basés sur les communautés produisent de moins bons résultats avec un écart important par rapport à ceux produits par les processus basés sur les individus. L'écart le plus important est observé dans l'échantillon avec des auteurs ayant moins de 10 co-auteurs. L'écart baisse quand on augmente le nombre de co-auteurs. Cela peut s'expliquer par le fait que l'algorithme de détection de communautés détectent certainement des communautés moins pertinentes quand la taille du réseau est très petit : il peut détecter une seule communauté et génère donc un biais lors du calcul de poids des éléments. Cela peut également être lié au fait que dans les petits réseaux, les informations partagées ne sont pas assez importantes pour retrouver les informations significatives pour l'utilisateur. Les

processus basés sur les communautés produisent des résultats plus satisfaisants que ceux issus des processus basés sur les individus quand le nombre de co-auteurs est supérieur à 100. Cependant quand le nombre de co-auteurs dépasse 250, les résultats de l'approche basée sur les communautés deviennent moins bons que ceux de l'approche basée sur les individus.

En appliquant la méthode temporelle proposée, presque toutes les techniques de calcul du poids temporel des intérêts améliorent les résultats dans tous les échantillons.

En ce qui concerne la densité du réseau, nous constatons que c'est dans les réseaux qui ont peu de densité (<10%) que l'on obtient les meilleurs résultats quel que soit le processus utilisé. Les résultats baissent pour des réseaux de densité plus importante (ce qui parait contradictoire avec notre hypothèse de départ). Dans les réseaux ayant une densité <30%, on ne voit pas trop de différence entre les résultats des processus basés sur les communautés et ceux des processus basés sur les individus. Dans les réseaux ayant une densité >30%, tous les résultats chutent dramatiquement et nous constatons que les processus basés sur les communautés donnent de moins bons résultats que ceux des processus basés sur les individus avec un écart assez important. Ceci peut s'expliquer 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) qui ne sera pas assez discriminante pour caractériser l'utilisateur. De plus, parmi les auteurs ayant des réseaux très denses, on peut retrouver des auteurs ayant une taille de réseau très petit (cf. étude par rapport à la taille du réseau expliquée précédemment).

Nous avons donc décidé d'étudier les résultats selon la taille et la densité du réseau en même temps. Nous fixons ici :

- En ce qui concerne la taille du réseau :
  - un réseau ayant moins de 50 individus sera considéré comme réseau de petite taille,
  - un réseau ayant plus de 50 individus sera considéré comme réseau de grande taille.
- En ce qui concerne la densité du réseau :
  - un réseau ayant moins de 10% de densité sera considéré comme un réseau épars,
  - un réseau ayant entre 10 et 30% sera considéré comme un réseau assez dense,
  - un réseau ayant plus de 30% sera considéré comme un réseau très dense.

Nous donnons les résultats en fonction de la taille et la densité ci-après.

❖ Réseau de petite taille et épars

Cet échantillon représente les réseaux des 46 auteurs ayant moins de 50 individus et ayant moins de 10% de densité. Les résultats sont présentés Figure 5.11.

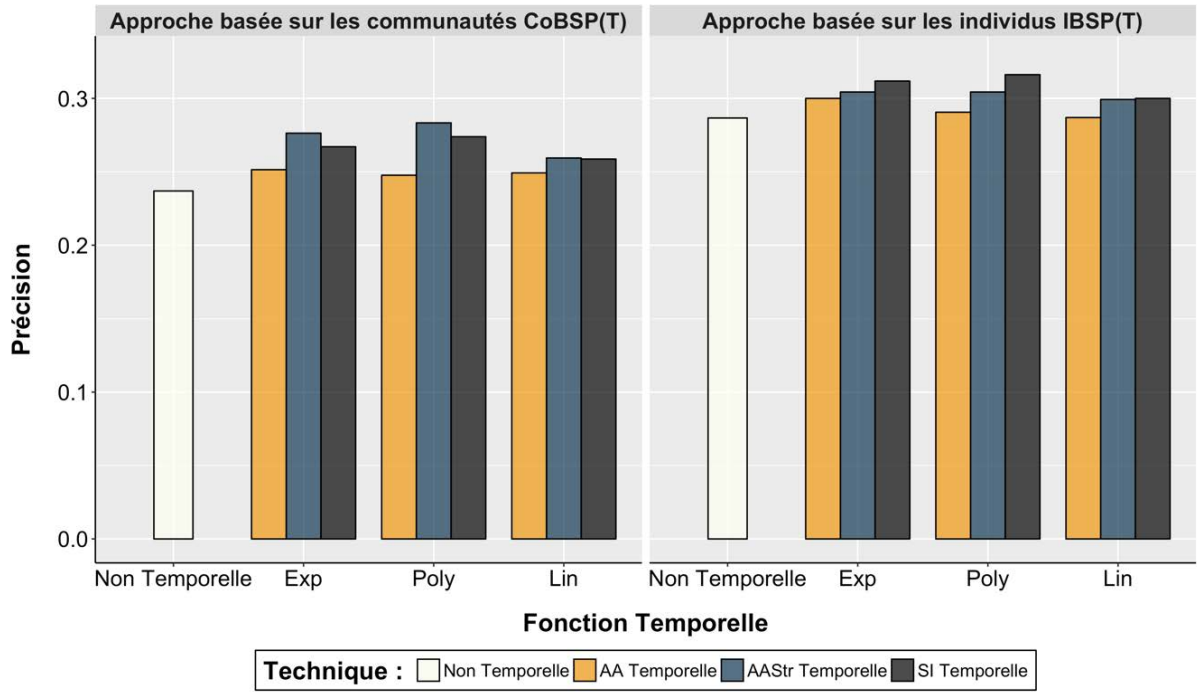


Figure 5.11 Comparaison de la meilleure précision moyenne pour les processus de construction du profil social, en appliquant la méthode temporelle (IBSPT et CoBSPT) avec ceux des processus qui ne prennent pas en compte de temps (IBSP et CoBSP) dans l'échantillon des utilisateurs d'un réseau de petite taille et épars

La méthode temporelle proposée améliore les résultats des processus existants quelles que soit l'approche de construction du profil, la technique de calcul du poids temporel des intérêts et la fonction temporelle appliquée. Les techniques appliquant la fonction linéaire produisent de moins bons résultats par rapport aux deux autres fonctions. La fonction temporelle polynomiale produit des résultats légèrement meilleurs que la fonction temporelle exponentielle. Concernant les techniques utilisées, AATemporelle produit généralement de moins bons résultats quelles que soient les approches (basée sur les communautés ou basée sur les individus). AAStrTemp produit généralement les meilleurs résultats pour l'approche basée sur les communautés et SITemp produit les meilleurs résultats pour l'approche basée sur les individus.

En ce qui concerne la comparaison des deux approches, nous observons que, dans cet échantillon, les résultats de l'approche basée sur les communautés sont moins bons que ceux de l'approche basée sur les individus quelle que soit la technique utilisée.

❖ Réseau de petite taille et assez dense

Cet échantillon représente les réseaux des 60 auteurs ayant moins de 50 individus et ayant entre de 10% et 30% de densité. Les résultats sont présentés Figure 5.12.

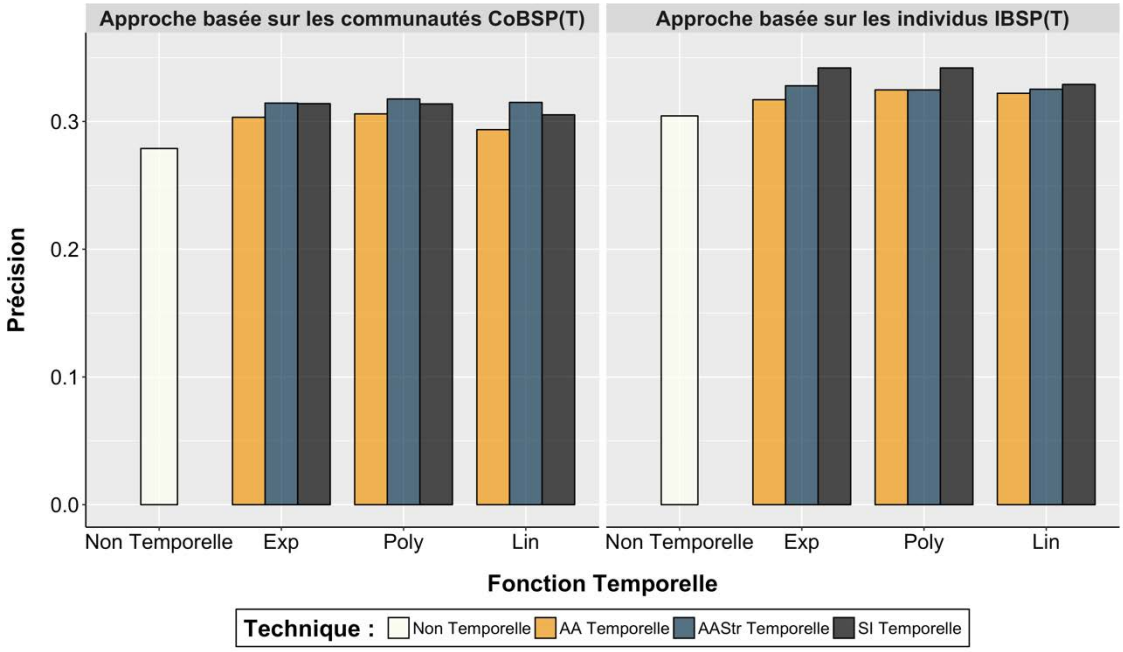


Figure 5.12 Comparaison de la meilleure précision moyenne pour les processus de construction du profil social, en appliquant la méthode temporelle (IBSPT et CoBSPT) avec celle des processus qui ne prennent pas en compte le temps (IBSP et CoBSP) dans l'échantillon des utilisateurs dans des réseaux de petite taille et assez denses

Dans cet échantillon, la méthode temporelle proposée, améliore les résultats des processus existants quelles que soient l'approche, la technique de calcul du poids temporel et la fonction temporelle appliquées. Nous obtenons généralement une meilleure précision par rapport à celle obtenue dans l'échantillon précédent.

En comparant les fonctions temporelles appliquées, nous n'obtenons pas beaucoup de différence dans les résultats. Les techniques appliquant la fonction linéaire produisent des résultats légèrement moins bons que ceux obtenus en appliquant les deux autres fonctions. La fonction temporelle polynomiale produit des résultats légèrement meilleurs que ceux produits par la fonction temporelle exponentielle.

L'observation des comparaisons des différentes techniques utilisées est similaire à celle de l'échantillon précédent. La technique AATemp donne globalement de moins bons résultats quelles que soient les approches (basée sur les communautés ou basée sur les individus). La technique AAStrTemp donne globalement les meilleurs résultats pour l'approche basée sur les communautés et la technique SITemp donne les meilleurs résultats pour l'approche basée sur les individus.

En ce qui concerne les approches, nous observons que dans cet échantillon, l'approche basée sur les communautés donne de moins bons résultats que ceux produits par l'approche basée sur les individus quelle que soit la technique utilisée. Cependant l'écart entre les résultats produits par les deux approches est moins important que celui obtenu dans l'échantillon précédent.

❖ Réseau de petite taille et très dense

Cet échantillon représente les réseaux des 21 auteurs ayant moins de 50 individus et ayant plus de 30% de densité. Les résultats sont présentés Figure 5.13.

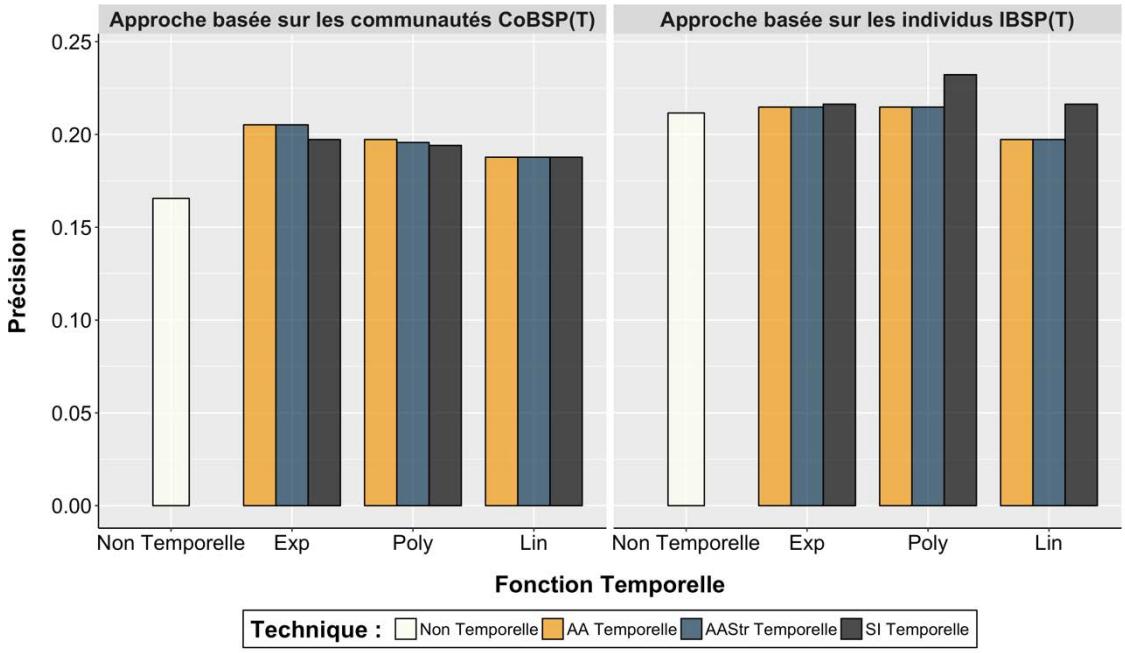


Figure 5.13 Comparaison de la meilleure précision moyenne pour les processus de construction du profil social, en appliquant la méthode temporelle (IBSPT et CoBSPT) avec celle des processus qui ne prennent pas en compte le temps (IBSP et CoBSP) dans l'échantillon des utilisateurs dans les réseaux de petite taille et très denses

Dans cet échantillon, la méthode temporelle proposée améliore encore les résultats des processus existants quelles que soient l'approche, la technique de calcul du poids temporel et la fonction temporelle utilisées.

Les précisions moyennes obtenues pour toutes les techniques sont moins importantes par rapport à celles obtenues dans l'échantillon précédent.

En comparant les fonctions temporelles utilisées, les techniques appliquant la fonction linéaire produisent de moins bons résultats par rapport à celles appliquant les deux autres fonctions. La fonction temporelle exponentielle a tendance à produire de meilleurs résultats par rapport à la fonction temporelle polynomiale.

En comparant les techniques de calcul utilisées, dans l'approche basée sur les communautés, la technique AATemp produit généralement de meilleurs résultats, ce qui est différent des observations faites dans les deux échantillons précédents. Dans l'approche basée sur les individus, la technique SITemp produit généralement de meilleurs résultats.

Nous observons que, dans cet échantillon, les résultats de l'approche basée sur les communautés sont moins bons par rapport à ceux de l'approche basée sur les individus quelle que soit la technique utilisée avec un écart qui devient important. Cela peut être expliqué par la densité de réseau qui amène à la détection de communautés non discriminantes (peu de communautés ou une seule grosse communauté), comme expliqué précédemment.

❖ Réseau de grande taille et épars

Cet échantillon représente les réseaux des 102 auteurs ayant plus de 50 individus et ayant moins de 10% de densité. Les résultats sont présentés Figure 5.14.

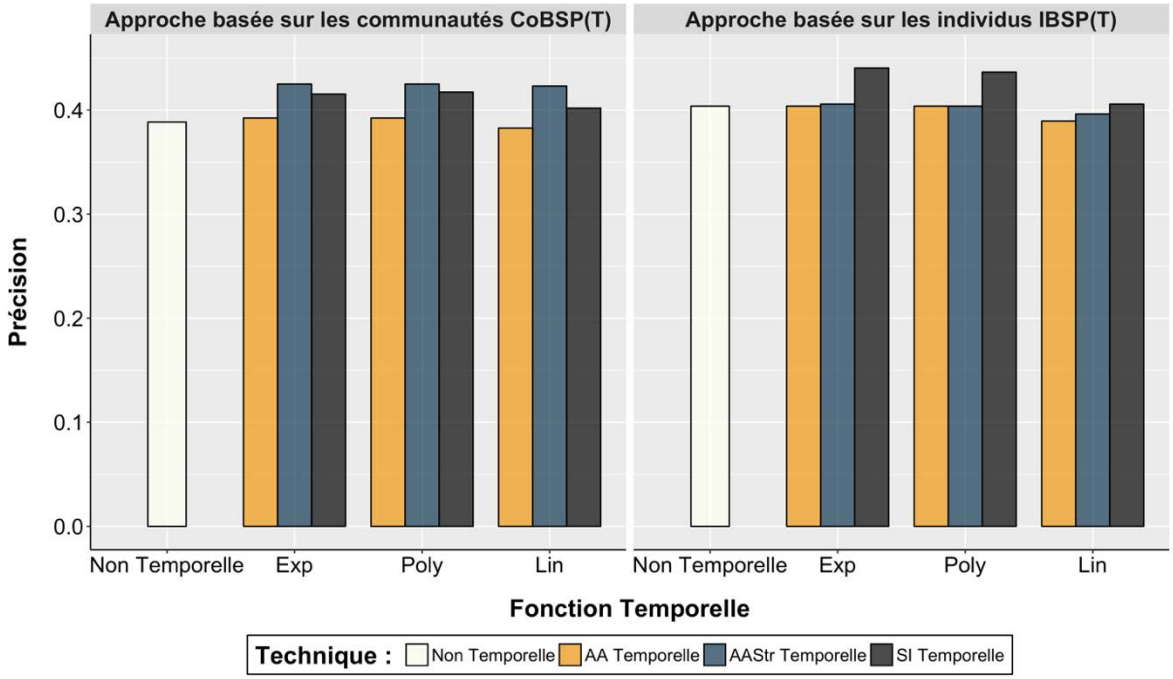


Figure 5.14 Comparaison de la meilleure précision moyenne pour les processus de construction du profil social, en appliquant la méthode temporelle (IBSPT et CoBSPT) avec celle des processus qui ne prennent pas en compte le temps (IBSP et CoBSP) dans l'échantillon des utilisateurs dans les réseaux de grande taille et éparse

Dans cet échantillon, la méthode temporelle proposée améliore globalement les résultats des processus existants pour toutes les techniques appliquant les fonctions temporelles exponentielle et polynomiale. La technique AATemp appliquant la fonction temporelle linéaire produit de mauvais résultats quelle que soit approche de construction de profil. La technique AAStrTemp appliquant la fonction temporelle linéaire donne de moins bons résultats dans l'approche basée sur les individus.

Nous observons que les précisions moyennes obtenues pour toutes les techniques augmentent par rapport à celles obtenues dans l'échantillon précédent.

En comparant les fonctions temporelles utilisées, les techniques appliquant la fonction linéaire produisent généralement de moins bons résultats par rapport aux deux autres fonctions. Les fonctions temporelles exponentielle et polynomiale donnent à peu près les mêmes résultats quelle que soit la technique de calcul du poids temporel utilisée.

En comparant les techniques de calcul utilisées, la technique AATemporelle produit généralement de moins bons résultats quelles que soient les approches (basée sur les communautés ou basée sur les individus). La technique AAStrTemp produit généralement les meilleurs résultats pour l'approche basée sur les communautés et la technique SITemp produit les meilleurs résultats pour l'approche basée sur les individus.

Nous observons que, dans cet échantillon, les résultats de l'approche basée sur les communautés sont légèrement moins bons par rapport à ceux de l'approche basée sur les individus.

- Réseau de grande taille et assez dense

Cet échantillon représente les réseaux des 7 auteurs ayant plus de 50 individus et ayant entre 10% et 30% de densité. Les résultats sont présentés Figure 5.15.

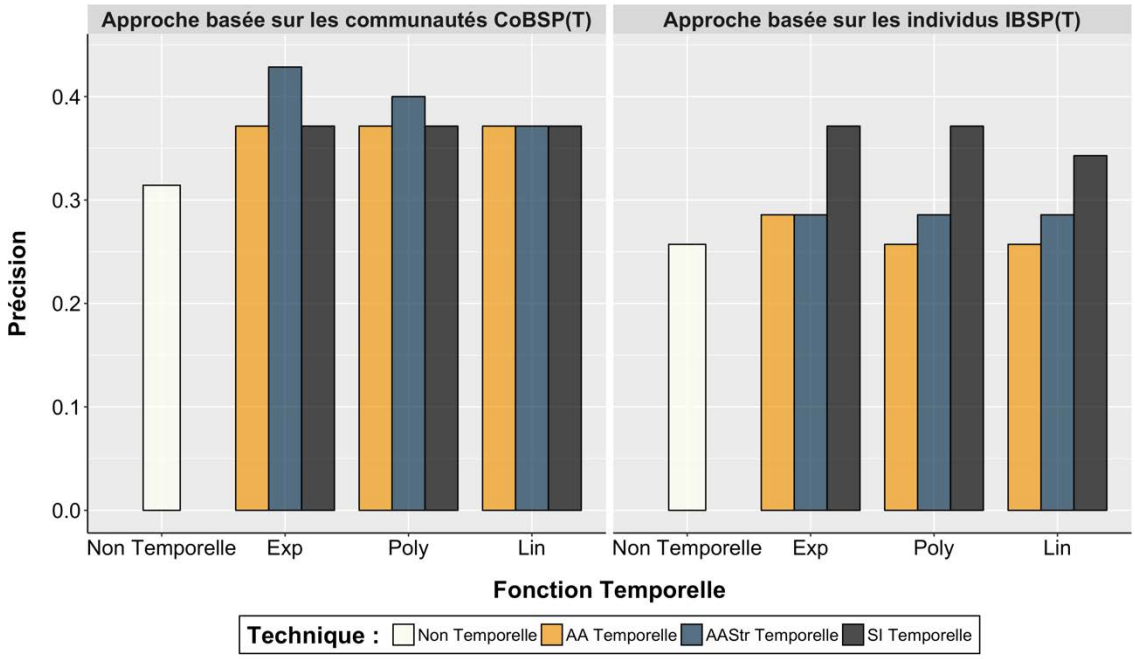


Figure 5.15 Comparaison de la meilleure précision moyenne pour les processus de construction du profil social, en appliquant la méthode temporelle (IBSPT et CoBSPT) avec celle des processus qui ne prennent pas en compte le temps (IBSP et CoBSP) dans l'échantillon des utilisateurs des réseaux de grande taille et assez denses

Dans cet échantillon, la méthode temporelle proposée améliore globalement les résultats des processus existants pour toutes les techniques appliquant les fonctions temporelles exponentielle ou polynomiale. La technique AATemp appliquant les fonctions temporelles linéaire ou polynomiale produit de mauvais résultats pour l'approche basée sur les individus.

En comparant les fonctions temporelles utilisées, les techniques appliquant la fonction linéaire produisent de moins bons résultats par rapport aux deux autres fonctions. La fonction temporelle exponentielle donne de meilleurs résultats par rapport aux deux autres fonctions.

En comparant les techniques appliquées, la technique AATemporelle produit de moins bons résultats quelles que soient les approches de construction du profil. La technique AAStrTemp produit les meilleurs résultats pour l'approche basée sur les communautés et la technique SITemp produit les meilleurs résultats pour l'approche basée sur les individus.

Nous observons que dans cet échantillon, les résultats de l'approche basée sur les communautés deviennent plus pertinents que ceux des processus de l'approche basée sur les individus quelle que soit la technique de calcul du poids temporel utilisée. Ce résultat est cohérent avec les observations de (Tchuente, 2013) donnant de très bons résultats sur les réseaux denses et ayant plus de 70 co-auteurs.

Notons que nous ne possédons pas d'échantillon représentant un réseau de grande taille et très dense. Ceci parait logique car, selon la formule de calcul de la densité du réseau, cette dernière diminue quand le nombre de relations augmente. Dans la vie réelle, pour un utilisateur qui possède beaucoup d'amis, il est difficile de penser que tous ses amis sont également amis entre eux.

d. Discussion des résultats

Nous avons montré, par les expérimentations sur DBLP, que, globalement, les processus appliquant la méthode temporelle proposée (comportant différentes techniques de calcul du poids temporel des intérêts) donnent de meilleurs résultats par rapport à ceux donnés par les processus existants quelle que soit l'approche de construction du profil social (basée sur les communautés ou les individus). En étudiant spécifiquement les résultats sur différents échantillons, on constate des améliorations dans chaque échantillon.

En analysant les résultats en fonction des techniques ou fonctions temporelles, nous avons montré que la fonction temporelle et la technique qui donnent les meilleurs résultats varient selon l'approche mais aussi selon les échantillons étudiés.

La méthode temporelle linéaire qui diminue l'importance des informations ou des relations de façon linéaire, donne généralement de moins bons résultats par rapport aux fonctions temporelles exponentielle et polynomiale qui diminuent l'importance des informations avec un taux de dépréciation assez petit.

Les valeurs de γ, qui montrent le rapport entre le poids temporel des individus et le poids temporel des informations, sont généralement hautes sur l'ensemble des échantillons (0.75 ou 0.95). Ceci montre l'importance de l'évolution de la force des relations entre l'utilisateur et les individus dans son réseau social par rapport à l'évolution des informations. Si nous nous focalisons sur l'observation des résultats obtenus dans le réseau de publications scientifiques, nous constatons que les relations entre les individus jouent un rôle plus important que l'évolution des informations.

Dans la section suivante, nous présentons les expérimentations effectuées dans un réseau de type différent de celui des publications scientifiques : Twitter. Nous évaluons les résultats obtenus en appliquant la méthode temporelle proposée dans ce type de réseau.

5.3.2. Expérimentation sur Twitter

5.3.2.1. Présentation du réseau social Twitter

Le deuxième domaine d'expérimentation choisi est Twitter : un réseau social de micro-blog (cf. section 3.1.2). Il permet de partager en ligne des messages courts appelés « tweets ». Ces messages sont limités à 140 caractères. Les tweets peuvent contenir des textes courts, des hyperliens vers d'autres contenus multimédia (images, vidéo) et des « *hashtags* » qui sont des tags préfixés par le caractère '#' (ex. #Euro2016, #France2016). Un *hashtag* est en effet un sujet explicite utilisé pour annoter le contenu du tweet. On peut utiliser les *hashtags* pour suivre ou parcourir les discussions sur un sujet donné.

Dans Twitter, un utilisateur peut utiliser la fonctionnalité « *follow* » pour suivre d'autres utilisateurs. Cette fonctionnalité lui permet de suivre les informations que les comptes twitter partagent/diffusent. Pour chaque utilisateur, les utilisateurs qu'il suit sont appelés ses « *followings* » et les utilisateurs qui le suivent sont appelés ses « *followers* ». Notons que la relation « *follow* » n'est pas une relation réciproque. Un utilisateur peut suivre une autre personne sans que cette dernière ne le suive. Le réseau social de Twitter est donc représenté par un graphe orienté. Les utilisateurs peuvent également créer leurs propres listes personnelles d'utilisateurs associées à un sujet. Ils peuvent également suivre (*follow*) des listes déjà créées par d'autres utilisateurs. Les *tweets* partagés par l'utilisateur ainsi que ceux partagés par ses « *following* » sont affichés dans un enchainement chronologique, appelé « *timeline* ».

En termes d'interactions, les utilisateurs peuvent :

- **Répondre** (**Reply**) à un tweet d'un autre utilisateur en ajoutant le caractère @ devant l'identité de cet utilisateur (@user)
- **Retransmettre** (**Retweet** ou **RT)** des tweets d'autres utilisateurs.
- **Mentionner** (**Mention**) d'autres utilisateurs dans les tweets en utilisant @user

Lancé en juillet 2006, Twitter est rapidement devenu populaire, jusqu'à réunir plus de 300 millions d'utilisateurs actifs publiant 500 millions de tweets chaque jour et dont plus de la moitié "tweetent" depuis leur téléphone mobile (Statista, 2016).

(Kwak et al., 2010) ont étudié les caractéristiques des données dans Twitter et ont montré que les sujets populaires (*trending topics*) ont tendance à évoluer et peuvent réapparaitre au fil du temps. Le cycle de vie d'un tweet est généralement plus court que des « posts » ou des contenus dans la plupart des autres RSN (souvent une semaine ou moins). Ils ont également trouvé que la moyenne du nombre de tweets par rapport au nombre de followers d'un utilisateur est généralement au-dessus de la médiane. Selon notre point de vue, Twitter peut être considéré comme un réseau de partage d'informations plutôt que comme un site de réseautage social.

Avec ses différentes fonctionnalités proposées (*follow*, *reply*, *mention*, *retweet*), Twitter est considéré comme un réseau social multi-dimensions. Dans notre travail, nous supposons qu'un utilisateur a tendance à suivre (*follow*) d'autres utilisateurs qui partagent des informations concernant ses intérêts. Nous considérons donc la relation « *following* » pour extraire le réseau égocentrique de l'utilisateur. La relation « *follower* » est moins pertinente car il peut s'agir de personnes qui suivent l'utilisateur sans que l'utilisateur ne les suive et donc sans que l'utilisateur ne s'intéresse aux informations que ces dernières partagent.

Dans l'expérimentation menée, nous considérons que les relations « *following* » représentent les liens entre les membres. Le réseau égocentrique d'un utilisateur u est donc l'ensemble de ses « *followings* » (membres). Le « *reply* », « *retweet* » et « *mention* » représentent les interactions entre les membres. Nous considérons que les utilisateurs *u1* et *u2* interagissent entre eux quand l'un retransmet un tweet de l'autre (retweet), répond (*reply*) à un tweet de l'autre ou mentionne (*mention*) l'autre dans un tweet.

Différents travaux ont prouvé la pertinence d'utiliser les *hashtags* pour identifier les intérêts de l'utilisateur (Abel et al., 2013 ; Kacem, Boughanem et Faiz, 2014 ; Liang et al., 2012 ; Mezghani et al., 2012). Dans ce travail, nous proposons donc d'extraire les intérêts sociaux depuis ce type d'information pour construire le profil social de l'utilisateur.

Nous considérons le « jour » comme la granularité de temps utilisée pour calculer le poids temporel des intérêts. Nous utilisons le *timestamp* des tweets pour calculer la fraicheur de chaque *hashtag* contenu dans le « post » afin de calculer le poids temporel de l'information. Pour calculer le poids temporel de chaque individu nous utilisons, comme *timestamp* d'interaction, le *timestamp* de « *retweet* », « *reply* » des tweets entre l'individu et l'utilisateur central ou le *timestamp* des tweets dans lesquels ils se mentionnent.

En résumé, le réseau égocentrique est construit de la façon suivante (cf. Figure 5.16) :

Nœuds : comptes twitter.

Relations : deux nœuds sont connectés entre eux si un nœud « *follow* » l'autre.

Informations : *hashtags* contenus dans les tweets.

Interactions : « *reply* », « *retweet* » et « *mention* ».

Granularité de temps : jour.

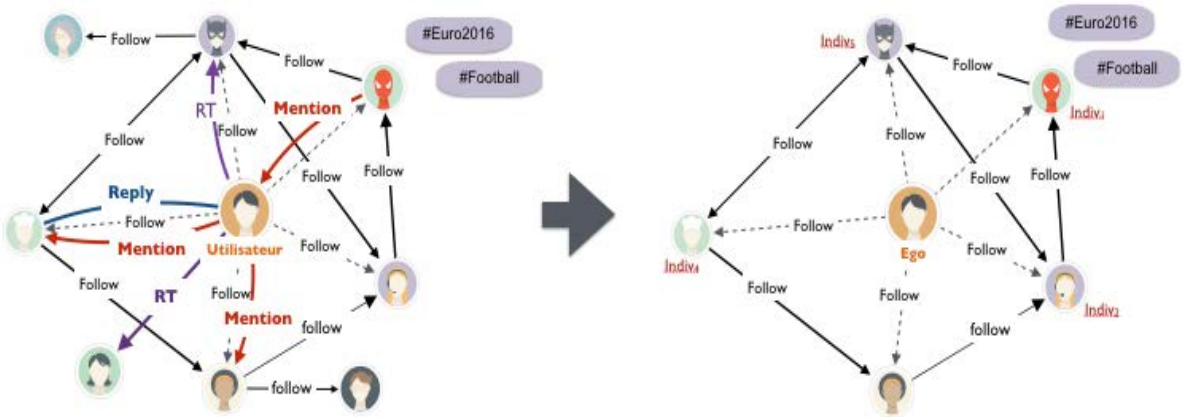


Figure 5.16 Extraction du réseau égocentrique de l'utilisateur dans Twitter

5.3.2.2. Accès aux donnés et présentation du dataset

L'échantillon de test a été récupéré à travers l'API officielle[70] fournie par Twitter en utilisant la bibliothèque Java indépendante Twitter4J[71]. Nous commençons par collecter les utilisateurs qui sont membres dans les listes publiques twitter sélectionnées sur les différents sujets suivants : Analyse des réseaux sociaux, Technologie, Jeux vidéo, Sport, Voyage.

Les listes sont généralement créées par un utilisateur pour suivre les informations partagées par les membres de la liste. Les membres d'une liste sont donc généralement ceux qui sont populaires et qui partagent des informations pertinentes pour le sujet concerné. Nous pouvons donc extraire les informations significatives depuis leurs tweets. Ceci nous permet d'éviter en grande partie les comptes de spam.

Notre échantillon de test contient 94 utilisateurs ayant entre 50 et 250 *followings*. La moyenne du nombre de *followings* de tous les utilisateurs (taille du réseau) est de 133. Au total, nous avons récupéré 10105 utilisateurs. En raison de la limite du nombre de tweets récupérables imposée par l'API, nous pouvons accéder à seulement 3200 tweets (les plus récents) pour chaque utilisateur. Les tweets collectés sont datés entre février 2009 et janvier 2016. Nous avons récupéré 6027624 tweets qui contiennent au moins un *hashtag*.

5.3.2.3. Evaluation

Nous présentons dans la Figure 5.17 notre protocole d'évaluation complet dans Twitter. Cette figure présente le protocole d'évaluation général (cf. Figure 5.1) instancié dans le contexte de Twitter.

[70] dev.twitter.com
[71] twitter4j.org

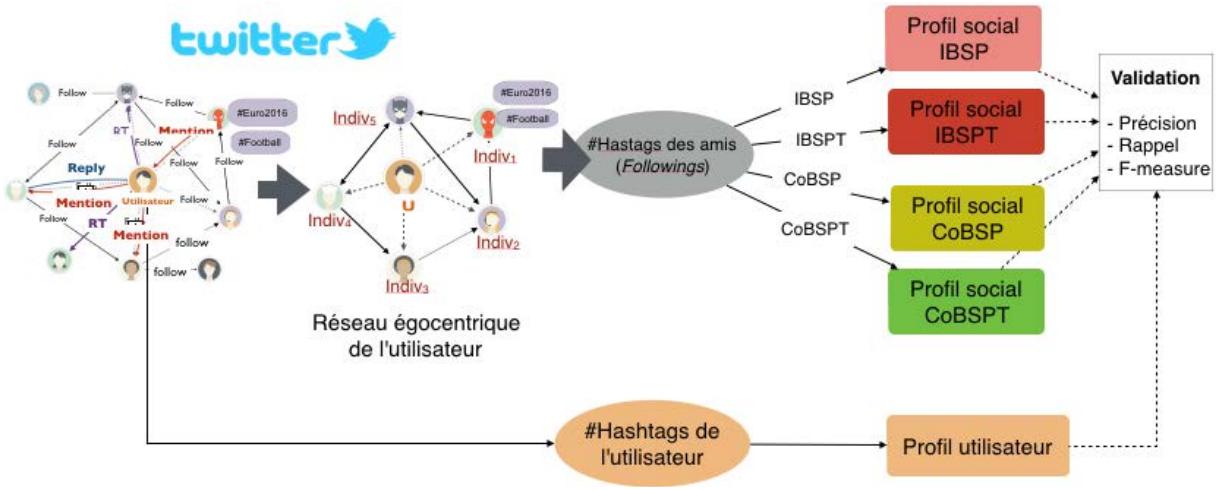


Figure 5.17 Protocole d'évaluation dans le réseau Twitter

Nous présentons dans ce qui suit, la construction du profil utilisateur et la construction du profil social dans cette évaluation.

a. Vérité de terrain (profil utilisateur)

La vérité de terrain dans Twitter peut être retrouvée à travers les intérêts indiqués par l'utilisateur dans son profil. Cependant, d'après les analyses de profil que nous avons effectuées, les utilisateurs dans Twitter ne mettent que très rarement leurs intérêts dans leur profil. Il est donc difficile dans ce contexte, d'extraire le profil explicite utilisateur comme vérité terrain. Nous avons donc construit le profil utilisateur de façon implicite par l'analyse des *hashtag*s dans les tweets de l'utilisateur. Nous considérons chaque *hashtag* comme un intérêt de l'utilisateur et extrayons les *hashtags* depuis ses tweets puis nous appliquons un dictionnaire thesaurus pour lemmatiser les *hashtags* qui ont le même sens (synonymes).

Ensuite, pour avoir un profil utilisateur significatif et à jour, nous prenons en compte un critère temporel en pondérant les intérêts dans le profil utilisateur en utilisant le « *a time-sensitive strategy* » proposé par (Abel et al., 2011) qui ont prouvé l'efficacité de cette stratégie dans Twitter par rapport aux approches qui ne prennent pas en compte le temps.

Dans (Abel et al., 2011), le profil d'un utilisateur *u* est un ensemble des concepts pondérés représentés par une entité ou un *hashtag*. Dans notre travail, nous considérons seulement les *hashtags* comme intérêts de l'utilisateur. Le profil de *u* est représenté comme suit :

$$P(u, time) = \{(h, w(h, time, T_{tweets,u}) | h \in CH\}$$

(5.4)

w(h, time, T_{tweets}, u) est une fonction temporelle qui permet de calculer, pour un instant *time* donné, le poids associé à un *hashtag h* contenu dans les tweets de u en se basant sur les *timestamp* T_{tweets} des tweets de *u* qui contiennent le *hashtag h*. *CH* représente l'ensemble des *hashtags* de *u*.

Cette fonction temporelle a pour objectif d'affaiblir la valeur de la fréquence d'une occurrence d'un *hashtag h* en se basant sur la distance temporelle entre le *timestamp* de cette occurrence de h et un *timestamp (time)* donné (généralement le *timestamp* actuel). La distance temporelle normalisée est calculée avec la formule (5.5).

$$normalized_{time}(tw, h) = \frac{|time - time(tw)|}{Max_{time} - Min_{time}}$$

(5.5)

Pour un tweet *tw* donné contenant le *hashtag h*, *time(tw)* représente son *timestamp*. Max_{time} et Min_{time} désignent respectivement le *timestamp* maximal (le temps le plus récent) et le *timestamp* minimal (le temps le plus ancien) des tweets dans $T_{tweets,u,h}$.

Le score temporel final d'un *hashtag h* est calculé comme suit :

$$w(h, time, T_{tweets,u}) = \sum_{t \in T_{tweets,u,h}} (1 - normalized_{time}(tw, h))^d$$

(5.6)

Le paramètre *d* est utilisé pour ajuster l'influence de la distance temporelle sur le poids temporel des *hashtags*. Il permet de donner plus ou moins d'importance à l'ancienneté des *hashtags* (taux de dépréciation, équivalent à λ dans notre travail). En se basant sur la valeur optimale trouvée dans le travail de (Abel et al., 2011), nous fixons *d* à 4.

Nous avons légèrement modifié l'équation (5.5) pour notre contexte. En effet, comme nous considérons seulement les *hashtags* pour construire le profil utilisateur, il se peut qu'il y ait des *hashtags* qui n'apparaissent qu'une seule fois dans le « *timeline* » de l'utilisateur. Dans ce cas, la distance entre max_{time} et $min_{time}(max_{time} - min_{time})$ devient 0, ce qui retourne une valeur infinie pour l'équation (5.5).

De plus, dans le cas de *hashtags* partagés pendant une courte période, la distance entre max_{time} et min_{time} peut être très petite. Cette valeur peut également être plus petite que la distance entre la *timestamp* donné *time* et le *timestamp* du *tweet time(tw)* (*time-time(tw)*). Ce qui induit que $normalized_{time}$ peut être supérieur à 1 et donc, $1 - normalized_{time}$ peut devenir négatif. Par conséquent, le poids final pourrait être non significatif. Pour pallier ce problème, nous modifions la formule pour calculer $normalized_{time}$ comme suit :

$$normalized'_{time} = \frac{|time - time(t)| + 1}{(Max_{time} - Min_{time}) + 1}$$

(5.7)

Nous avons également normalisé la valeur de $normalized'_{time}$ pour qu'elle soit comprise entre 0 et 1 comme suit :

$$w(h, time, T_{tweets,u}) = \sum_{t \in T_{tweets,u,h}} (1 - \frac{normalized'_{time}}{Max\ (NORM)})^d$$

(5.8)

NORM représente l'ensemble des poids $normalized'_{time}$ de tous les *hashtags* trouvés dans les tweets de l'utilisateur.

b. Construction du profil social avec l'étude paramétrique

La méthodologie est similaire à celle de DBLP mais au lieu d'utiliser tous les « *tweets* » et « *retweets* » complets pour extraire les mots-clés, nous extrayons seulement les *hashtags* depuis les tweets partagés dans le réseau égocentrique de l'utilisateur comme source d'informations.

L'expérimentation a été programmée dans le même environnement que celui utilisé pour DBLP (cf. section 5.3.1.3.b). Les données correspondent à 2 GO d'espace disque.

5.3.2.4. Résultats

Cette section présente les résultats de nos expérimentations. La procédure d'évaluation consiste à comparer les résultats en appliquant la méthode temporelle proposée avec les résultats ne prenant pas en compte le temps sur l'approche basée sur les individus et sur l'approche basée sur les communautés.

Tous les résultats présentés dans cette section sont calculés en prenant seulement les top 5 des intérêts. Notons que les algorithmes appliqués construisent des profils de potentiellement plusieurs dizaines ou centaines de *hashtags* et nous ne retenons ici que les 5 les plus significatifs. Notons également que dans Twitter, la taille de chaque top N pour le profil social et pour le profil utilisateur, sera identique dans notre expérimentation. Les résultats en termes de précision et de rappel sont par conséquent similaires. Le calcul de la F-mesure ne sera donc pas significatif. Dans cette partie nous ne présenterons donc que les résultats en termes de précision.

a. Résultats globaux

Nous présentons d'abord les résultats globaux de l'étude paramétrique pour les 94 utilisateurs. Les résultats sont calculés pour chaque combinaison de paramètres (α, λ et γ) et sont présentés par la précision moyenne pour l'échantillon c'est-à-dire par la moyenne de précision de tous les utilisateurs dans l'échantillon.

- Résultats sans la prise en compte du poids de centralité ($\alpha = 0$)

Comme dans le cas de DBLP, nous décidons d'étudier dans un premier temps, les résultats des processus de construction du profil social en fixant la valeur $\alpha = 0.0$. Cela montre les résultats sans prendre en compte le poids structurel (centralité de degré des communautés dans les cas des processus basés sur les communautés et centralité de degré des individus dans le cas de ceux basés sur les individus).

La Figure 5.18 ci-après présente la comparaison des résultats, en termes de précision des profils sociaux construits par les différentes techniques de calcul du poids temporel pour l'approche de construction de profil social basée sur les individus (processus IBSP et IBSPT) et pour celle basée sur les communautés (processus CoBSP et CoBSPT). Les résultats des chaque technique sont présentés par la précision moyenne correspondant à la meilleure combinaison de paramètres (γ et λ) ; la combinaison qui a obtenu la meilleure moyenne.

Pour distinguer les profils construits à partir de différentes techniques/fonctions temporelles, nous utilisons les mêmes notations pour les processus IBSPT et CoBSPT que dans le cas de DBLP (section 5.3.1.4.b).

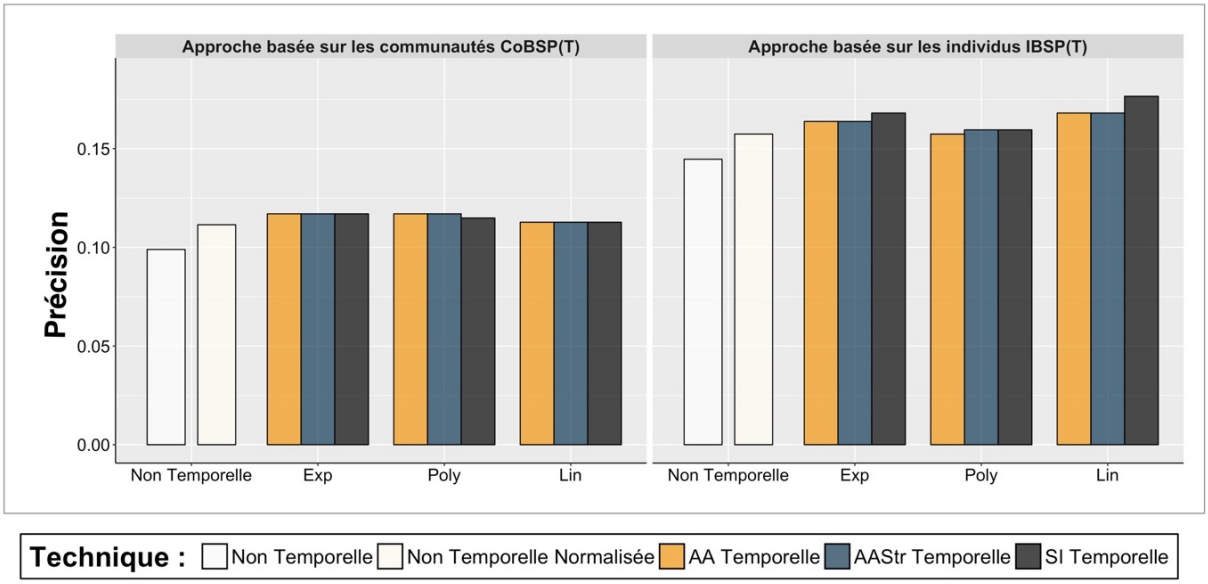


Figure 5.18 Comparaison de la meilleure précision moyenne, pour les résultats des processus de construction du profil social, en appliquant la méthode temporelle (IBSPT et CoBSPT) avec celle des processus qui ne prennent pas en compte le temps (IBSP et CoBSP)

Pour **l'approche basée sur les communautés**, avec le processus CoBSP, qui ne prend pas en compte le temps, la meilleure précision est 0.10296.

Pour le processus *CoBSPT*, nous présentons la précision par rapport à la technique utilisée : nous comparons les résultats selon la méthode temporelle appliquée et pour chacune, selon la technique de calcul du poids temporel des intérêts.

- Pour les techniques appliquant la fonction temporelle exponentielle (Exp) :
    - la technique ExpAATemp produit la meilleure précision (0.10957) lorsque $\lambda = 0.001$ et $\gamma = 0.95$ avec 6.429 % d'amélioration par rapport au processus CoBSP.
    - la technique ExpAAStrTemp produit la meilleure précision (0.10745) lorsque $\lambda = 0.001$ et $\gamma \in \{0.0, 0.001, 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.95\}$ avec 4.363 % d'amélioration par rapport au processus CoBSP.
    - la technique ExpSITemp produit la meilleure précision (0.10745) lorsque $\lambda = 0.001$ et $\gamma \in \{0.0, 0.001, 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.95\}$ avec 4.363 % d'amélioration par rapport au processus CoBSP.
- Pour les techniques appliquant la fonction temporelle polynomiale (Poly) :
    - la technique PolyAATemp produit la meilleure précision (0.10816) lorsque $\lambda \in \{0.25, 0.5, 0.75, 0.95, 1.0\}$ et $\gamma = 0.95$ avec 5.052 % d'amélioration par rapport au processus CoBSP.
    - la technique PolyAAStrTemp produit la meilleure précision (0.10816) lorsque $\lambda \in \{0.001, 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.95, 1.0\}$ et $\gamma = 0.95$ avec 5.052 % d'amélioration par rapport au processus CoBSP.
    - la technique PolySITemp produit la meilleure précision (0.1039) $\lambda \in \{0.001, 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.95, 1.0\}$ et $\gamma \in \{0.0, 0.001, 0.01,$

0.05, 0.1, 0.25, 0.5, 0.75, 0.95} avec 0.918 % d'amélioration par rapport au processus CoBSP.

- Pour les techniques appliquant la fonction temporelle linéaire (Lin) :
    - la technique LinAATemp produit la meilleure précision (0.09149) lorsque $\gamma \in \{0.0, 0.001, 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.95\}$ avec une perte de 11.137 % par rapport au processus CoBSP.
    - la technique LinAAStrTemp produit la meilleure précision (0.09149) $\gamma \in \{0.0, 0.001, 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.95\}$ avec une perte de 11.137 % par rapport au processus CoBSP.
    - la technique LinSITemp produit la meilleure précision (0.09149) lorsque $\gamma \in \{0.0, 0.001, 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.95\}$ avec une perte de 11.137 % par rapport au processus CoBSP.

Pour **l'approche basée sur les individus**, pour le processus IBSP (non prise en compte du temps), le meilleur résultat en termes de précision est 0.15106.

- Pour les techniques appliquant la fonction temporelle exponentielle (Exp) :
    - la technique ExpAATemp produit la meilleure précision (0.16383) lorsque $\lambda = 0.05$ et $\gamma \in \{0.01, 0.05\}$ avec 8.451 % d'amélioration par rapport au processus IBSP.
    - la technique ExpAAStrTemp produit la meilleure précision (0.1617) lorsque $\lambda \in \{0.05, 0.1\}$ et $\gamma \in \{0.0, 0.001, 0.01, 0.05, 0.1, 0.25\}$ avec 7.042 % d'amélioration par rapport au processus IBSP.
    - la technique ExpSITemp produit la meilleure précision (0.16383) lorsque $\lambda \in \{0.05, 0.1, 0.25\}$ et $\gamma \in \{0.1, 0.25, 0.5, 0.75\}$ avec 8.451 % d'amélioration par rapport au processus IBSP.
- Pour les techniques appliquant la fonction temporelle polynomiale (Poly) :
    - la technique PolyAATemp produit la meilleure précision (0.15532) lorsque $\lambda \in \{0.0, 0.001, 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.95, 1.0\}$ et $\gamma = 0.25$ avec 2.817 % d'amélioration par rapport au processus IBSP.
    - la technique PolyAAStrTemp produit la meilleure précision (0.15957) lorsque $\lambda = 1$ et $\gamma = 0.25$ avec 5.634 % d'amélioration par rapport au processus IBSP.
    - la technique PolySITemp produit la meilleure précision (0.15957) lorsque $\lambda \in \{0.1, 0.25, 0.5, 0.95, 1.0\}$ et $\gamma = 0.5$ avec 5.634 % d'amélioration par rapport au processus IBSP.
- Pour les techniques appliquant la fonction temporelle linéaire (Lin) :
    - la technique LinAATemp produit la meilleure précision (0.16809) lorsque $\gamma = 0.1$ avec 11.268 % d'amélioration par rapport au processus IBSP.
    - la technique LinAAStrTemp produit la meilleure précision (0.16809) lorsque $\gamma = 0.05$ avec 11.268 % d'amélioration par rapport au processus IBSP.
    - la technique LinSITemp produit la meilleure précision (0.1766) lorsque $\gamma = 0.5$ avec 16.901 % d'amélioration par rapport au processus IBSP.

Nous remarquons que pour les deux approches de construction du profil social (basée sur les individus ou sur les communautés), en appliquant les fonctions temporelles exponentielle et polynomiale, la méthode temporelle proposée produit de meilleurs résultats par rapport à ceux des processus qui ne prennent pas en compte le temps, quelle que soit la technique de calcul du poids temporel utilisée. Avec la fonction temporelle linéaire inverse, la méthode temporelle produit de plus mauvais résultats quelle que soit la technique de calcul du poids temporel utilisée pour l'approche basée sur les communautés. A l'inverse, pour l'approche de basée sur les individus, la fonction linéaire produit de meilleurs résultats par rapport aux autres fonctions temporelles quelle que soit la technique utilisée. L'approche basée sur les communautés donne de moins bons résultats avec une perte très importante (presque 100% pour chaque technique) par rapport à l'approche basée sur les individus quelle que soit la technique et les fonctions temporelles utilisées.

- Résultats en fonction de γ et λ

Pour montrer les résultats en fonction de γ et λ, la Figure 5.19 représente les résultats obtenus en termes de précision pour l'approche basée sur les individus IBSPT, où les valeurs de γ varient (axe horizontal) et où les différentes valeurs de λ sont représentées par différents courbes de couleurs pour les techniques appliquant les fonctions exponentielle et polynomiale.

Avec la fonction linéaire inverse qui ne dépend pas de la valeur de λ, les meilleurs résultats sont trouvés quand γ est fixé à 0.1 pour la technique AATemp, 0.05 pour la technique AAStrTemp et 0.5 pour la technique SITemp et au-delà de ces valeurs, les résultats deviennent de moins en moins bons.

Pour les fonctions exponentielle et polynomiale, les résultats varient de façon différente selon la technique appliquée, avec les valeurs de γ et λ fixées.

Pour les techniques appliquant la fonction temporelle exponentielle, les résultats en fixant la valeur de λ très bas (près de 0) sont meilleurs que ceux où λ est fixé plus haut (près de 1). Les courbes des résultats avec $\lambda = 1$ sont généralement plus basses que les autres courbes quelle que soit la valeur de γ. Les meilleurs résultats sont trouvés quand λ est fixé très bas. La valeur optimale de λ est généralement 0.05 et donne de meilleurs résultats par rapport à $\lambda=0$ quelles que soient la valeur de γ et la technique appliquée. En ce qui concerne γ, pour les techniques AATemp et AAStrTem, nous obtenons les meilleurs résultats quand γ est inférieur à 0.5. Pour la technique SITemp nous obtenons les meilleurs résultats quand γ est entre 0.1 et 0.75. Au-delà de ces valeurs, les résultats diminuent.

Pour les techniques appliquant la fonction temporelle polynomiale, nous ne voyons pas de grande différence entre les résultats en fonction de valeur de λ quand γ est fixé très bas (entre 0 et 0.25). Au-delà, nous voyons que les meilleurs résultats sont trouvés quand λ est fixé assez haut (0.95 ou 1). En ce qui concerne γ, pour les techniques AATemp et AAStrTem, nous obtenons les meilleurs résultats quand γ est inférieur à 0.25. Pour la technique SITemp, nous obtenons les meilleurs résultats quand $\gamma= 0.5$.

Nous pouvons observer globalement que, quand $\gamma = 1$, les résultats baissent dramatiquement quelles que soient la technique appliquée et la valeur de λ.

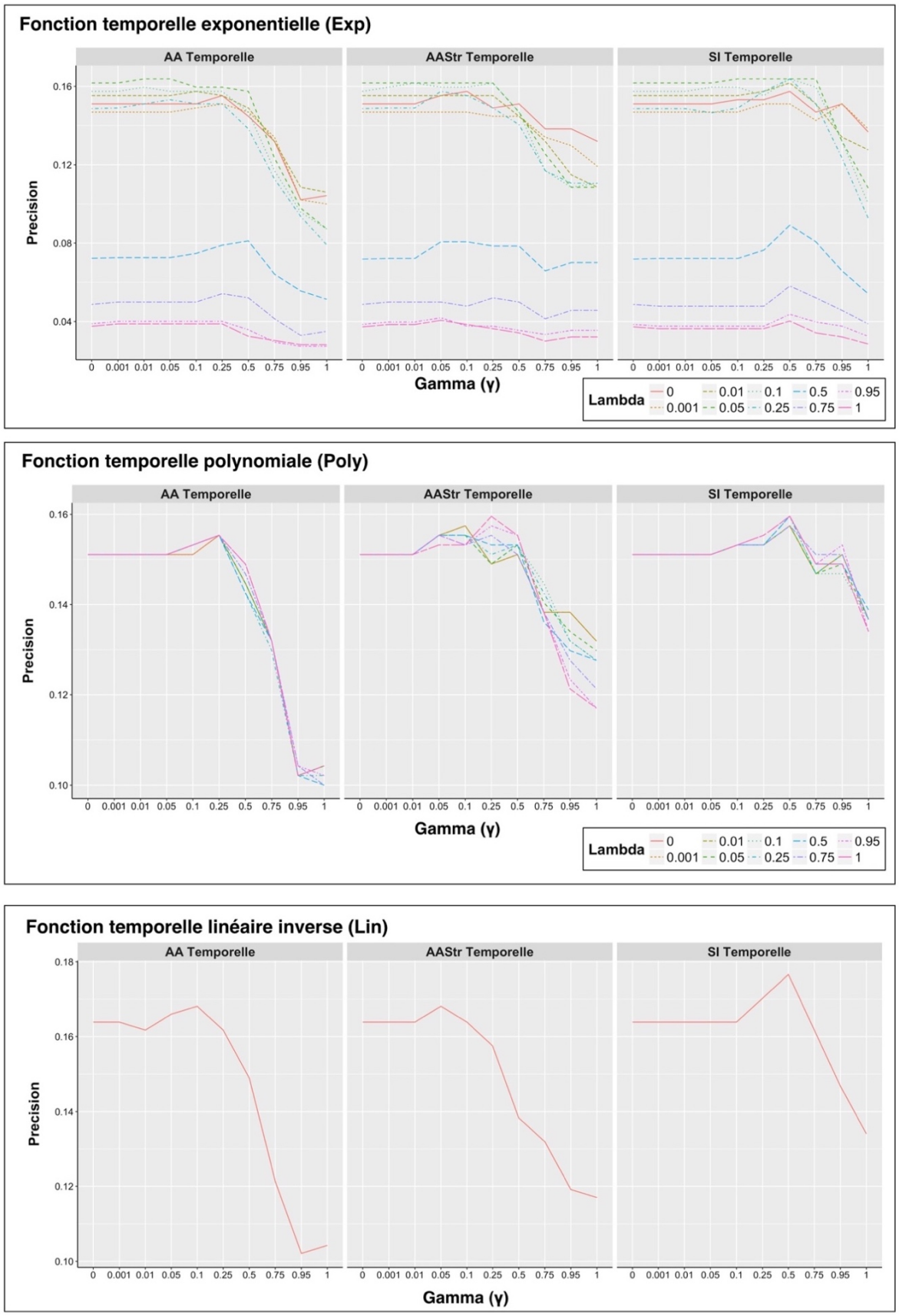


Figure 5.19 Comparaison des résultats en fonction de γ en axe horizontal et λ (différentes courbes) des processus basés sur les individus (IBSPT), pour les techniques appliquant les trois fonctions temporelles exponentielle, polynomiale et linéaire.

Nous montrons dans la Figure 5.20, les résultats en termes de précision pour l'approche basée sur les communautés (processus CoBPT), avec les valeurs de γ en axe horizontal et les valeurs de λ représentées par différentes courbes de couleurs pour les techniques appliquant la fonction exponentielle et polynomiale.

Avec la fonction linéaire inverse qui ne dépend pas de la valeur de λ, les résultats ne varient pas beaucoup en fonction de γ, sauf quand γ est = 1 où les résultats chutent dramatiquement.

Pour les fonctions temporelles exponentielle et polynomiale, avec la technique AATemp et AAStrTemp, les résultats restent stables quand γ est fixé entre 0 et 0.75, augmentent quand γ = 0.95 et rechutent quand γ = 1. Avec la technique SITemp, sauf dans le cas de γ = 1 les résultats restent stables quelle que soit la valeur de γ.

En ce qui concerne les valeurs de λ, nous observons une variation des résultats selon la fonction temporelle appliquée. Pour les techniques appliquant la fonction temporelle exponentielle, quand la valeur de λ est fixée très bas (près de 0), les résultats sont généralement meilleurs que quand la valeur de λ est fixée plus haut (près de 1). Les courbes des résultats avec λ = 1 sont généralement plus basses que les autres courbes quelle que soit la valeur de γ. Les meilleurs résultats sont trouvés quand la valeur de λ est fixée très bas. La valeur optimale de λ est 0.001 qui donne globalement de meilleurs résultats par rapport à ceux obtenus en fixant λ = 0, quelle que soit la valeur de γ.

Pour les techniques appliquant la fonction temporelle polynomiale, nous ne voyons pas clairement de différence entre les résultats en fonction de la valeur de λ. La valeur optimale de λ change en fonction de γ.

Au final, nous pouvons synthétiser les résultats en fonction de γ et λ comme suit :

Les valeurs optimales de γ qui montrent l'importance du poids temporel des individus par rapport au poids temporel des informations, sont : moins de 0.75 pour la technique SITemp et moins de 0.5 pour les autres techniques.

Nous observons également que les résultats varient en fonction de λ et la valeur optimale dépend de la technique utilisée mais aussi de la valeur de γ. Les valeurs optimales de λ généralement différentes de 0 montrent que la prise en compte du temps est importante pour améliorer les résultats des processus basés sur les individus même si le taux de dépréciation (λ) est assez bas dans certains cas.

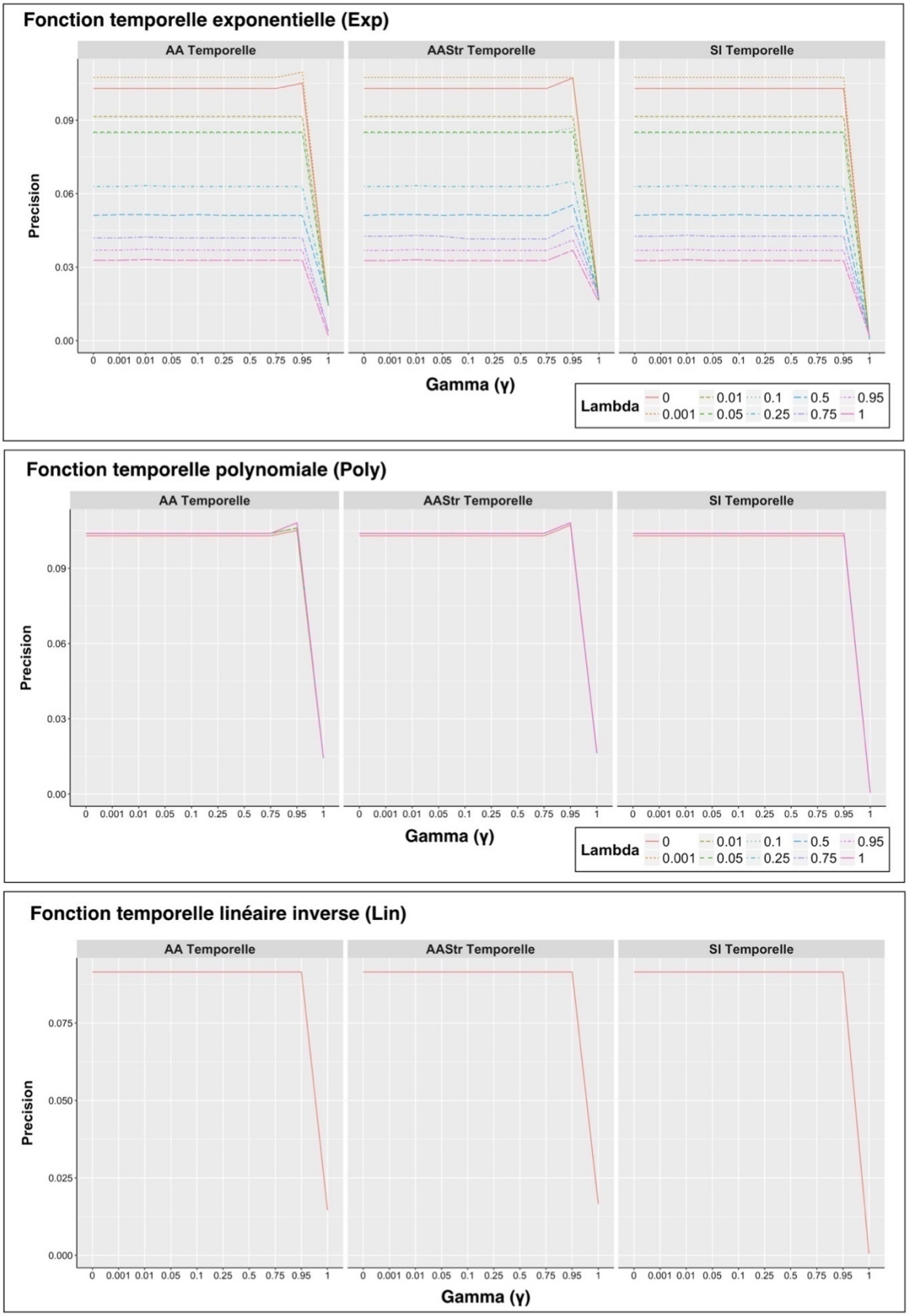


Figure 5.20 Comparaison en termes de précision des résultats en fonction de γ en axe horizontal et λ (différentes courbes) des processus basés sur les communautés (CoBSPT), pour les techniques appliquant les trois fonctions temporelles exponentielle, polynomiale et linéaire.

❖ Résultats en fonction de α

Nous avons ensuite, étudié les résultats en prenant en compte le poids de structure de centralité (en faisant varier différentes valeurs de α).

La Figure 5.21 représente les meilleurs résultats en termes de précision, correspondant à la meilleure combinaison de paramètres (γ, λ), des techniques pour l'approche basée sur les communautés (processus CoBSP et CoBSPT) et pour l'approche basée sur les individus (processus IBSP et IBSPT) en fonction des valeurs de α.

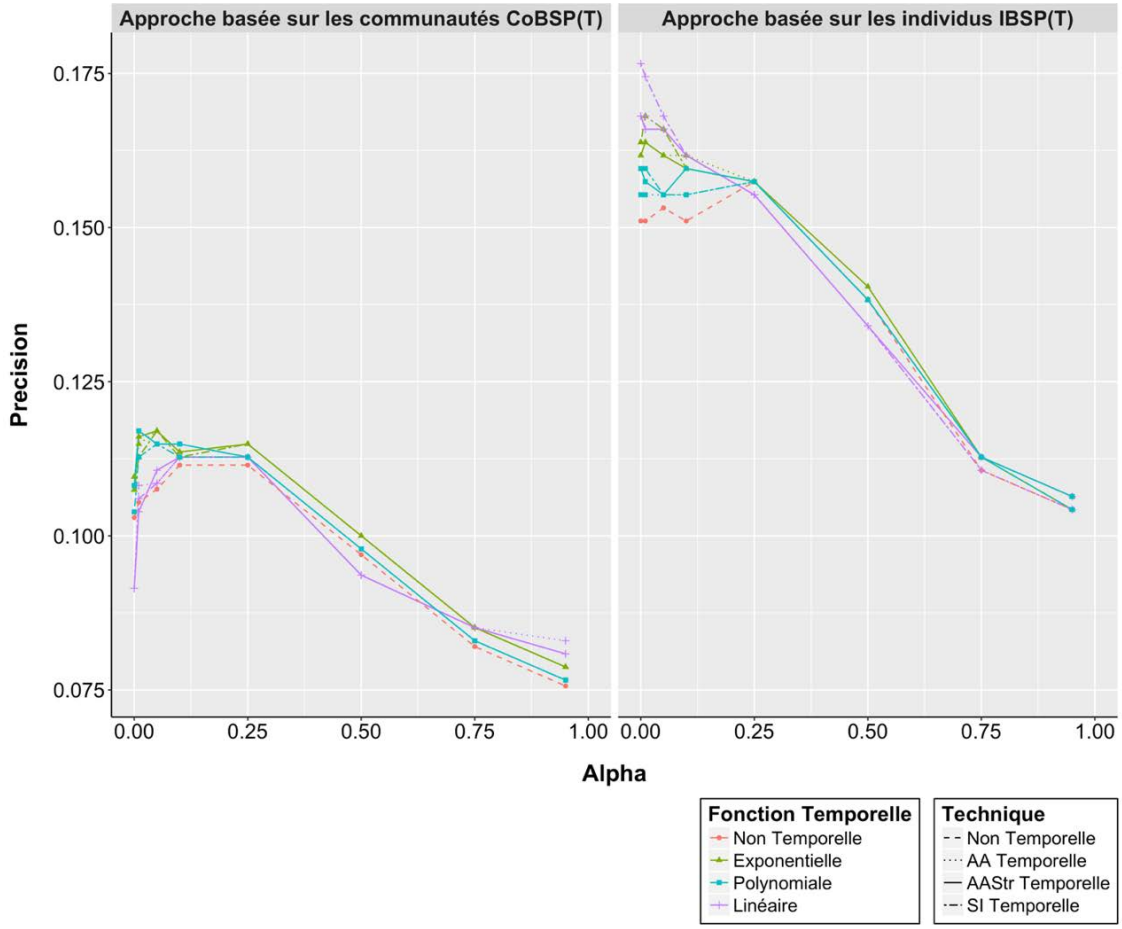


Figure 5.21 Comparaison de la meilleure précision moyenne pour les résultats des différentes techniques utilisées pour l'approche basée sur les communautés (processus CoBSP et CoBSPT) et pour l'approche basée sur les individus (processus IBSP et IBSPT) en fonction de la valeur de α.

Pour l'approche basée sur les communautés, le processus CoBSP (non prise en compte du temps) produit le meilleur résultat en termes de précision (0.11147) quand $\alpha \in \{0.1, 0.25\}$.

Pour le processus *CoBSPT*, nous présentons les résultats par rapport à la technique utilisée : nous comparons les résultats selon la méthode temporelle appliquée et pour chacune, selon la technique de calcul du poids temporel des intérêts.

- Pour les techniques appliquant la fonction temporelle exponentielle (Exp) :
    - la technique ExpAATemp produit la meilleure précision (0.11702) lorsque $\alpha = 0.05$, $\lambda = 0.001$ et $\gamma = 0.5$ avec 4.984 % d'amélioration par rapport au processus CoBSP.

- la technique ExpAAStrTemp produit la meilleure précision (0.11702) lorsque $\alpha = 0.05$, $\lambda = 0.001$ et $\gamma = 0.5$ avec 4.984 % d'amélioration par rapport au processus CoBSP.
- la technique ExpSITemp produit la meilleure précision (0.11702) lorsque $\alpha = 0.05$, $\lambda = 0.001$ et $\gamma = 0.5$ avec 4.984 % d'amélioration par rapport au processus CoBSP.

- Pour les techniques appliquant la fonction temporelle polynomiale (Poly) :
  - la technique PolyAATemp produit la meilleure précision (0.11702) lorsque $\alpha = 0.01$, $\lambda \in \{0.1, 0.25, 0.5\}$ et $\gamma = 0.95$ avec 4.984 % d'amélioration par rapport au processus CoBSP.
  - la technique PolyAAStrTemp produit la meilleure précision (0.11702) lorsque $\alpha = 0.01$, $\lambda \in \{0.001, 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.95, 1.0\}$ et $\gamma = 0.95$ avec 4.984 % d'amélioration par rapport au processus CoBSP.
  - la technique PolySITemp produit la meilleure précision (0.11489) lorsque $\alpha = 0.05$, $\lambda \in \{0.1, 0.25\}$ et $\gamma = 0.5$ avec 3.075 % d'amélioration par rapport au processus CoBSP.
- Pour les techniques appliquant la fonction temporelle linéaire (Lin) :
  - la technique LinAATemp produit la meilleure précision (0.11277) lorsque $\alpha \in \{0.1, 0.25\}$ et $\gamma \in \{0.0, 0.001, 0.01, 0.05, 0.1, 0.5\}$ avec 1.166 % d'amélioration par rapport au processus CoBSP.
  - la technique LinAAStrTemp produit la meilleure précision (0.11277) lorsque $\alpha \in \{0.1, 0.25\}$ et $\gamma \in \{0.0, 0.001, 0.01, 0.05, 0.1, 0.5\}$ avec 1.166 % d'amélioration par rapport au processus CoBSP.
  - la technique LinSITemp produit la meilleure précision (0.11277) lorsque $\alpha \in \{0.1, 0.25\}$ et $\gamma \in \{0.0, 0.001, 0.01, 0.05, 0.1, 0.5\}$ avec 1.166 % d'amélioration par rapport au processus CoBSP.

Pour l'approche basée sur les individus, le processus IBSP (non prise pas en compte du temps) produit les meilleurs résultats en termes de précision (0.15745) quand $\alpha = 0.25$.

- Pour les techniques appliquant la fonction temporelle exponentielle (Exp) :
  - la technique ExpAATemp produit la meilleure précision (0.16383) lorsque $\alpha \in \{0.0, 0.001, 0.01\}$, $\lambda = 0.05$ et $\gamma \in \{0.0, 0.001, 0.01, 0.05\}$ avec 4.054 % d'amélioration par rapport au processus IBSP.
  - la technique ExpAAStrTemp produit la meilleure précision (0.16383) lorsque $\alpha = 0.01$, $\lambda = 0.05$ et $\gamma \in \{0.0, 0.001, 0.01, 0.25\}$ avec 4.054 % d'amélioration par rapport au processus IBSP.
  - la technique ExpSITemp produit la meilleure précision (0.16809) lorsque $\alpha = 0.01$, $\lambda = 0.25$ et $\gamma = 0.5$ avec 6.757 % d'amélioration par rapport au processus IBSP.
- Pour les techniques appliquant la fonction temporelle polynomiale (Poly) :

- la technique PolyAATemp produit la meilleure précision (0.15745) lorsque $\alpha = 0.25$, $\lambda \in \{0.0, 0.001, 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.95, 1.0\}$ et $\gamma \in \{0.0, 0.001, 0.01\}$ avec 0 % d'amélioration par rapport au processus IBSP.
- la technique PolyAAStrTemp produit la meilleure précision (0.15957) lorsque $\alpha \in \{0.0, 0.001, 0.1\}$, $\lambda \in \{0.95, 1.0\}$ et $\gamma = 0.25$ avec 1.351 % d'amélioration par rapport au processus IBSP.
- la technique PolySITemp produit la meilleure précision (0.15957) lorsque $\alpha \in \{0.0, 0.001, 0.01\}$, $\lambda \in \{0.1, 0.25, 0.5, 0.95, 1.0\}$ et $\gamma = 0.5$ avec 1.351 % d'amélioration par rapport au processus IBSP.

- Pour les techniques appliquant la fonction temporelle linéaire (Lin) :
  - la technique LinAATemp produit la meilleure précision (0.16809) lorsque $\alpha \in \{0.0, 0.001, 0.01\}$ et $\gamma = 0.1$ avec 6.757 % d'amélioration par rapport au processus IBSP.
  - la technique LinAAStrTemp produit la meilleure précision (0.16809) lorsque $\alpha \in \{0.0, 0.001\}$ et $\gamma = 0.05$ avec 6.757 % d'amélioration par rapport au processus IBSP.
  - la technique LinSITemp produit la meilleure précision (0.1766) lorsque $\alpha \in \{0.0, 0.001\}$ et $\gamma = 0.5$ avec 12.162 % d'amélioration par rapport au processus IBSP.

Nous remarquons que dans les deux approches de construction du profil social (basée sur les individus ou sur les communautés), pour les techniques appliquant les fonctions temporelles exponentielle et polynomiale, la méthode temporelle proposée produit généralement de meilleurs résultats quelle que soit la technique de calcul du poids temporel utilisée et pour toutes les valeurs de α. Pour les techniques appliquant la fonction temporelle linéaire, la méthode temporelle proposée produit une baisse des résultats avec la technique AATemp pour l'approche basée sur les communautés mais produit de meilleurs résultats pour l'approche basée sur les individus. Les résultats ont tendance à être moins bons quand α diminue.

En comparant les résultats en fonction des poids temporels des intérêts, nous observons que, pour l'approche basée sur les communautés, la fonction temporelle exponentielle donne généralement de meilleurs résultats par rapport aux fonctions temporelles polynomiales et linaire. Cette dernière donne de mauvais résultats pour l'approche basée sur les communautés mais donne de meilleurs résultats sur l'approche basée sur les individus.

En comparant les techniques de calcul du poids, nous constatons que la technique AATemp donne généralement de meilleurs résultats par rapport aux techniques AAStrTemp et SITemp pour l'approche basée sur les communautés. Avec le processus basé sur les individus IBSPT, les meilleurs résultats sont obtenus en appliquant la technique SITemp.

En comparant les résultats des deux processus (*IBSPT* vs *CoBSPT*), le processus IBSPT produit globalement de meilleurs résultats avec un gain important.

Comme dans DBLP, nous supposons aussi que les résultats peuvent être différents en fonction de la taille et la densité du réseau. La sous-section suivante étudie ce point.

b. Résultats selon la taille et la densité du réseau

Dans le cas de DBLP, nous avons étudié les résultats selon la taille et la densité du réseau égocentrique de l'utilisateur pouvoir mieux observer les résultats par rapport au nombre de co-

auteurs des auteurs étudiés mais aussi par rapport à la densité de leur réseau égocentrique. Dans le cas de Twitter, nous ne possédons pas d'utilisateur ayant moins de 50 *following*s. Après plusieurs essais, nous avons donc découpé l'échantillon en plusieurs intervalles en fonction de la taille et la densité du réseau et appliqué l'étude paramétrique à chaque échantillon.

Pour chaque échantillon, les résultats pour chaque technique de calcul du poids temporel sont présentés par la précision moyenne pour l'échantillon, correspondant à la meilleure combinaison de paramètres (γ, λ, α); la combinaison qui a obtenu la meilleure moyenne.

Les résultats intéressants des échantillons étudiés se trouvent dans l'annexe. Dans Annexe 2-a sont présentés les résultats en fonction de taille du réseau et dans Annexe 2-b ceux en fonction de la densité du réseau.

Si l'on considère la taille du réseau, on a à étudier ici un réseau de grande taille et donc on ne peut pas faire varier la taille de réseau comme dans le cas de DBLP.

Nous donnons les résultats en fonction de la densité du réseau comme suit.

- Réseau de grande taille et épars

Cet échantillon représente les réseaux des 75 utilisateurs ayant plus de 50 individus et ayant moins de 10% de densité.

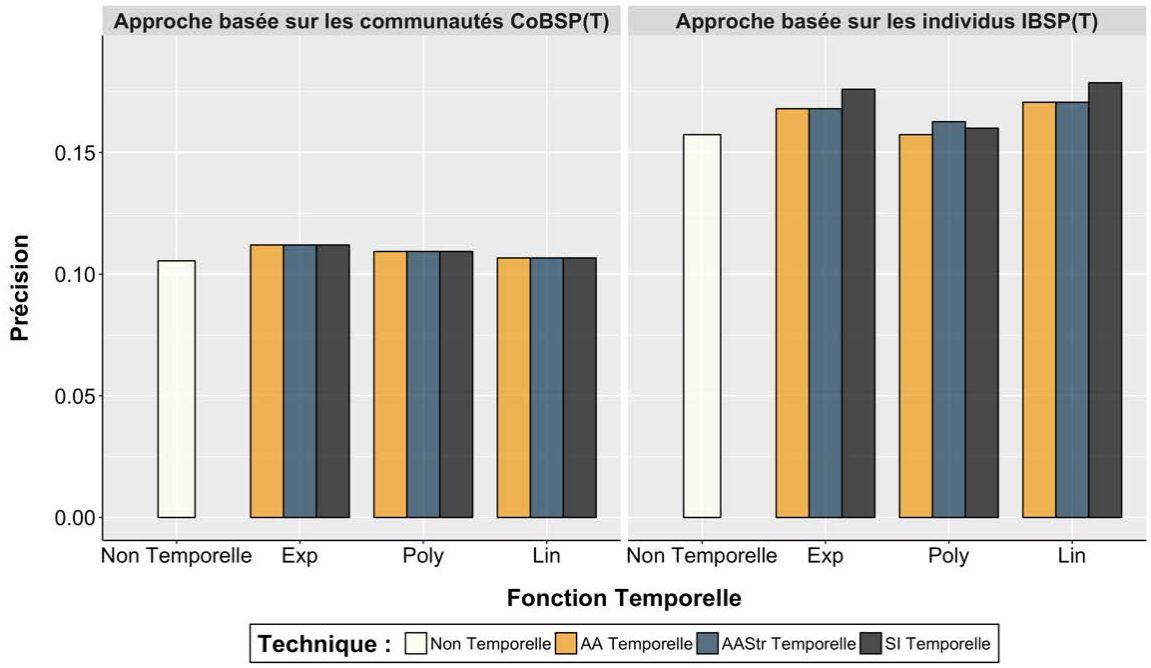


Figure 5.22 Comparaison de la meilleure précision moyenne des résultats obtenus pour les processus de construction du profil social, en appliquant la méthode temporelle proposée (IBSPT et CoBSPT) avec celle des résultats obtenus par les processus qui ne prennent pas en compte le temps (IBSP et CoBSP) dans l'échantillon des utilisateurs dans un réseau de grand taille et épars

Dans cet échantillon, pour l'approche basée sur les communautés, la méthode temporelle proposée améliore les résultats des processus existants avec les techniques appliquant la fonction temporelle exponentielle et produit de moins bons résultats avec les techniques appliquant la fonction linéaire. Le gain ou la perte ne sont pas importants. Pour l'approche basée sur les individus, les techniques appliquant les fonctions linéaire et exponentielle

améliorent globalement les résultats. Avec la fonction polynomiale, nous obtenons de moins bons résultats par rapport aux processus existants.

Nous n'observons pas de différence de résultats en fonction de la technique appliquée pour l'approche basée sur les communautés. Pour l'approche basée sur les individus, la technique SITemp produit généralement les meilleurs résultats.

Concernant l'approche, nous observons que dans cet échantillon, les résultats de l'approche basée sur les communautés sont moins bons (baisse importante) par rapport à ceux de l'approche basée sur les individus quelle que soit la technique utilisée.

- ❖ Réseau de grande taille et assez dense

Cet échantillon représente des réseaux des 19 utilisateurs ayant plus de 50 individus et ayant entre de 10% et 30% de densité.

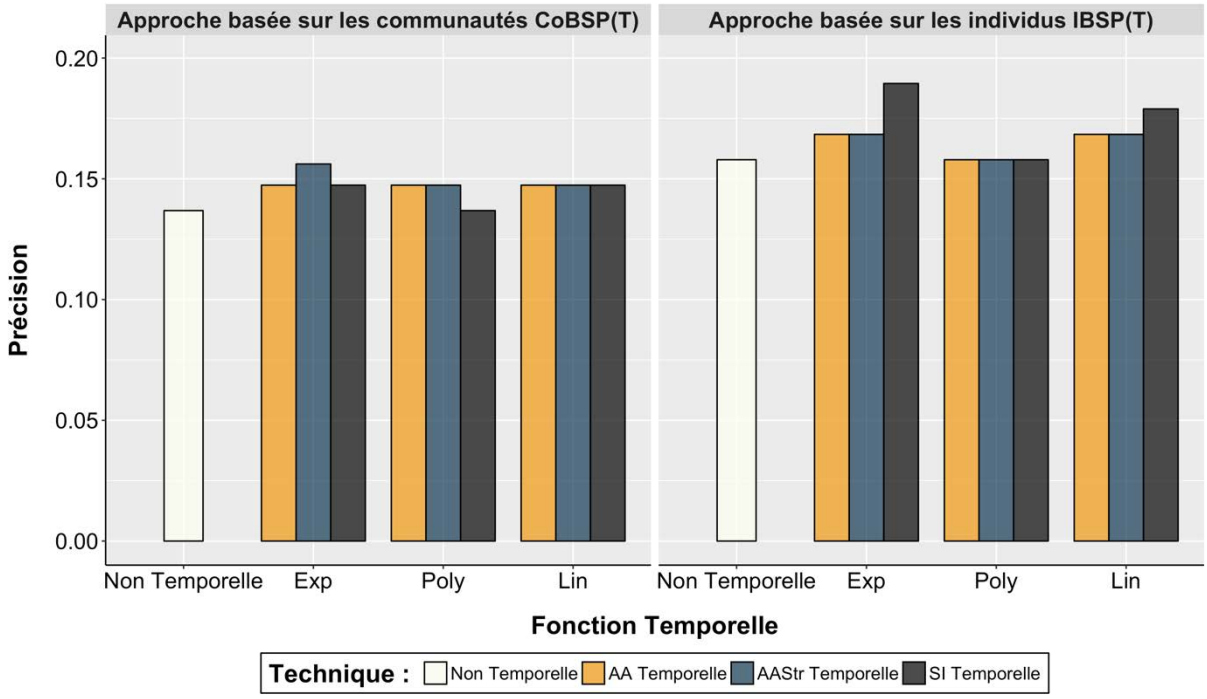


Figure 5.23 Comparaison de la meilleure précision moyenne pour les résultats des processus de construction du profil social, en appliquant la méthode temporelle proposée (IBSPT et CoBSPT) avec celle des résultats des processus qui ne prennent pas en compte le temps (IBSP et CoBSP) dans l'échantillon des utilisateurs dans des réseaux de grande taille et assez denses

Dans cet échantillon, pour l'approche basée sur les communautés, la méthode temporelle proposée améliore les résultats des processus existants quelle que soit la technique utilisée en appliquant la fonction temporelle exponentielle et produit de moins bons résultats avec les techniques appliquant la fonction linéaire. Les gains sont plus importants par rapport aux gains obtenus dans l'échantillon précédent. Pour l'approche basée sur les communautés, la technique AAStrTemp appliquant la fonction exponentielle donne de meilleurs résultats. Avec la fonction polynomiale, nous obtenons de moins bons résultats avec la technique SITemp. Avec la fonction temporelle linéaire, nous ne voyons pas de différence dans les résultats quelle que soit la technique appliquée.

Pour l'approche basée sur les individus, les meilleurs résultats sont obtenus en appliquant la technique SITemp avec la fonction exponentielle. En appliquant la fonction temporelle

polynomiale, la méthode temporelle proposée produit de moins bons résultats par rapport aux processus existants quelle que soit la technique appliquée.

Nous observons que dans cet échantillon, les résultats de l'approche basée sur les communautés sont moins bons que ceux de l'approche basée sur les individus quelle que soit la technique utilisée. Cependant les précisions moyennes obtenues pour toutes les techniques sont plus grandes que celles obtenues dans l'échantillon précédent. L'écart des résultats entre les deux approches devient également moins important.

c. Discussion des résultats

Dans Twitter, nous avons montré empiriquement que, globalement, les processus appliquant la méthode temporelle proposée (comportant différentes techniques de calcul du poids temporel des intérêts) donnent de meilleurs résultats que les processus existants quelle que soit l'approche de construction du profil social utilisée (basée sur les communautés ou sur les individus). En étudiant spécifiquement les résultats sur différents échantillons, nous trouvons de meilleurs résultats dans tous les échantillons étudiés.

Nous constatons que l'approche basée sur les individus est plus efficace que l'approche basée sur les communautés dans le contexte de Twitter. Le gain est important quand le réseau est peu dense et diminue quand le réseau est assez dense. Cela peut s'expliquer par le fait que dans l'approche basée sur les communautés, l'algorithme exploité pour extraire les communautés depuis le réseau égocentrique de l'utilisateur est basé sur les relations entre les membres du réseau. Or, la relation « *following* » n'est pas considérée pertinente pour extraire les communautés dans Twitter selon (Leicht et Newman, 2008 ; Liang et al., 2012). Cela peut donc entrainer des communautés non pertinentes et une mauvaise interprétation des résultats finaux.

En étudiant les résultats en fonction des techniques ou fonctions temporelles utilisées, nous trouvons que celles qui fournissent les meilleurs résultats varient en fonction de l'approche de construction du profil social utilisée mais aussi des échantillons étudiés.

La méthode temporelle linéaire qui diminue l'importance des informations ou des relations de façon linéaire donne généralement de moins bons résultats pour l'approche basée sur les communautés mais donne de bons résultats pour l'approche basée sur les individus.

Les valeurs de γ qui donnent plus ou moins d'importance à la prise en compte du poids temporel des individus par rapport au poids temporel des informations sont généralement bas (comprises entre 0.1 et 0.5).

Comme Twitter est un réseau qui possède des caractéristiques spécifiques et complexes, il pourrait y avoir également d'autres facteurs (par exemple le taux d'activité dans le réseau de l'utilisateur, la notoriété de ses amis, …) qui peuvent influencer la pertinence des processus de construction de profil présentés. Nous avons donc besoin de mener des études supplémentaires pour déterminer ces différents facteurs et pour trouver les meilleures combinaisons de ces facteurs dans le but d'obtenir des résultats plus pertinents.

5.4. Bilan des expérimentations des évaluations dans DBLP et Twitter

Les expérimentations menées dans DBLP et Twitter montrent l'efficacité de notre proposition par rapport à l'approche existante et montre par conséquent l'importance de la prise en compte de critères temporels non seulement dans la sélection des sources d'informations liées à la structure du réseau social de l'utilisateur mais aussi dans le traitement des informations partagées dans le réseau.

Les résultats obtenus dans DBLP sont généralement meilleurs que ceux obtenus dans Twitter (meilleure précision 0.36 contre 0.17). Cependant, dans Twitter, l'amélioration est plus significative quand on applique la méthode temporelle proposée. Cette observation peut être expliquée par la caractéristique de ces deux types de réseaux et les natures des liens considérés pour construire les réseaux égocentriques de l'utilisateur.

Dans le réseau DBLP, les relations considérées pour extraire les réseaux égocentriques de l'utilisateur se basent sur la co-publication. Ce type de relation peut fortement refléter l'influence sociale et la corrélation sociale entre les individus dans le réseau. Un auteur a tendance à publier avec les auteurs qui travaillent dans le même domaine et donc ont les mêmes intérêts. L'influence sociale a donc beaucoup d'importance dans ce type de réseau social. De plus, les intérêts dans ce type de réseau se limitent au domaine informatique. Il n'y a donc pas beaucoup de variation d'intérêts pour un utilisateur. Il est donc plus facile de trouver des mots-clés qui peuvent intéresser l'utilisateur principal à partir des informations partagées (dans ce cas, les titres de publications) par leurs co-auteurs.

Comme Twitter est un réseau de partage d'informations, les relations sociales entre individus sont moins importantes. Les relations « *following* » que nous considérons pour extraire les réseaux égocentriques de l'utilisateur peuvent ne pas être assez discriminantes pour refléter les influences sociales entre l'utilisateur et ses amis. De plus, les domaines des intérêts sur Twitter sont très variés par rapport à ceux de DBLP. Un utilisateur peut partager des informations sur le foot, la musique, la politique en même temps. Les informations partagées par un ami ne sont donc pas tout le temps significatives pour l'utilisateur. Un utilisateur peut s'intéresser au foot partagé par un de ses *followings* mais peut-être à d'autres informations. L'application de la méthode temporelle proposée permet dans un premier temps d'éliminer les individus qui ne sont pas pertinents pour l'utilisateur et de mettre en avant les informations qui sont partagées au moment d'une interaction entre l'utilisateur et ses amis. Ceci explique l'amélioration significative des résultats dans ce type du réseau. Cependant, comme la méthode temporelle proposée repose sur les interactions et que, dans Twitter, les interactions ne sont pas l'élément le plus important de la plateforme, il se peut qu'un utilisateur n'interagisse que très rarement avec ces *followings* ce qui fausse alors le calcul du poids des intérêts tel que nous l'avons proposé.

Pour obtenir de meilleurs résultats dans DBLP, la valeur optimale de γ doit être fixée très haut (0.75- 0.95) pour les deux approches de construction du profil social (individuel/communautés). Cela signifie que le poids temporel des individus est plus important que le poids temporel des informations dans ce dataset. Ceci montre l'importance de l'évolution de la force des relations entre l'utilisateur et les individus dans son réseau social par rapport à l'évolution des informations. Si nous nous focalisons sur l'observation des résultats obtenus dans le réseau de publications scientifiques, nous constatons que les relations entre les individus jouent un rôle plus important que l'évolution des informations. Ceci peut être expliqué par le fait que, dans le réseau de publications scientifiques, les auteurs ont tendance à rester sur le même domaine de recherche et à publier avec les mêmes personnes. Le changement de co-auteurs peut être lié à

un changement de domaine de recherche : un nouveau co-auteur peut indiquer un nouvel axe de recherche pour l'auteur. Par exemple, un auteur qui travaille dans le domaine de l'informatique et qui, à un moment donné, collabore avec un auteur qui travaille dans le domaine de la bio-informatique pourrait avoir un intérêt en bio-informatique en plus dans son profil.

Dans le cas de Twitter, la valeur optimale de γ est plus faible (entre 0.1 et 0.5) pour les deux approches de construction du profil. Cela signifie que le poids temporel des individus a moins d'importance dans le cas de Twitter. Cela peut provenir des caractéristiques du réseau Twitter qui est considéré comme un réseau de partage d'informations. Les informations s'échangent plus vite que sur le réseau DBLP. Le poids temporel des informations doit donc être pris en compte de manière plus importante que dans le cas de DBLP. Enfin, les relations twitter utilisées pour construire les relations entre personnes (« *followings* ») suggèrent un attachement pour « suivre l'information diffusée » par la personne d'où, sans doute, l'importance à accorder au poids de l'information.

D'après les résultats en fonction des λ et γ présentés dans la Figure 5.8 et la Figure 5.9 pour DBLP et dans la Figure 5.19 et la Figure 5.20 pour Twitter, nous observons le changement de valeur optimale de λ selon γ. Quand γ est fixé à 1 ou est à proximité de 1 (lors que le poids temporel de l'individu a beaucoup plus d'importance que le poids temporel de l'information), la valeur optimale de λ est différente de celle obtenue quand γ est fixé à 0 ou proximité de 0 (lors que le poids temporel de l'information a beaucoup plus d'importance que le poids temporel de l'individu). Cela nous emmène à penser qu'il serait judicieux de fixer différentes valeurs de λ dans le calcul du poids temporel des individus et celui du poids temporel des informations pour obtenir de meilleurs résultats. Cette hypothèse suit également les caractéristiques évolutives des relations et des informations dans un réseau social qui peuvent être différentes comme expliqué précédemment (sur DBLP les relations (co-publication) ont tendance à évoluer plus que les informations (sujets de recherche) alors que sur Twitter les sujets des informations partagées (*tweets*) évoluent plus que les relations (*following*)).

En étudiant les résultats en fonction des différentes techniques de calcul du poids temporel ou fonctions temporelles utilisées, nous constatons que, dans les deux types de dataset DBLP ou Twitter, la fonction temporelle et/ou la technique utilisée qui fournissent les meilleurs résultats varient selon l'approche utilisée mais aussi selon les échantillons étudiés. Ceci montre l'importance de la sélection des approches de construction du profil, des techniques de calcul du poids temporel et de la fonction temporelle appliquée selon les caractéristiques du réseau (taille et densité).

Dans DBLP, l'approche basée sur les communautés peut produire de meilleurs résultats par rapport à l'approche basée sur les individus dans un réseau de taille grande et assez dense. Dans Twitter, les approches basées sur les communautés donnent de meilleurs résultats dans un réseau grand et assez dense mais produisent toujours de moins bons résultats par rapport à l'approche basée sur les individus. Ceci peut provenir de la nature des liens dans Twitter qui est basée sur les relations « *following* » qui n'est peut-être pas assez discriminante pour extraire les communautés pertinentes.

6. CONCLUSION ET PERSPECTIVES

6.1. Conclusion

Nos travaux se situent dans le contexte de la construction du profil utilisateur, élément essentiel dans un système de personnalisation d'informations et dont les domaines d'application sont très nombreux (recommandation, recherche d'information personnalisée, …).

Le profil de l'utilisateur est généralement construit à partir de l'historique des activités de l'utilisateur (approche classique). Ces dernières années, plusieurs travaux ont proposé d'exploiter les informations à partir du réseau social de l'utilisateur pour construire ce profil. Le terme « profilage social » désignant les techniques associées à cette approche de construction de profil, nous avons utilisé le terme « profil social » pour représenter le profil construit à partir de cette approche et utilisé comme un profil complémentaire au profil utilisateur existant ou remplaçant le profil utilisateur dans le cas où celui-ci s'avère vide ou trop lacunaire.

Notre objectif dans cette thèse était la proposition d'une technique de profilage social efficace, intégrant donc l'évolution des relations et des informations dans le réseau social qui peut générer des biais d'informations et amener vers des sources d'informations non pertinentes en raison d'intérêts obsolètes. Prendre en compte le caractère hétérogène du réseau social nous a permis d'aborder le problème de la généricité du profilage proposé, une technique pouvant bien fonctionner dans un type de réseau social mais pas sur les autres.

Nous avons proposé une démarche générique de profilage social efficace permettant de retourner un profil social représentatif de l'utilisateur prenant en compte différents types de réseau ainsi que leurs caractéristiques évolutives.

Pour prendre en compte l'évolution des intérêts dans le profil social, nous avons proposé d'améliorer l'efficacité du processus de construction du profil social en intégrant la prise en compte de l'évolution du réseau social de l'utilisateur, via un facteur temporel ajouté aux processus existants de construction du profil social (approche basée sur des individus et approche basée sur les communautés). La solution développée permet de privilégier les intérêts provenant des informations significatives et à jour ; elle est basée sur une mesure temporelle dans l'étape d'extraction et de pondération des intérêts. Cette mesure est calculée d'une part, à partir de la pertinence temporelle des informations utilisées pour extraire cet intérêt et d'autre part, à partir de la pertinence temporelle de l'individu qui partage ces informations. La pertinence temporelle de l'information représente son importance, au moment t, vis-à-vis de l'utilisateur. De la même manière, la pertinence temporelle d'un individu représente l'importance de sa relation avec l'utilisateur, au moment t.

Eliciter la technique et la fonction temporelle appropriées selon le type et les caractéristiques du réseau social a conduit à proposer une méthode temporelle suffisamment générique pour être appliquée sur différents types du réseau.

Nous avons intégré cette méthode dans les deux processus de construction du profil social existants (approche basée sur les individus et approche basée sur les communautés), pour améliorer leur pertinence.

Nous avons mis en œuvre la méthode proposée sur différents réseaux sociaux : *DBLP* qui est un réseau de publications scientifiques et *Twitter* qui est un réseau de micro-blogs. Ces deux réseaux sociaux possèdent des caractéristiques différentes en termes d'objectif d'utilisation, de type d'informations partagées, de type de relations et interactions entre les individus dans le réseau et enfin en termes de caractéristiques d'évolution. Ces expérimentations ont permis d'étudier l'impact des caractéristiques du réseau social sur la pertinence du processus de construction du profil social.

Enfin, par rapport aux questions posées dans l'introduction générale :

- Comment construire un profil social à la fois pertinent et à jour dans le contexte des RSNs ?

Les résultats des expérimentations nous ont permis de montrer l'efficacité des processus intégrant la méthode temporelle proposée par rapport à ceux ne prennant pas compte des critères temporels. Ceci montre l'importance de la prise en compte de l'évolution du réseau social (temporalité) de l'utilisateur dans le processus de construction du profil social.

- Comment étudier l'influence des caractéristiques des RSNs sur la pertinence du profil social construit ?

Dans les expérimentations dans DBLP et Twitter, nous avons pu constater que, pour obtenir de meilleurs résultats, l'approche de la construction du profil, les techniques de calcul du poids temporel et les fonctions temporelles sont différentes selon le type et les caractéristiques du réseau social étudié. Ceci montre l'importance de la sélection de l'approche de construction du profil social, de la technique de calcul du poids temporel et de la fonction temporelle en fonction du type et des caractéristiques du réseau (taille et densité).

En conclusion, nous avons montré l'apport de notre contribution dans le processus de construction du profil social. Les méthodes et techniques proposées peuvent être encore améliorées mais nous espérons que cette préfiguration puisse ouvrir des pistes de recherche pour aller plus loin dans de futurs travaux.

6.2. Perspectives

La problématique abordée dans cette thèse étant relativement nouvelle dans le contexte des systèmes de personnalisation de l'information à l'utilisateur, plusieurs perspectives sont envisageables.

Un phénomène important qui peut avoir un impact sur la pertinence du profil social construit est l'apparition et la propagation de buzz ou de spams dans le réseau social. En effet, dans (Mezghani et al., 2015), nous avons montré que la construction du profil social peut permettre d'anticiper le « parasitage », ou « bruitage » lié à la propagation de buzz ou rumeurs dans un système de recommandation. Ne pas distinguer les rumeurs des vraies informations amène à les exploiter pour extraire des intérêts et peut conduire à proposer de « faux » intérêts dans le profil social construit, et par conséquent donner de fausses interprétations dans les mécanismes d'adaptation de l'information. Il serait nécessaire d'étudier une méthode permettant d'analyser la qualité des sources d'informations pour éviter de prendre en compte des informations « fausses » dans le processus de construction du profil social.

Sur un autre volet de travail, nous envisageons d'appliquer la méthode temporelle proposée, dans d'autres réseaux sociaux qui possèdent des caractéristiques différentes pour évaluer son efficacité. Nous envisageons d'étudier dans un premier temps StackOverflow, un site de questions/réponses qui ne représente pas explicitement des liens entre les utilisateurs mais dans lequel on peut extraire des informations importantes que l'on pourrait considérer comme des intérêts de l'utilisateur. Dans un second temps, nous étudierons un réseau de type Facebook qui, par opposition, est considéré comme un réseau social plus « orienté relations » que « partage d'informations » et montre plus explicitement et plus significativement les liens entre les utilisateurs.

A moyen terme, comme nous avons montré qu'il existe plusieurs facteurs qui peuvent impacter les résultats du processus de construction du profil social (la taille du réseau, la densité du réseau, la fonction temporelle appliquée, la technique de calcul du poids temporel utilisée, la valeur de gamma, lambda et alpha fixé,…), une autre perspective est d'étudier et d'appliquer des méthodes d'apprentissage qui permettraient de trouver les combinaisons optimales ou potentiellement par sélection des facteurs pertinents pour la construction du « meilleur » profil social : fonction temporelle/technique de calcul du poids/coefficients de calcul gamma, lambda et alpha.

Nous projetons aussi d'étudier le processus de mise à jour continue du profil social, afin d'avoir un profil pertinent et à jour à tout moment. Une fois le profil construit, le maintenir à jour en appliquant la même méthode de calcul complet à chaque fois pour prendre en compte les nouvelles informations partagées dans le réseau égocentrique de l'utilisateur est inenvisageable. Nous proposons donc de maintenir le profil à jour par un processus de mise à jour de profil, qui, à partir d'un profil déjà pertinent et à jour, ajustera à chaque mise à jour, les intérêts considérés pertinents en se basant sur les anciens intérêts de la période précédente. Les techniques de mise à jour de profil utilisateur présentées dans l'état de l'art peuvent être adaptées dans ce contexte. Comme dans l'étape de construction du profil, le processus de mise à jour ne sera pas centré sur l'utilisateur lui-même mais sur les informations partagées par les individus dans son réseau égocentrique et sur les relations de ces derniers avec l'utilisateur.

Nous pourrions envisager également d'évaluer les profils sociaux construits dans le cadre de notre proposition dans une application réelle, un système de recommandation par exemple (campagne de tests).

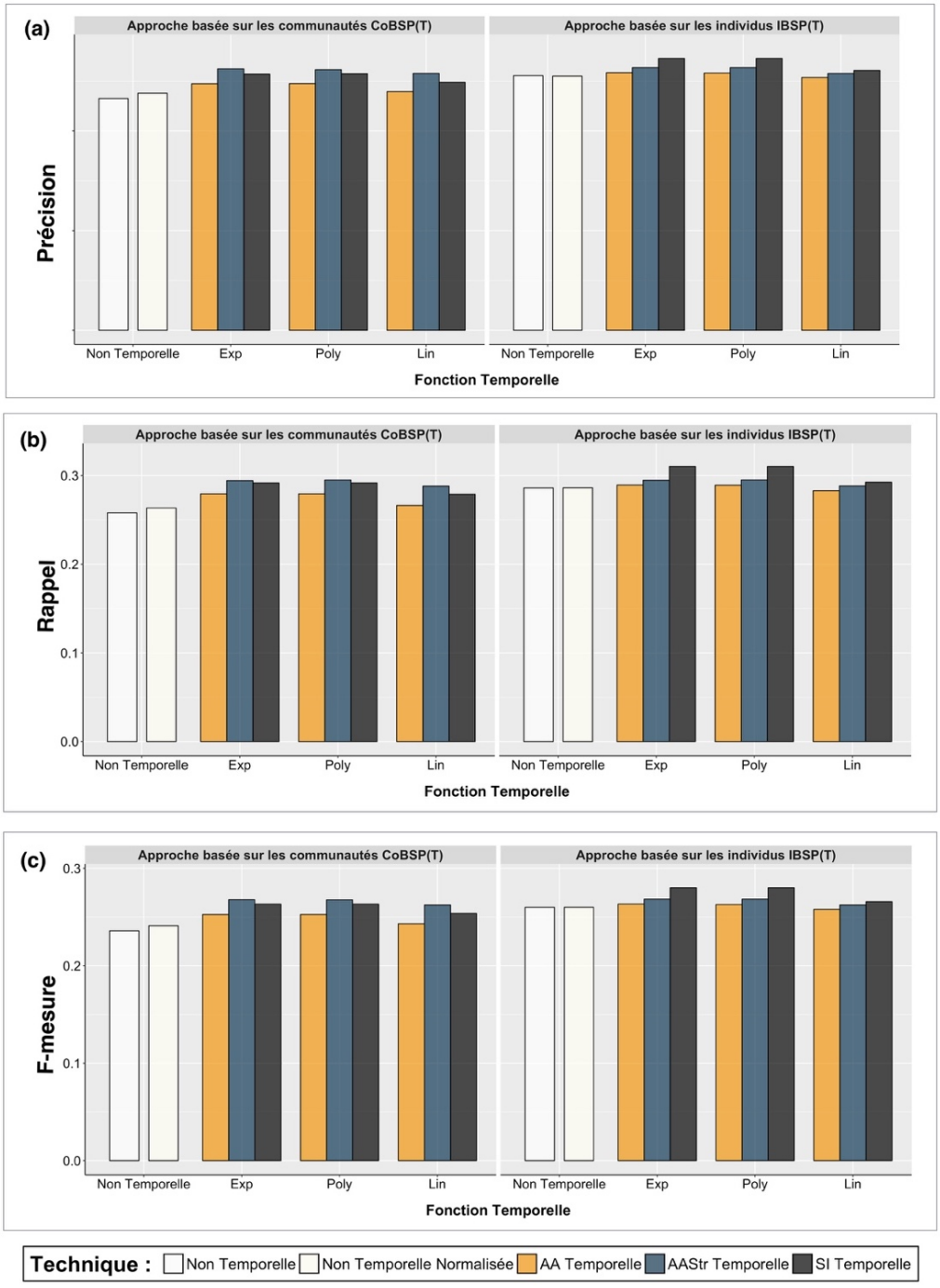
A plus long terme, il pourrait être intéressant de concevoir une plateforme de construction du profil social, s'appuyant sur nos travaux, qui permet en fonction d'un réseau social donné, de construire le profil en appliquant l'approche de construction du profil et la technique de calcul du poids temporel des intérêts appropriées. Le travail à réaliser s'appuiera sur la détection des techniques appropriées en analysant les caractéristiques du réseau social de chaque utilisateur, sur l'application de ces techniques pour construire son profil social et enfin sur la définition des métadonnées utilisées en fonction du type des informations disponibles dans le réseau étudié.

Enfin, nous envisageons de prendre en compte l'écosystème complet de l'utilisateur. En effet, dans le contexte des média sociaux, un utilisateur peut faire partie de plusieurs réseaux sociaux à la fois. Les utilisateurs mobilisent leurs réseaux sociaux à différents propos : Facebook pour contacter et partager des activités quotidiennes avec des amis proches, Twitter pour suivre des informations concernant leurs intérêts, Instagram pour partager des photos, vidéo, LinkedIn pour rester en contact et suivre des informations professionnelles. De ce fait, nous pouvons penser agréger les informations à partir de différents réseaux sociaux de l'utilisateur pour avoir une source de données plus complète mais aussi plus complexe, c'est-à-dire disposer d'intérêts à partir de différents domaines et selon différents points de vue, ce qui rendrait le profil social de l'utilisateur plus riche. La diversité des informations des différents réseaux ainsi que le volume de données extraites rendront sans doute le processus de traitement des informations beaucoup plus complexe. Ceci relève d'une problématique de passage à l'échelle (Volumétrie) et d'hétérogénéité des informations (Variété) qui nous projette dans les dimensions sociales du « big data » pour prendre en compte toutes celles d'un écosystème social généralisé.

ANNEXE

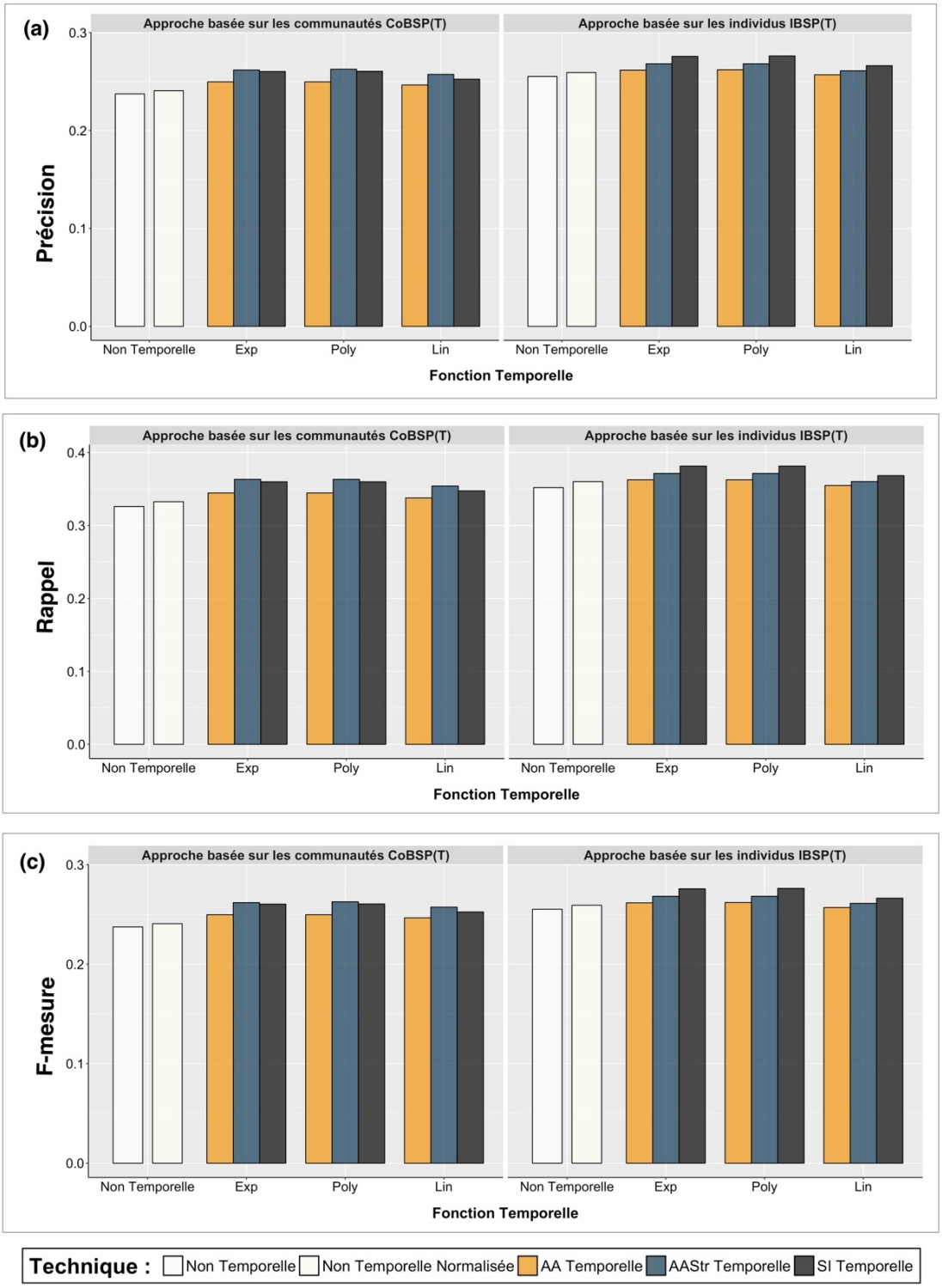
1. Résultats expérimentation DBLP

a. Résultats sur DBLP sur les top10



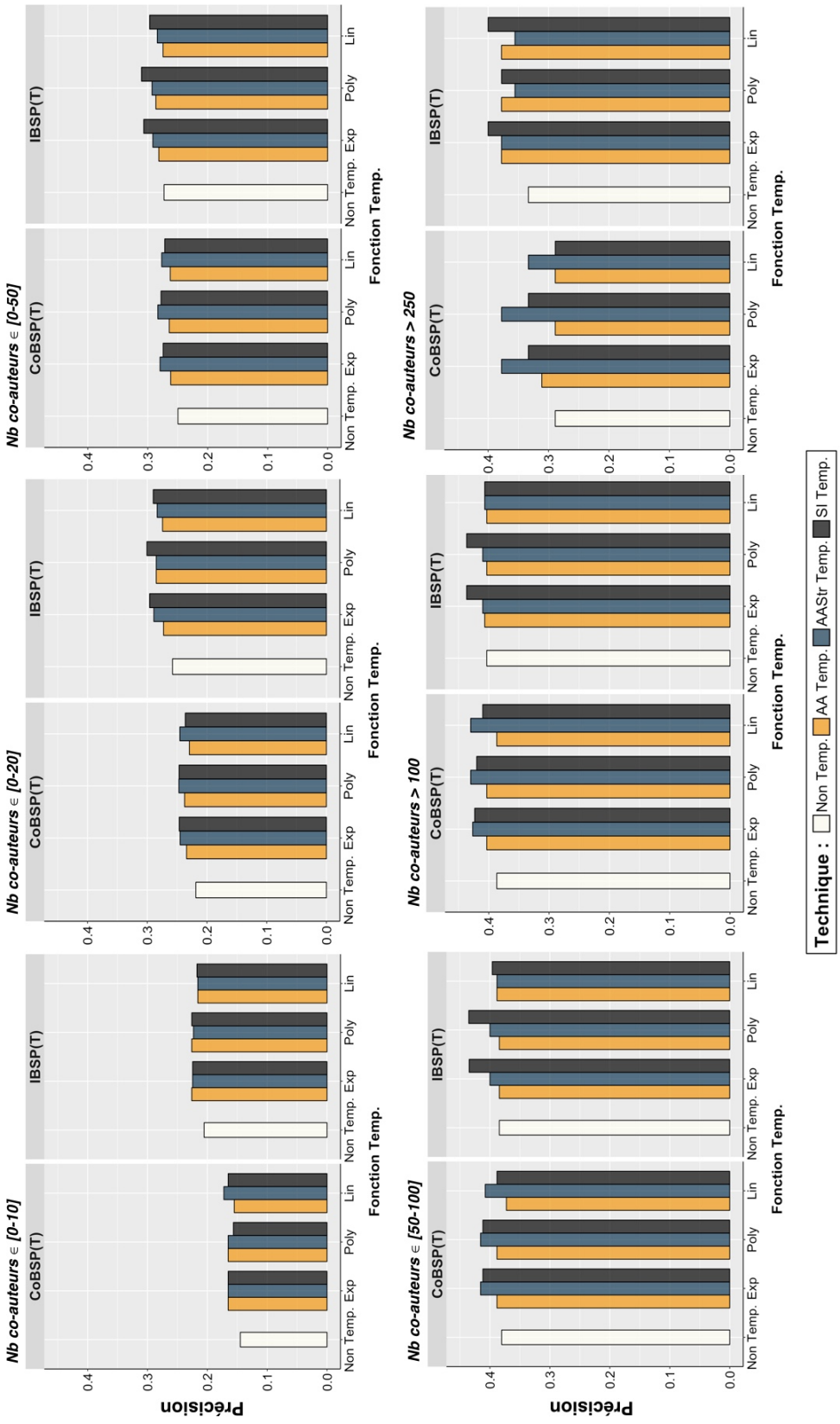
Annexe 1-a Comparaison de la meilleure précision moyenne (a), meilleur rappel moyen (b) et meilleure F-mesure moyenne (c), des résultats des processus de construction du profil social en appliquant les différentes méthodes temporelles aux processus (IBSPT et CoBSPT) avec ceux des processus qui ne prennent pas en compte les critères temporels (IBSP et CoBSP) sur les top 10 des intérêts

b. Résultats sur DBLP sur les top15



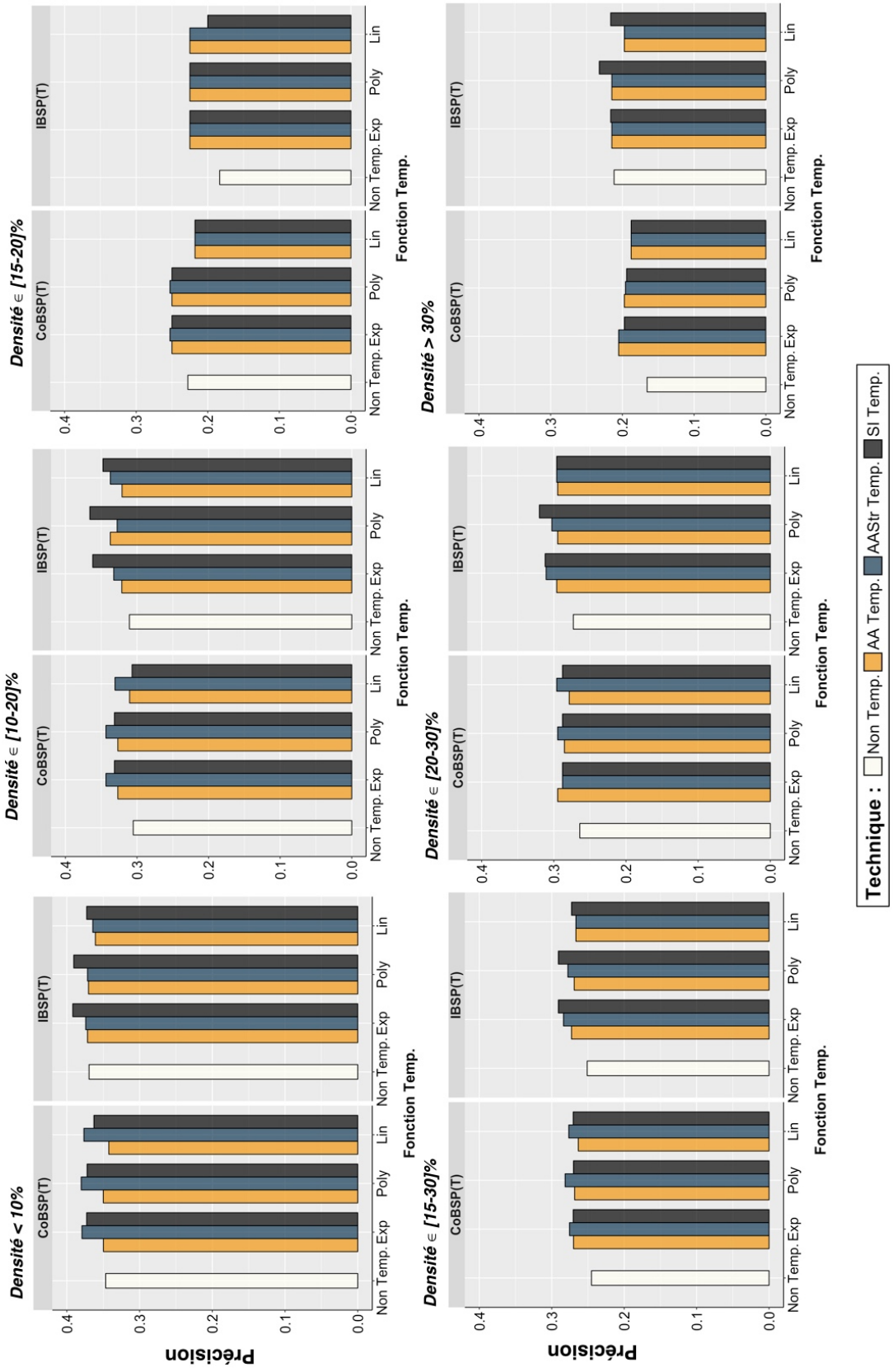
Annexe 1-b Comparaison de la meilleure précision moyenne (a), meilleur rappel moyen (b) et meilleure F-mesure moyenne (c), des résultats des processus de construction du profil social en appliquant les différentes méthodes temporelles aux processus (IBSPT et CoBSP) avec ceux des processus qui ne prennent pas en compte les critères temporels (IBSP et CoBSP) sur les top 15 des intérêts

c. Résultats sur DBLP selon la taille du réseau (nb co-auteurs)



Annexe 1-c Comparaison de la meilleure précision moyenne des résultats des processus de construction du profil social en appliquant les différentes méthodes temporelles aux processus (IBSPT et CoBSPT) selon la taille du réseau

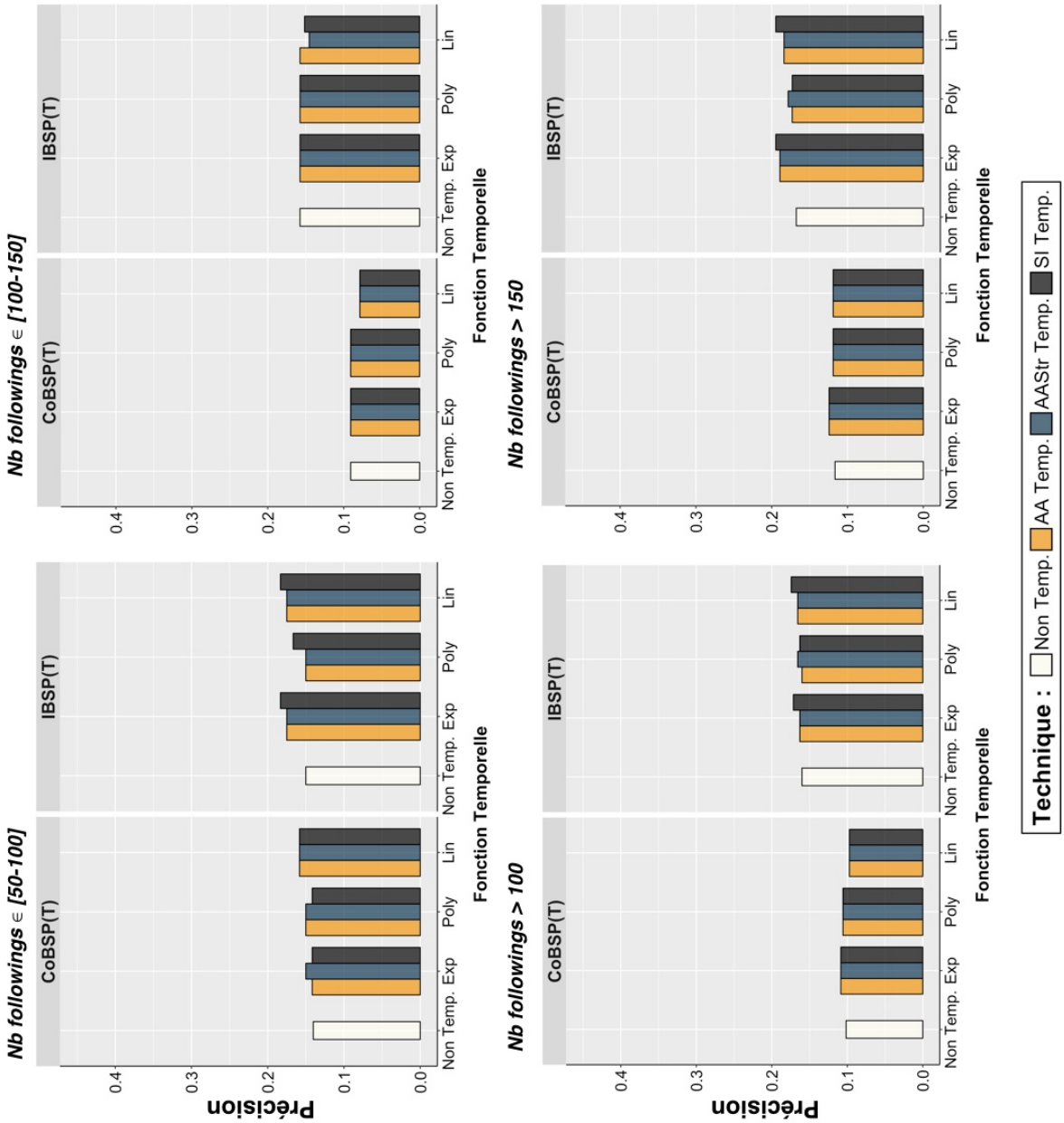
d. Résultats sur DBLP selon la densité du réseau



Annexe 1-d Comparaison de la meilleure précision moyenne des résultats des processus de construction du profil social en appliquant les différentes méthodes temporelles aux processus (IBSPT et CoBSPT) selon la densité du réseau

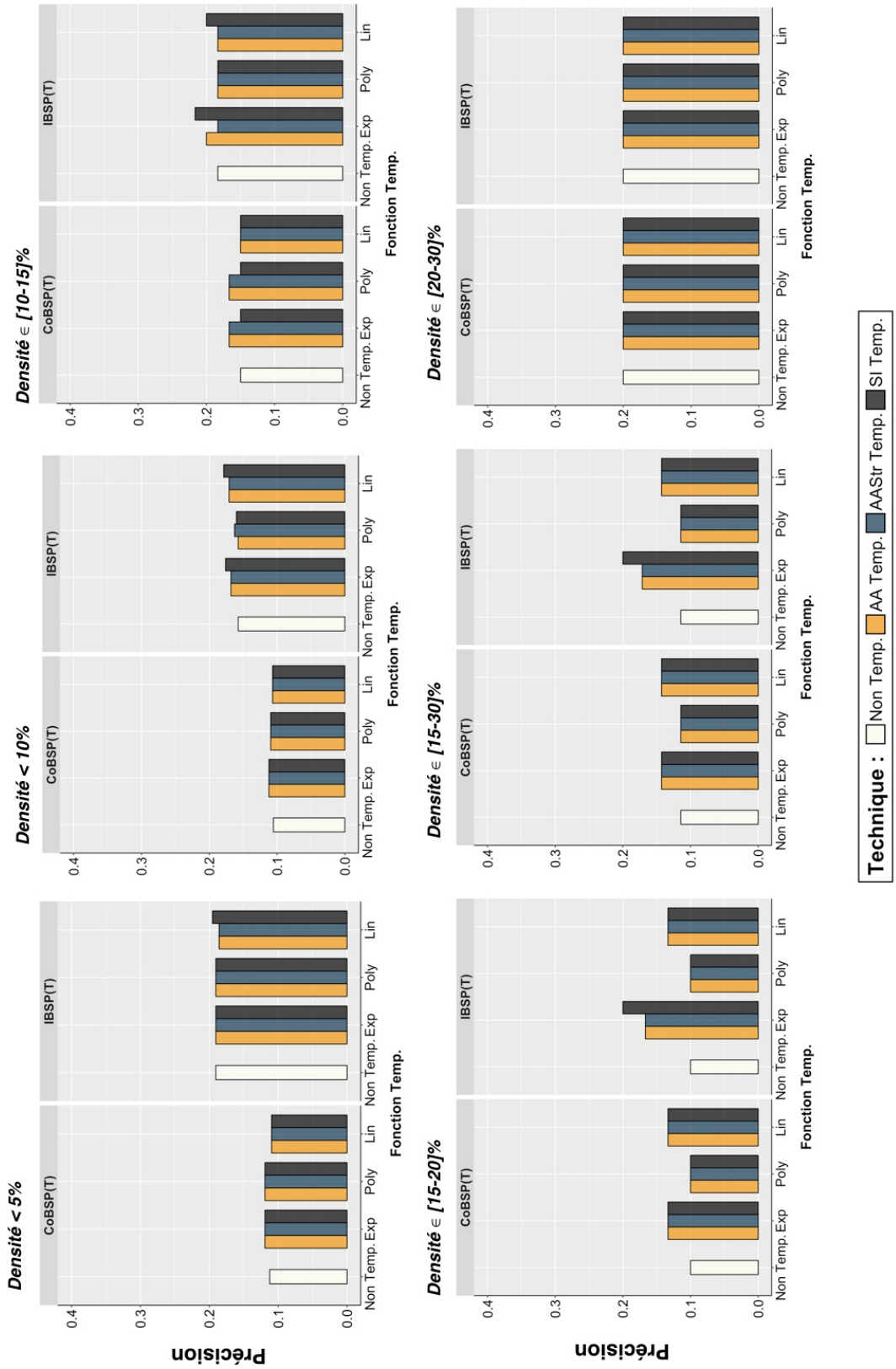
2. Résultats Twitter

a. Résultats sur Twitter selon la taille du réseau (nb followings)



Annexe 2-a Comparaison de la meilleure précision moyenne des résultats des processus de construction du profil social en appliquant les différentes méthodes temporelles aux processus (IBSPT et CoBSPT) selon la taille du réseau

b. Résultats sur Twitter selon la densité du réseau



Annexe 2-b Comparaison de la meilleure précision moyenne des résultats des processus de construction du profil social en appliquant les différentes méthodes temporelles aux processus (IBSPT et CoBSPT) selon la densité du réseau.

TABLE DES FIGURES

BIBLIOGRAPHIE

ABEL F., GAO Q., HOUBEN G.-J., TAO K., 2011, « Analyzing Temporal Dynamics in Twitter Profiles for Personalized Recommendations in the Social Web », *Proceedings of the 3rd International Web Science Conference*, p. 2:1–2:8.

ABEL F., GAO Q., HOUBEN G.-J., TAO K., 2013, « Twitter-based User Modeling for News Recommendations », *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, p. 2962–2966.

ADAMIC L.A., ADAR E., 2003, « Friends and neighbors on the Web », *Social Networks*, *25*, 3, p. 211‑230.

ADOMAVICIUS G., SANKARANARAYANAN R., SEN S., TUZHILIN A., 2005, « Incorporating Contextual Information in Recommender Systems Using a Multidimensional Approach », *ACM Trans. Inf. Syst.*, *23*, 1, p. 103–145.

AGGARWAL C., SUBBIAN K., 2012, « Event Detection in Social Streams », dans *Proceedings of the 2012 SIAM International Conference on Data Mining*, Society for Industrial and Applied Mathematics (Proceedings), p. 624‑635.

AGGARWAL C., SUBBIAN K., 2014, « Evolutionary Network Analysis: A Survey », *ACM Comput. Surv.*, *47*, 1, p. 10:1–10:36.

AIELLO L.M., PETKOS G., MARTIN C., CORNEY D., PAPADOPOULOS S., SKRABA R., GÖKER A., KOMPATSIARIS I., JAIMES A., 2013, « Sensing Trending Topics in Twitter », *IEEE Transactions on Multimedia*, *15*, 6, p. 1268‑1282.

ALBANO A., 2014, *Dynamique des graphes de terrain : analyse en temps intrinsèque*, thèse, Université Pierre et Marie Curie - Paris VI.

ALLEN J.F., 1979, *A Plan-based Approach to Speech Act Recognition*, thèse, Toronto, Canada, University of Toronto.

AMATO G., STRACCIA U., 1999, « User Profile Modeling and Applications to Digital Libraries », dans ABITEBOUL S., VERCOUSTRE A.-M. (dirs.), *Research and Advanced Technology for Digital Libraries*, Springer Berlin Heidelberg (Lecture Notes in Computer Science), p. 184‑197.

ANAND S.S., MOBASHER B., 2005, « Intelligent Techniques for Web Personalization », dans *Intelligent Techniques for Web Personalization*, Springer, Berlin, Heidelberg, p. 1‑36.

ARAL S., WALKER D., 2013, « Tie Strength, Embeddedness & Social Influence: Evidence from a Large Scale Networked Experiment », SSRN Scholarly Paper, ID 2197972, Rochester, NY, Social Science Research Network.

ARMSTRONG R., FREITAG D., JOACHIMS T., MITCHELL T., 1995, « WebWatcher: A Learning Apprentice for the World Wide Web », p. 6‑12.

ARNABOLDI V., CONTI M., PASSARELLA A., DUNBAR R., 2013, « Dynamics of Personal Social Relationships in Online Social Networks: A Study on Twitter », *Proceedings of the First ACM Conference on Online Social Networks*, p. 15–26.

ATTIAS C., BRAYER C., BRUNO S., JACQUOT C., STRUL R., THOBELLEM A., VILLALBA A., 2010, « Les médias sociaux », Paris, IAB France.

AUDEH B., 2014, *Reformulation sémantique des requêtes pour la recherche d'information ad hoc sur le Web*, thèse, Ecole Nationale Supérieure des Mines de Saint-Etienne.

BARABÁSI A.-L., ALBERT R., 1999, « Emergence of Scaling in Random Networks », *Science*, *286*, 5439, p. 509‑512.

BELKIN N.J., CROFT W.B., 1992, « Information Filtering and Information Retrieval: Two Sides of the Same Coin? », *Commun. ACM*, *35*, 12, p. 29–38.

BENESTY P.D.J., CHEN J., HUANG Y., COHEN P.I., 2009, « Pearson Correlation Coefficient », dans *Noise Reduction in Speech Processing*, Springer Berlin Heidelberg (Springer Topics in Signal Processing), p. 1‑4.

BENNETT P.N., WHITE R.W., CHU W., DUMAIS S.T., BAILEY P., BORISYUK F., CUI X., 2012, « Modeling the Impact of Short- and Long-term Behavior on Search Personalization », *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 185–194.

BERLINGERIO M., COSCIA M., GIANNOTTI F., MONREALE A., PEDRESCHI D., 2013, « Multidimensional networks: foundations of structural analysis », *World Wide Web*, *16*, 5‑6, p. 567‑593.

BERRUT C., DENOS N., 2003, « Filtrage collaboratif », dans *Assistance intelligente à la recherche d'informations*, Hermes - Lavoisier, p. 30.

BHAGAT S., CORMODE G., MUTHUKRISHNAN S., 2011, « Node Classification in Social Networks », dans AGGARWAL C.C. (dir.), *Social Network Data Analytics*, Springer US, p. 115‑148.

BHATTACHARYYA P., GARG A., WU S.F., 2011, « Analysis of user keyword similarity in online social networks », *Social Network Analysis and Mining*, *1*, 3, p. 143‑158.

BOCCALETTI S., LATORA V., MORENO Y., CHAVEZ M., HWANG D.-U., 2006, « Complex networks: Structure and dynamics », *Physics Reports*, *424*, 4–5, p. 175‑308.

BOND R.M., FARISS C.J., JONES J.J., KRAMER A.D.I., MARLOW C., SETTLE J.E., FOWLER J.H., 2012, « A 61-million-person experiment in social influence and political mobilization », *Nature*, *489*, 7415, p. 295‑298.

BORGATTI S.P., 2012, « Social Network Analysis, Two-Mode Concepts in », dans PH. D R.A.M. (dir.), *Computational Complexity*, Springer New York, p. 2912‑2924.

BORGATTI S.P., JONES C., EVERETT M.G., 1998, « Network measures of social capital », *Connections*, *21*, 2, p. 27–36.

BRA P.D., AERTS A., SMITS D., STASH N., 2002, « AHA! Meets AHAM », *Adaptive Hypermedia and Adaptive Web-Based Systems*, p. 388‑391.

BRESLIN J., DECKER S., 2007, « The Future of Social Networks on the Internet: The Need for Semantics », *IEEE Internet Computing*, *11*, 6, p. 86‑90.

BRUSILOVSKY P., 1996, « Methods and techniques of adaptive hypermedia », *User Modeling and User-Adapted Interaction*, *6*, 2‑3, p. 87‑129.

BURKE R., 2002, « Hybrid Recommender Systems: Survey and Experiments », *User Modeling and User-Adapted Interaction*, *12*, 4, p. 331–370.

BURT R.S., 1978, « Applied Network Analysis An Overview », *Sociological Methods & Research*, *7*, 2, p. 123‑130.

BURT R.S., MINOR M.J., 1983, *Applied Network Analysis: A Methodological Introduction*, Beverly Hills, SAGE Publications Ltd, 350 p.

BURT R.S., 2004, « Structural Holes and Good Ideas », *American Journal of Sociology*, *110*, 2, p. 349‑399.

CABANAC G., 2011, « Accuracy of Inter-researcher Similarity Measures Based on Topical and Social Clues », *Scientometrics*, *87*, 3, p. 597–620.

CARMEL D., ZWERDLING N., GUY I., OFEK-KOIFMAN S., HAR'EL N., RONEN I., UZIEL E., YOGEV S., CHERNOV S., 2009, « Personalized Social Search Based on the User's Social Network », *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, p. 1227–1236.

CATTUTO C., QUAGGIOTTO M., PANISSON A., AVERBUCH A., 2013, « Time-varying Social Networks in a Graph Database: A Neo4J Use Case », *First International Workshop on Graph Data Management Experiences and Systems*, p. 11:1–11:6.

CAZABET R., AMBLARD F., HANACHI C., 2010, « Detection of Overlapping Communities in Dynamical Social Networks », *2010 IEEE Second International Conference on Social Computing (SocialCom)*, p. 309‑314.

CHENG Y., QIU G., BU J., LIU K., HAN Y., WANG C., CHEN C., 2008, « Model Bloggers' Interests Based on Forgetting Mechanism », *Proceedings of the 17th International Conference on World Wide Web*, p. 1129–1130.

CHEVALIER M., 2011, « Usagers & Recherche d'Information », HDR, Université Paul Sabatier - Toulouse III.

CHO Y.H., KIM J.K., KIM S.H., 2002, « A personalized recommender system based on web usage mining and decision tree induction », *Expert Systems with Applications*, *23*, 3, p. 329‑342.

COHEN P.R., PERRAULT C.R., 1979, « Elements of a plan-based theory of speech acts », *Cognitive Science*, *3*, 3, p. 177‑212.

CORMODE G., SHKAPENYUK V., SRIVASTAVA D., XU B., 2009, « Forward Decay: A Practical Time Decay Model for Streaming Systems », *2009 IEEE 25th International Conference on Data Engineering*, p. 138‑149.

COUTANT A., STENGER T., 2013, « Médias sociaux: clarification et cartographie pour une approche sociotechnique », *Décisions Marketing*, 70, p. 107‑117.

CRANDALL D., COSLEY D., HUTTENLOCHER D., KLEINBERG J., SURI S., 2008, « Feedback Effects Between Similarity and Social Influence in Online Communities », *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 160–168.

CRUZ J.D., BOTHOREL C., POULET F., 2011, « Entropy based community detection in augmented social networks », *2011 International Conference on Computational Aspects of Social Networks (CASoN)*, p. 163‑168.

DAOUD M., 2009, *Accès personnalisé à l'information : approche basée sur l'utilisation d'un profil utilisateur sémantique dérivé d'une ontologie de domaines à travers l'historique des sessions de recherche*, Toulouse 3.

DAVIS JR. C.A., PAPPA G.L., OLIVEIRA D.R.R. DE, L. ARCANJO F. DE, 2011, « Inferring the Location of Twitter Messages Based on User Relationships », *Transactions in GIS*, *15*, 6, p. 735‑751.

DENOS N., 1997, *Modélisation de la pertinence en recherche d'information : modèle conceptuel, formalisation et application*, thèse, Université Joseph-Fourier - Grenoble I.

DING Y., LI X., 2005, « Time Weight Collaborative Filtering », *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, p. 485–492.

DING Y., 2011, « Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks », *Journal of informetrics*, *5*, 1, p. 187‑203.

EKSTRAND M.D., RIEDL J.T., KONSTAN J.A., 2011, « Collaborative Filtering Recommender Systems », *Found. Trends Hum.-Comput. Interact.*, *4*, 2, p. 81–173.

EVERETT M.G., BORGATTI S.P., 1999, « The centrality of groups and classes », *The Journal of Mathematical Sociology*, *23*, 3, p. 181‑201.

FINK J., KOBSA A., 2000, « A Review and Analysis of Commercial User Modeling Servers for Personalization on the World Wide Web », *User Modeling and User-Adapted Interaction*, *10*, 2‑3, p. 209‑249.

FORTUNATO S., 2010, « Community detection in graphs », *Physics Reports*, *486*, 3–5, p. 75‑174.

FREEMAN L.C., 1978, « Centrality in social networks conceptual clarification », *Social Networks*, *1*, 3, p. 215‑239.

FREY D., JÉGOU A., KERMARREC A.-M., 2011, « Social Market: Combining Explicit and Implicit Social Networks », dans DÉFAGO X., PETIT F., VILLAIN V. (dirs.), *Stabilization, Safety, and Security of Distributed Systems*, Springer Berlin Heidelberg (Lecture Notes in Computer Science), p. 193‑207.

FRIEDKIN N., 1980, « A test of structural features of granovetter's strength of weak ties theory », *Social Networks*, *2*, 4, p. 411‑422.

FUKUYAMA F., 1996, *Trust: The Social Virtues and The Creation of Prosperity*, 1st Free Press Pbk. Ed edition, New York, Free Press, 480 p.

GAO M., LIU K., WU Z., 2010, « Personalisation in Web Computing and Informatics: Theories, Techniques, Applications, and Future Research », *Information Systems Frontiers*, *12*, 5, p. 607–629.

GARCÍA S., LUENGO J., HERRERA F., 2015, *Data Prepocessing in Data Mining*, Springer International Publishing (Intelligent Systems Reference Library).

GAUCH S., SPERETTA M., CHANDRAMOULI A., MICARELLI A., 2007, « User Profiles for Personalized Information Access », dans BRUSILOVSKY P., KOBSA A., NEJDL W. (dirs.), *The Adaptive Web*, Springer Berlin Heidelberg (Lecture Notes in Computer Science), p. 54‑89.

GILBERT E., KARAHALIOS K., 2009, « Predicting Tie Strength with Social Media », *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, p. 211–220.

GILBERT E., KARAHALIOS K., SANDVIG C., 2008, « The Network in the Garden: An Empirical Analysis of Social Media in Rural Life », *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, p. 1603–1612.

GODOY D., AMANDI A., 2008, « Hybrid Content and Tag-based Profiles for Recommendation in Collaborative Tagging Systems », *Web Conference, 2008. LA-WEB '08., Latin American*, p. 58‑65.

GODOY D., 2006, « Learning user interests for user profiling in personal information agents », *AI Communications*, *19*, 4, p. 391‑394.

GRANOVETTER M.S., 1973, « The Strength of Weak Ties », *American Journal of Sociology*, *78*, 6, p. 1360‑1380.

GRINDROD P., HIGHAM D.J., 2010, « Evolving graphs: dynamical models, inverse problems and propagation », *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *466*, 2115, p. 753‑770.

GUARINO N., MASOLO C., VETERE G., 1999, « OntoSeek: content-based access to the Web », *IEEE Intelligent Systems and their Applications*, *14*, 3, p. 70‑80.

GUILLE A., 2014, *Diffusion de l'information dans les médias sociaux : modélisation et analyse*, thèse, Université Lumière Lyon 2.

GUY I., ZWERDLING N., CARMEL D., RONEN I., UZIEL E., YOGEV S., OFEK-KOIFMAN S., 2009, « Personalized Recommendation of Social Software Items Based on Social Relations », *Proceedings of the Third ACM Conference on Recommender Systems*, p. 53–60.

HANANI U., SHAPIRA B., SHOVAL P., 2001, « Information Filtering: Overview of Issues, Research and Systems », *User Modeling and User-Adapted Interaction*, *11*, 3, p. 203–259.

HOLME P., SARAMÄKI J., 2012, « Temporal networks », *Physics Reports*, *519*, 3, p. 97‑125.

HUBERT G., LOISEAU Y., MOTHE J., 2007, « 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*, p. 199‑207.

ISOZAKI H., KAZAWA H., 2002, « Efficient Support Vector Classifiers for Named Entity Recognition », *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, p. 1–7.

JACCARD P., 1901, « Étude comparative de la distribution florale dans une portion des Alpes et du Jura », *Bulletin de la Societe Vaudoise des Sciences Naturelles*, *37*, 142, p. 547‑579.

JIANG C., CHEN Y., LIU K.J.R., 2014, « Evolutionary Dynamics of Information Diffusion Over Social Networks », *IEEE Transactions on Signal Processing*, *62*, 17, p. 4573‑4586.

JOACHIMS T., 1997, « A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization », *Proceedings of the Fourteenth International Conference on Machine Learning*, p. 143–151.

JURGENS D., 2013, « That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships », *Seventh International AAAI Conference on Weblogs and Social Media*.

KACEM A., BOUGHANEM M., FAIZ R., 2014, « Time-Sensitive User Profile for Optimizing Search Personlization », dans DIMITROVA V., KUFLIK T., CHIN D., RICCI F., DOLOG P., HOUBEN G.-J. (dirs.), *User Modeling, Adaptation, and Personalization*, Springer International Publishing (Lecture Notes in Computer Science), p. 111‑121.

KAJDANOWICZ T., KAZIENKO P., DOSKOCZ P., 2010, « Label-Dependent Feature Extraction in Social Networks for Node Classification », dans BOLC L., MAKOWSKI M., WIERZBICKI A. (dirs.), *Social Informatics*, Springer Berlin Heidelberg (Lecture Notes in Computer Science), p. 89‑102.

KAPLAN A.M., HAENLEIN M., 2010, « Users of the world, unite! The challenges and opportunities of Social Media », *Business Horizons*, *53*, 1, p. 59‑68.

KATZ L., 1953, « A new status index derived from sociometric analysis », *Psychometrika*, *18*, 1, p. 39‑43.

KELLY D., TEEVAN J., 2003, « Implicit Feedback for Inferring User Preference: A Bibliography », *SIGIR Forum*, *37*, 2, p. 18–28.

KIM M., LESKOVEC J., 2010, « Multiplicative Attribute Graph Model of Real-World Networks », dans KUMAR R., SIVAKUMAR D. (dirs.), *Algorithms and Models for the Web-Graph*, Springer Berlin Heidelberg (Lecture Notes in Computer Science), p. 62‑73.

KOBSA A., 2001, « Generic User Modeling Systems », *User Modeling and User-Adapted Interaction*, *11*, 1‑2, p. 49‑63.

KOBSA A., 2007, « Generic User Modeling Systems », dans BRUSILOVSKY P., KOBSA A., NEJDL W. (dirs.), *The Adaptive Web*, Springer Berlin Heidelberg (Lecture Notes in Computer Science), p. 136‑154.

KREBS V., 2002, « Mapping Networks of Terrorist Cells », *CONNECTIONS*, *24*, 3, p. 43–52.

KUMAR R., NOVAK J., TOMKINS A., 2006, « Structure and Evolution of Online Social Networks », *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 611–617.

KWAK H., LEE C., PARK H., MOON S., 2010, « What is Twitter, a Social Network or a News Media? », *Proceedings of the 19th International Conference on World Wide Web*, p. 591–600.

LANCICHINETTI A., FORTUNATO S., RADICCHI F., 2008, « Benchmark graphs for testing community detection algorithms », *Physical Review E*, *78*, 4, p. 46110.

LANDAUER T.K., FOLTZ P.W., LAHAM D., 1998, « An Introduction to Latent Semantic Analysis », *Discourse Processes*, *25*, p. 259‑284.

LEENDERS R.T.A.J., 2002, « Modeling social influence through network autocorrelation: constructing the weight matrix », *Social Networks*, *24*, 1, p. 21‑47.

LEICHT E.A., NEWMAN M.E.J., 2008, « Community structure in directed networks », *Physical Review Letters*, *100*, 11, p. 118703.

LEROY V., CAMBAZOGLU B.B., BONCHI F., 2010, « Cold Start Link Prediction », *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 393–402.

LESKOVEC J., HUTTENLOCHER D., KLEINBERG J., 2010, « Predicting Positive and Negative Links in Online Social Networks », *Proceedings of the 19th International Conference on World Wide Web*, p. 641–650.

LEY M., 2009, « DBLP: Some Lessons Learned », *Proc. VLDB Endow.*, *2*, 2, p. 1493–1500.

LI D., CAO P., GUO Y., LEI M., 2013, « Time Weight Update Model Based on the Memory Principle in Collaborative Filtering », *Journal of Computers*, *8*, 11, p. 2763‑2767.

LI J., HUANG L., BAI T., WANG Z., CHEN H., 2012, « CDBIA: A dynamic community detection method based on incremental analysis », *2012 International Conference on Systems and Informatics (ICSAI)*, p. 2224‑2228.

LI L., YANG Z., WANG B., KITSUREGAWA M., 2007, « Dynamic Adaptation Strategies for Long-term and Short-term User Profile to Personalize Search », *Proceedings of the Joint 9th Asia-Pacific Web and 8th International Conference on Web-age Information Management Conference on Advances in Data and Web Management*, p. 228–240.

LI R., WANG C., CHANG K.C.-C., 2014, « User Profiling in an Ego Network: Co-profiling Attributes and Relationships », *Proceedings of the 23rd International Conference on World Wide Web*, p. 819–830.

LIANG H., XU Y., TJONDRONEGORO D., CHRISTEN P., 2012, « Time-aware Topic Recommendation Based on Micro-blogs », *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, p. 1657–1661.

LIBEN-NOWELL D., KLEINBERG J., 2003, « The Link Prediction Problem for Social Networks », *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, p. 556–559.

LIN N., 1995, « Les ressources sociales: une théorie du capital social », *Revue Française de Sociologie*, *36*, 4, p. 685.

LIN N., DAYTON P.W., GREENWALD P., 1978, « Analyzing the Instrumental Use of Relations in the Context of Social Structure », *Sociological Methods & Research*, *7*, 2, p. 149‑166.

LIN N., ENSEL W.M., VAUGHN J.C., 1981, « Social Resources and Strength of Ties: Structural Factors in Occupational Status Attainment », *American Sociological Review*, *46*, 4, p. 393‑405.

LIN Y.-R., CHI Y., ZHU S., SUNDARAM H., TSENG B.L., 2009, « Analyzing Communities and Their Evolutions in Dynamic Social Networks », *ACM Trans. Knowl. Discov. Data*, *3*, 2, p. 8:1–8:31.

LIU B., 2007, *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, Springer Science & Business Media, 544 p.

LIU F., YU C., MENG W., 2004, « Personalized Web Search For Improving Retrieval Effectiveness », *IEEE Trans. on Knowl. and Data Eng.*, *16*, 1, p. 28–40.

LOPS P., GEMMIS M. DE, SEMERARO G., 2011, « Content-based Recommender Systems: State of the Art and Trends », dans RICCI F., ROKACH L., SHAPIRA B., KANTOR P.B. (dirs.), *Recommender Systems Handbook*, Springer US, p. 73‑105.

MALOOF M.A., MICHALSKI R.S., 2000, « Selecting Examples for Partial Memory Learning », *Machine Learning*, *41*, 1, p. 27‑52.

MASSA P., AVESANI P., 2007, « Trust-aware Recommender Systems », *Proceedings of the 2007 ACM Conference on Recommender Systems*, p. 17–24.

MASUDA N., HOLME P., 2013, « Predicting and controlling infectious disease epidemics using temporal networks », *F1000Prime Reports*, *5*.

MAYFIELD J., MCNAMEE P., 2003, « Single N-gram Stemming », *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, p. 415–416.

MEHRA A., 2005, « Review of The Development of Social Network Analysis: A Study in the Sociology of Science », *Administrative Science Quarterly*, *50*, 1, p. 148‑151.

MELVILLE P., MOONEY R.J., NAGARAJAN R., 2002, « Content-boosted Collaborative Filtering for Improved Recommendations », *Eighteenth National Conference on Artificial Intelligence*, p. 187–192.

MEZGHANI M., 2015, *Analyse des réseaux sociaux : vers une adaptation de la navigation sociale*, thèse, Université de Toulouse, Université Toulouse III - Paul Sabatier.

MEZGHANI M., ON-AT S., PÉNINOU A., CANUT M.-F., ZAYANI C.A., AMOUS I., SEDES F., 2015, « A Case Study on the Influence of the User Profile Enrichment on Buzz Propagation in Social Media: Experiments on Delicious », dans MORZY T., VALDURIEZ P., BELLATRECHE L. (dirs.), *New Trends in Databases and Information Systems*, Springer International Publishing (Communications in Computer and Information Science), p. 567‑577.

MEZGHANI M., ZAYANI C.A., AMOUS I., GARGOURI F., 2012, « A User Profile Modelling Using Social Annotations: A Survey », *Proceedings of the 21st International Conference Companion on World Wide Web*, p. 969–976.

MOODY J., MCFARLAND D., BENDER-DEMOLL S., 2005, « Dynamic Network Visualization », *American Journal of Sociology*, *110*, 4, p. 1206‑1241.

MUNASINGHE L., ICHISE R., 2011, « Time Aware Index for Link Prediction in Social Networks », dans CUZZOCREA A., DAYAL U. (dirs.), *Data Warehousing and Knowledge Discovery*, Springer Berlin Heidelberg (Lecture Notes in Computer Science), p. 342‑353.

NEVILLE J., SIMSEK O., JENSEN D., 2004, « Autocorrelation and Relational Learning: Challenges and Opportunities ».

NEWMAN M.E.J., 2001a, « Scientific collaboration networks. I. Network construction and fundamental results », *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, *64*, 1 Pt 2, p. 16131.

NEWMAN M.E.J., 2001b, « Clustering and preferential attachment in growing networks », *Physical Review E*, *64*, 2.

NEWMAN M.E.J., 2004a, « Coauthorship networks and patterns of scientific collaboration », *Proceedings of the National Academy of Sciences of the United States of America*, *101 Suppl 1*, p. 5200‑5205.

NEWMAN M.E.J., 2004b, « Analysis of weighted networks », *Physical Review E*, *70*, 5, p. 56131.

O'MADADHAIN J., HUTCHINS J., SMYTH P., 2005, « Prediction and Ranking Algorithms for Event-based Network Data », *SIGKDD Explor. Newsl.*, *7*, 2, p. 23–30.

PALLA G., BARABÁSI A.-L., VICSEK T., 2007, « Quantifying social group evolution », *Nature*, *446*, 7136, p. 664‑667.

PALLA G., DERÉNYI I., FARKAS I., VICSEK T., 2005, « Uncovering the overlapping community structure of complex networks in nature and society », *Nature*, *435*, 7043, p. 814‑818.

PAPADOGIORGAKI M., PAPASTATHIS V., NIDELKOU E., WADDINGTON S., BRATU B., RIBIERE M., KOMPATSIARIS I., 2008, « Two-Level Automatic Adaptation of a Distributed User Profile for Personalized News Content Delivery », *International Journal of Digital Multimedia Broadcasting*, *2008*, p. e863613.

PAVLOV M., 2007, « Finding Experts by Link Prediction in Co-authorship Networks », *FEWS*, *290*, p. 42‑45.

PAZZANI M., BILLSUS D., 1997, « Learning and Revising User Profiles: The Identification ofInteresting Web Sites », *Mach. Learn.*, *27*, 3, p. 313–331.

PAZZANI M.J., BILLSUS D., 2007, « Content-Based Recommendation Systems », dans *The Adaptive Web*, Springer Berlin Heidelberg, p. 325‑341.

PEMPEK T.A., YERMOLAYEVA Y.A., CALVERT S.L., 2009, « College students' social networking experiences on Facebook », *Journal of Applied Developmental Psychology*, *30*, 3, p. 227‑238.

PON R.K., CÁRDENAS A.F., BUTTLER D.J., CRITCHLOW T.J., 2011, « Measuring the interestingness of articles in a limited user environment », *Information Processing & Management*, *47*, 1, p. 97‑116.

PORTER M. F., 2006, « An algorithm for suffix stripping », *Program*, *40*, 3, p. 211‑218.

PRETSCHNER A., GAUCH S., 1999, « Ontology based personalized search », *Proceedings 11th International Conference on Tools with Artificial Intelligence*, p. 391‑398.

PUTNAM R., 1995, « Bowling Alone: America's Declining Social Capital », *Journal of Democracy*, *6*, 1, p. 65‑78.

RATTIGAN M.J., JENSEN D., 2005, « The Case for Anomalous Link Discovery », *SIGKDD Explor. Newsl.*, *7*, 2, p. 41–47.

REN X., ZENG Y., QIN Y., ZHONG N., HUANG Z., WANG Y., WANG C., 2010, « Social Relation Based Search Refinement: Let Your Friends Help You! », dans AN A., LINGRAS P., PETTY S., HUANG R. (dirs.), *Active Media Technology*, Springer Berlin Heidelberg (Lecture Notes in Computer Science), p. 475‑485.

RESNICK P., IACOVOU N., SUCHAK M., BERGSTROM P., RIEDL J., 1994, « GroupLens: An Open Architecture for Collaborative Filtering of Netnews », *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, p. 175–186.

RIJSBERGEN C.J.V., 1979, *Information Retrieval*, 2nd édition, Newton, MA, USA, Butterworth-Heinemann.

RIVERA M.T., SODERSTROM S.B., UZZI B., 2010, « Dynamics of Dyads in Social Networks: Assortative, Relational, and Proximity Mechanisms », *Annual Review of Sociology*, *36*, 1, p. 91‑115.

ROBERTSON S.E., SPARCK JONES K., 1988, « Document Retrieval Systems », dans WILLETT P. (dir.), London, UK, UK, Taylor Graham Publishing, p. 143–160.

ROCCHIO J.J., 1971, « Relevance Feedback in Information Retrieval », dans *The SMART Retrieval System—Experiments in Automatic Document Processing*, Upper Saddle River, NJ, USA, Prentice-Hall, Inc.

ROSVALL M., BERGSTROM C.T., 2007, « An information-theoretic framework for resolving community structure in complex networks », *Proceedings of the National Academy of Sciences*, *104*, 18, p. 7327‑7331.

ROWE M., STANKOVIC M., ALANI H., 2012, « Who Will Follow Whom? Exploiting Semantics for Link Prediction in Attention-Information Networks », dans CUDRÉ-MAUROUX P., HEFLIN J., SIRIN E., TUDORACHE T., EUZENAT J., HAUSWIRTH M., PARREIRA J.X., HENDLER J., SCHREIBER G., BERNSTEIN A., BLOMQVIST E. (dirs.), *The Semantic Web – ISWC 2012*, Springer Berlin Heidelberg (Lecture Notes in Computer Science), p. 476‑491.

SALTON G., MCGILL M.J., 1986, *Introduction to Modern Information Retrieval*, New York, NY, USA, McGraw-Hill, Inc.

SARWAR B., KARYPIS G., KONSTAN J., RIEDL J., 2001, « Item-based Collaborative Filtering Recommendation Algorithms », *Proceedings of the 10th International Conference on World Wide Web*, p. 285–295.

SCHIFANELLA R., BARRAT A., CATTUTO C., MARKINES B., MENCZER F., 2010, « Folks in Folksonomies: Social Link Prediction from Shared Metadata », *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, p. 271–280.

SEHABA K., 2012, « Système d'aide adaptatif à base de traces », *Revue internationale des technologies en pédagogie universitaire / International Journal of Technologies in Higher Education*, *9*, 3, p. 55‑70.

SHANG J., LIU L., XIE F., CHEN Z., MIAO J., FANG X., WU C., 2012, « A real-time detecting algorithm for tracking community structure of dynamic networks »,.

SHAW J.A., FOX E.A., SHAW J.A., FOX E.A., 1994, « Combination of Multiple Searches », *The Second Text REtrieval Conference (TREC-2*, p. 243–252.

SHI X., ADAMIC L.A., STRAUSS M.J., 2007, « Networks of strong ties », *Physica A: Statistical Mechanics and its Applications*, *378*, 1, p. 33‑47.

SINGLA P., RICHARDSON M., 2008, « Yes, There is a Correlation: - from Social Networks to Personal Behavior on the Web », *Proceedings of the 17th International Conference on World Wide Web*, p. 655–664.

SINHA R.R., SWEARINGEN K., 2002, « Comparing Recommendations Made by Online Systems and Friends. », *DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries*.

SPILIOPOULOU M., 2011, « Evolution in Social Networks: A Survey », dans AGGARWAL C.C. (dir.), *Social Network Data Analytics*, Springer US, p. 149‑175.

STATISTA, 2016, « Statistics and facts about Twitter », *www.statista.com*.

STATTNER E., COLLARD M., VIDOT N., 2013, « D2SNet: Dynamics of diffusion and dynamic human behaviour in social networks », *Computers in Human Behavior*, *29*, 2, p. 496‑509.

SU X., KHOSHGOFTAAR T.M., 2009, « A Survey of Collaborative Filtering Techniques », *Adv. in Artif. Intell.*, *2009*, p. 4:2–4:2.

SUGIYAMA K., HATANO K., YOSHIKAWA M., 2004, « Adaptive Web Search Based on User Profile Constructed Without Any Effort from Users », *Proceedings of the 13th International Conference on World Wide Web*, p. 675–684.

SUN J., TANG J., 2011, « A Survey of Models and Algorithms for Social Influence Analysis », dans AGGARWAL C.C. (dir.), *Social Network Data Analytics*, Springer US, p. 177‑214.

TAN B., SHEN X., ZHAI C., 2006, « Mining Long-term Search History to Improve Search Accuracy », *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 718–723.

TANUDJAJA F., MUI L., 2002, « Persona: a contextualized and personalized web search », *Proceedings of the 35th Annual Hawaii International Conference on System Sciences*, p. 1232‑1240.

TCHUENTE D., 2013, *Modélisation et dérivation de profils utilisateurs à partir de réseaux sociaux : approche à partir de communautés de réseaux k-égocentriques*, thèse, Université de Toulouse, Université Toulouse III - Paul Sabatier.

TCHUENTE D., CANUT M.-F., JESSEL N., PENINOU A., SÈDES F., 2013, « A community-based algorithm for deriving users' profiles from egocentrics networks: experiment on Facebook and DBLP », *Social Network Analysis and Mining*, *3*, 3, p. 667‑683.

TIAN DAI B., CHONG TAT CHUA F., LIM E., 2012, « Structural Analysis in Multi-Relational Social Networks », dans *Proceedings of the 2012 SIAM International Conference on Data Mining*, Society for Industrial and Applied Mathematics (Proceedings), p. 451‑462.

TRIER M., 2008, « Research Note—Towards Dynamic Visualization for Understanding Evolution of Digital Communication Networks », *Information Systems Research*, *19*, 3, p. 335‑350.

TYLENDA T., ANGELOVA R., BEDATHUR S., 2009, « Towards Time-aware Link Prediction in Evolving Social Networks », *Proceedings of the 3rd Workshop on Social Network Mining and Analysis*, p. 9:1–9:10.

WANG C., SHEN Y., YANG H., GUO M., 2013, « Improving Rocchio Algorithm for Updating User Profile in Recommender Systems », dans LIN X., MANOLOPOULOS Y., SRIVASTAVA D., HUANG G. (dirs.), *Web Information Systems Engineering – WISE 2013*, Springer Berlin Heidelberg (Lecture Notes in Computer Science), p. 162‑174.

WASSERMAN S., FAUST K., 1994, *Social network analysis: methods and applications*, Cambridge; New York, Cambridge University Press.

WATTS D.J., STROGATZ S.H., 1998, « Collective dynamics of 'small-world' networks », *Nature*, *393*, 6684, p. 440‑442.

WELLMAN B., WORTLEY S., 1990, « Different Strokes from Different Folks: Community Ties and Social Support », *American Journal of Sociology*, *96*, 3, p. 558‑588.

WEN Z., LIN C.-Y., 2010, « On the Quality of Inferring Interests from Social Neighbors », *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 373–382.

WEN Z., LIN C.-Y., 2011, « Improving User Interest Inference from Social Neighbors », *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, p. 1001–1006.

WENG L., RATKIEWICZ J., PERRA N., GONÇALVES B., CASTILLO C., BONCHI F., SCHIFANELLA R., MENCZER F., FLAMMINI A., 2013, « The Role of Information Diffusion in the Evolution of Social Networks », *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 356–364.

XIANG R., NEVILLE J., ROGATI M., 2010, « Modeling Relationship Strength in Online Social Networks », *Proceedings of the 19th International Conference on World Wide Web*, p. 981–990.

XU J., ZHU Z., REN X., TIAN Y., LUO Y., 2007, « Personalized Web Search Using User Profile », *2007 International Conference on Computational Intelligence and Security (CIS 2007)*, p. 222‑226.

YAHIA S.A., BENEDIKT M., BOHANNON P., 2007, « Challenges in searching online communities », *IEEE Data Eng. Bull*, *30*.

YANG J., MCAULEY J., LESKOVEC J., 2013, « Community Detection in Networks with Node Attributes », *2013 IEEE 13th International Conference on Data Mining*, p. 1151‑1156.

YIN Z., GUPTA M., WENINGER T., HAN J., 2010, « LINKREC: A Unified Framework for Link Recommendation with User Attributes and Graph Structure », *Proceedings of the 19th International Conference on World Wide Web*, p. 1211–1212.

ZEMIRLI W.N., TAMINE-LECHANI L., 2007, « A Personalized Retrieval Model based on Influence Diagrams », *Sixth International and Interdisciplinary Conference on Modeling and Using Context*.

ZEMIRLI W.N., 2008, *Modèle d'accès personnalisé à l'information basé sur les Diagrammes d'Influence intégrant un profil utilisateur évolutif*, thèse, Toulouse, France, Université de Toulouse, Université Toulouse III - Paul Sabatier.

ZENG Y., YAO Y., ZHONG N., 2009, « DBLP-SSE: A DBLP Search Support Engine », *IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT '09*, *1*, p. 626‑630.

ZHANG Z.-K., LIU C., ZHANG Y.-C., ZHOU T., 2010, « Solving the cold-start problem in recommender systems with social tags », *EPL (Europhysics Letters)*, *92*, 2, p. 28002.

ZHENG N., LI Q., 2011, « A recommender system based on tag and time information for social tagging systems », *Expert Systems with Applications*, *38*, 4, p. 4575‑4587.