Apports des réseaux sociaux pour la gestion de la relation client

Ian Basaille-Gahitte, Lylia Abrouk, Nadine Cullot, Eric Leclercq

Apports des réseaux sociaux pour la gestion de la relation client

Ian Basaille — Lylia Abrouk — Nadine Cullot — Éric Leclerq

Laboratoire Électronique, Informatique et Image
UMR CNRS 6306 - Université de Bourgogne
9 Avenue Alain Savary
21078 Dijon CEDEX

prenom.nom@u-bourgogne.fr

RÉSUMÉ. *Depuis quelques années, le Web s'est transformé en une plateforme d'échanges. La gestion de relation client doit évoluer pour tirer partie des données disponibles sur les réseaux sociaux et mettre l'entreprise au cœur des échanges. Nous proposons dans cet article une approche générique de détection de communautés de clients d'une entreprise, basée sur leur comportement explicite et implicite, intégrant des données de sources diverses. Nous définissons une mesure de similarité, entre un utilisateur et un tag, prenant en compte la notation et la consultation des ressources et le réseau social de l'utilisateur. Nous validons cette approche sur une base exemple en utilisant deux méthodes de détection de communautés pour trois cas d'utilisation.*

ABSTRACT. *In recent years, the Web has evolved into an exchange platform. Customer relationship management must follow this evolution and provide users with tools to integrate data from social networks in order to place companies at the heart of the exchanges. We propose in this paper a generic approach to community detection of customers of a company based on their explicit and implicit behavior, integrating data from various sources. For this, we define a similarity measure, between a user and a tag, that takes into account the rating and consultation of resources and the social network of the user. We validate this approach against a test database by using two different community detection methods for three use cases.*

MOTS-CLÉS : *communautés, réseaux sociaux, relation client, gestion de la relation client, Social CRM, découverte de communautés, profils, modélisation d'utilisateur, tags, Web 2.0.*

KEYWORDS: *communities, social networks, client relationship, client relationship management, Social CRM, community detection, profiles, user modeling, tags, Web 2.0.*

1. Introduction

Le métier de la Gestion de la Relation Client (GRC) ou *Customer Relationship Management* (CRM) est en pleine évolution, impulsée par l'arrivée des technologies du Web 2.0 et la popularité croissante des réseaux sociaux. Les médias sociaux grand public, comme *Facebook*[1] ou *Twitter*[2] sont maintenant utilisés quotidiennement par les salariés et les entreprises développent de plus en plus leur propre réseau social en partageant et élaborant du contenu ou en créant des liens entre les clients. De plus, la démocratisation du e-commerce depuis le début des années 2000 a bouleversé le commerce en général. Le Social Commerce, c'est-à-dire, l'achat de biens et services utilisant le levier d'un réseau social (Rignault *et al.,* 2012) est un évolution du e-commerce, s'appuyant sur l'évolution du Web en Web Social. Ces nouveaux usages du Web et la richesse des données sociales apportent aux professionnels du CRM des promesses importantes en termes de développement de la connaissance sur les clients et les consommateurs (Melville *et al.,* 2009). En effet, depuis quelques années, le Web s'est transformé en une plateforme d'échange générique, où tout utilisateur devient fournisseur de contenu via des outils comme les blogs avec commentaires, les wikis avec les fonctionnalités de collaboration et de contribution, ou encore les réseaux sociaux avec le partage de ressources, de contenu et les mécanismes d'annotation.

La compréhension des mécanismes d'interaction entre une entreprise et ses clients mais aussi de la diffusion de l'information entre les clients ou futurs clients (via les forums par exemple), ainsi que la connaissance sur les profils des consommateurs sont des éléments essentiels pour la croissance et la compétitivité. Les entreprises utilisent le Web pour différentes finalités, principalement axées sur le marketing et la vente de produits. Pour le marketing, elles utilisent les technologies Web à travers des sites dédiés, ou au moyen de publicités incluses dynamiquement sur d'autres sites, ou encore avec des campagnes de diffusion de courriers électroniques généralement peu ciblées. Pour leurs activités de commerce électronique, les entreprises développent également des applications riches incluant par exemple des fonctionnalités de notation des produits ou de recommandation. Ainsi, les consommateurs sont amenés à interagir avec des sites marchands non seulement pour connaître les produits ou réaliser des achats, mais aussi pour correspondre avec les entreprises à travers les services en ligne. Par conséquent, les consommateurs deviennent acteurs pour la marque ou la société en donnant des notes, avis et commentaires sur les produits et informations mises à leur disposition, voire même en enrichissant ces informations que ce soit sur des sites marchands, sur le site dédié de la marque ou encore sur les réseaux sociaux. Pour ces entreprises, l'usage des médias sociaux comme moyen de communication avec les consommateurs et, au-delà, comme moyen d'investigation pour étudier le comportement des consommateurs, est un enjeu important.

Du point de vue de l'entreprise, l'offre d'outils de CRM est vaste. Ces outils visent à proposer des fonctionnalités permettant d'améliorer la gestion des services commer-

1. https://www.facebook.com/
2. https://twitter.com/

ciaux, marketing ou après-vente et incluent généralement des méthodes d'analyse, notamment statistiques. Depuis quelques années, les outils CRM évoluent vers une meilleure prise en compte de la dimension sociale des échanges entre les clients et la société ou entre les clients eux-mêmes vis-à-vis de la société. Ansi, le terme *Social CRM* est associé à l'utilisation des médias sociaux dans le cadre d'outils de gestion de la relation client (Mohan *et al.,* 2008). On distingue généralement trois catégories d'outils CRM : généralistes, intégrables et génériques.

Les outils *généralistes* sont des logiciels « sur étagère » plutôt destinés aux petites et moyennes entreprises. Les logiciel *SugarCRM* et *SalesForce* sont des exemples de cette catégorie ayant des fonctionnalités proches. *SugarCRM* propose des fonctionnalités classiques des CRM (gestion de la relation commerciale, marketing, service client, outils d'analyse) mais aussi quelques fonctionnalités collaboratives pour la vente et l'intégration de contacts depuis les médias sociaux.

Les outils *intégrables* sont des modules logiciels destinés à s'interconnecter avec le système d'information (SI) de l'entreprise. Ils peuvent avoir des fonctionnalités plus ciblées comme l'analyse ou la fouille de données. C'est le cas par exemple de logiciels comme *Smarter Analytics d'IBM* qui permet l'analyse de données à des fins décisionnelles au sein d'une entreprise.

Les outils *génériques* sont des logiciels développés pour être paramétrables et adaptables aux besoins des sociétés. L'offre peut être modulaire et toucher tous les domaines de la gestion de la relation client. Ces derniers peuvent plus facilement être étendus pour évoluer vers le *Social CRM*.

Les logiciels CRM doivent évoluer pour pouvoir analyser le comportement des consommateurs aussi bien en tant que clients, prospects mais aussi comme acteurs qui participent à la *e-réputation* d'une entreprise. Le suivi de cette *e-réputation* concerne aussi bien un marque que ses produits et services et ses dirigeants, ainsi il est important de savoir qui parle d'une entreprise, la portée des opinions émises et de pouvoir réagir le cas échéant (Cordina *et al.,* 2013). Les impacts sur l'architecture logicielle des CRM concernent non seulement l'interconnexion avec les applications traditionnelles des SI d'entreprise, avec des outils collaboratifs tels les réseaux sociaux d'entreprises[3] mais aussi l'interconnexion avec les médias sociaux grand public (Ajmera *et al.,* 2013). Les motivations de l'intégration des médias sociaux dans les CRM peuvent être classées en trois catégories principales :

– **la détection de communautés** dans les populations de clients ou de prospects. D'un point de vue algorithmique, la détection de communautés produira une définition en extension permettant d'obtenir une vue macroscopique d'un système complexe, utilisée ensuite dans la compréhension et l'analyse du système. L'approche duale est l'identification des personnes extérieures à une communauté, afin d'analyser leurs profils ou afin de déterminer la communauté qui leur est la plus proche ;

3. Il existe plusieurs outils spécifiques qualifiés de réseaux sociaux d'entreprise, comme par exemple *Yammer* (`http://www.yammer.com`) qui permet de travailler en réseau avec ses collègues ou *Bluekiwi* (`http://www.bluekiwi-software.com`), une plateforme de collaboration et de dialogue pour les échanges internes et externes de l'entreprise.

– **l'analyse communautés** ou leur caractérisation par une définition en intention, c'est-à-dire au moyen de propriétés caractéristiques éventuellement hiérarchisées. Les méthodes et algorithmes utilisés doivent permettre de faire émerger une sémantique à partir des données des éléments de la communauté. L'analyse peut faire appel à des concepts de socio-psychologie pour la définition de profils utilisateur comme par exemple l'endoreprésentation (représentation qu'un groupe se fait de lui-même, indépendamment de celle que les autres groupes se font de lui) ou l'exoreprésentation (représentation qu'un autre groupe se fait de ce groupe) (Perrin, 2011) ;

– **l'analyse des flux d'information** intra-communauté et inter-communautés pour identifier les personnes influentes afin de diffuser au mieux des informations de l'entreprise mais aussi afin d'effectuer une prédiction.

Nos travaux se concentrent sur la détection et l'analyse des communautés. Cependant, ces communautés peuvent aussi bien se trouver dans les différents médias sociaux publics que dans le réseau social professionnel d'une entreprise. De plus, les communautés d'utilisateurs existent de manière implicite et sont le résultat des comportements des utilisateurs par rapport aux applications, mais aussi le résultat des interactions entre utilisateurs. La contribution principale décrite dans cet article concerne la modélisation des utilisateurs via des profils, incluant des éléments explicites et implicites, basée sur les usages et comportements des utilisateurs. Les données recueillies pour constituer le profil sont issues d'applications aussi bien à l'intérieur qu'à l'extérieur du SI (activités des internautes sur un site dédié ou sur des médias sociaux publics). Nous proposons également une plateforme générique de détection de communautés, DisCoCRM (*Discovering Communities for CRM*), à destination des *community manager* ou des conseillers clients. Les utilisateurs de la plateforme ont des connaissances *a priori* du domaine de la marque. Ainsi, les analyses qu'ils effectuent se basent sur un vocabulaire contrôlé connu à l'avance. Nous décrirons trois expérimentations. La première utilise l'algorithme des k-means avec un *community manager* n'ayant pas de connaissance précise sur la population qu'il étudie. La seconde utilise la méthode de Louvain (Blondel *et al.,* 2008) avec l'hypothèse d'un *community manager* ayant une connaissance empirique de la population qu'il analyse. Dans ce cas, l'algorithme proposera des communautés et des tags les caractérisant, que le *community manager* confrontera à sa connaissance. Enfin, la dernière expérience pilote la méthode de Louvain par une connaissance formalisée du domaine. Dans notre cas, nous utiliserons une hiérarchie de termes issue du domaine de la nutrition-santé.

L'article est organisé de la façon suivante : la section 2 présente une synthèse des états de l'art sur la détection de communautés et les travaux concernant la modélisation de profil utilisateur. La section 3 présente l'approche DisCoCRM, au travers de l'architecture générale d'un système de type *Social CRM*, la construction du profil client et la détection de communautés. Avant de conclure et de présenter plusieurs perspectives dans la section 5, nous testons notre approche sur un jeu de données réduit mais représentatif dans la section 4.

2. État de l'art

Les médias sociaux constituent un domaine de recherche très actif. (Quan, 2011) présente un état de l'art technique sur les réseaux sociaux, leurs fonctionnalités, les plateformes et les nouvelles problématiques de recherche comme la structure distribuée des informations, l'interopérabilité des plateformes, la recherche d'identité, la propriété des données utilisateur et la sécurité. Les médias sociaux partagent avec de nombreux autres domaines applicatifs des problématiques de type graphe. En effet, ils traitent d'entités en relation qui peuvent être modélisées par des graphes dans lesquels les entités (utilisateurs ou ressources) sont des sommets et les relations des arcs ou arêtes. Les relations peuvent décrire des liens d'affinité entre les utilisateurs, des similarités thématiques entre des ressources, etc. Ces graphes, appelés graphes de terrain ou réseaux complexes, ont la particularité d'être produits par des interactions/interventions humaines et s'opposent aux graphes théoriques produits à partir de modèles. Ces graphes se rencontrent par exemple dans les sciences humaines et sociales, les réseaux informatiques ou la biologie. De plus, ils ont les particularités d'être de grande taille et d'évoluer au cours du temps, sans montrer de propriété structurelle évidente. Cependant les graphes de terrain sont généralement composés de sous-graphes, nommés communautés, denses et faiblement inter-connectés (Girvan *et al.,* 2002, Yang *et al.,* 2010).

2.1. *Découverte des communautés Web*

Depuis les débuts du Web, la recherche de communautés a fortement évolué. Elle a d'abord concerné l'étude des liens entre des documents ayant un même contenu thématique (Rome *et al.,* 2005) pour aboutir, depuis quelques années, à l'étude des liens entre des individus en fonction de leurs inter-actions. Dans les médias sociaux, les communautés sont généralement des ensembles de ressources, d'utilisateurs et/ou de tags (Papadopoulos *et al.,* 2010).

Un grand nombre de méthodes de détection de communautés ont récemment été publiées, ainsi que plusieurs travaux de classification. Les plus significatifs pour notre domaine sont (Plantié *et al.,* 2013, Fortunato, 2010, Porter *et al.,* 2009). Les classifications s'appuient sur des définitions similaires de la notion de communauté mais avec des points de vue différents. (Fortunato, 2010) propose, selon une approche orientée graphe et algorithmes, trois niveaux de définition de la notion de communauté : une définition locale par la structure interne, une globale définie à partir d'un critère de regroupement et une basée sur la similarité des sommets. Les algorithmes existants sont mis en perspective des trois niveaux pour aboutir à huit classes. (Porter *et al.,* 2009) adoptent le point de vue des sciences sociales pour définir les communautés au moyen de la cohésion de groupe, traduite par une notion de densité, permettant d'identifier cinq classes d'algorithmes :

– le clustering par regroupement hiérarchique ou par partitionnement ;

– la centralité qui mesure l'importance relative d'un sommet dans le graphe (Newman *et al.,* 2004) ;

– la percolation de cliques qui fusionne des k-cliques[4] adjacentes, c'est-à-dire qui partagent $k-1$ sommets, pour constituer une communauté (Palla *et al.,* 2005) ;

– les méthodes spectrales de partitionnement de graphe utilisant des outils algébriques développés pour étudier les espaces multi-dimensionnels ;

– l'optimisation de modularité se base sur le nombre d'arêtes incidentes à un ensemble de sommets par rapport à un graphe aléatoire (Newman, 2006).

(Plantié *et al.,* 2013) après avoir présenté les classification existantes, adoptent le point de vue de la structure des graphes et de leurs capacités de modélisation des relations complexes. Ainsi il obtiennent trois grandes catégories d'approches :

– celles considérant le réseau social comme un graphe avec des relations simples/binaires entre les sommets ;

– celles considérant le réseau comme un hypergraphe, permettant de modéliser des relations plus complexes, où les communautés apparaissent comme des (hyper)arêtes ;

– celles utilisant les concepts de treillis traduisant une relation d'ordre ou faisant émerger des propriétés sémantiques avec les treillis de Galois.

Parmi l'ensemble des algorithmes présentés dans les trois états de l'art, on peut noter certaines singularités. Les approches de percolation de cliques et les hypergraphes permettent notamment le traitement des communautés recouvrantes. De plus, les travaux utilisant les hypergraphes sont peu nombreux comparés aux travaux sur les graphes. Les treillis permettent de faire émerger des caractéristiques et participent donc à une analyse sémantique des communautés. Les fonctions de coûts ou de distances sont utilisées dans de nombreux algorithmes, (Cohen *et al.,* 2012) présentent un état de l'art des mesures de proximité permettant de quantifier le degré de similarité entre deux utilisateurs.

Afin de mesurer la qualité des algorithmes, c'est-à-dire la qualité de la partition produite, la notion de modularité a été introduite par Newman dans (Newman *et al.,* 2004) et étendue pour traiter des graphes avec des liens valués (Blondel *et al.,* 2008), en se basant sur la notion intuitive qu'une communauté est un ensemble de sommets dont la densité des connexions internes est plus importante que la densité des connexions externes. En considérant un graphe $G = (V, E)$ et une partition en communautés disjointes $\mathcal{P} = \{C_i, i = 1, .., n\}$, la fraction de liens situés à l'intérieur des communautés de $\mathcal{P}$ est notée $\sum_{i=1}^{n} e_{C_i}$. Cette fraction est comparée à un graphe aléatoire (qui ne possède pas de structure communautaire) ayant le même nombre de sommets et la même distribution des degrés. Dans le graphe aléatoire, pour $\mathcal{P}$, la probabilité qu'un lien ait une extrémité dans la communauté C_i est $a_{C_i} = (\sum_{j \in C_i} d(j))/2|E|$ avec $d(j)$ le degré du sommet j. La probabilité que les deux extrémités du lien soient dans la communauté C_i est $a_{C_i}^2$ et par conséquent la

4. Ensemble de sommets complètement connectés

modularité est $Q(\mathcal{P}) = \sum_{i=1}^{n}(e_{C_i} - a_{C_i}^2)$. Cependant, maximiser le modularité d'un graphe est un problème NP-difficile. La modularité est à la base de la méthode de Louvain (Blondel *et al.,* 2008), qui est une heuristique, composée de deux phases qui sont répétées jusqu'à obtenir un maximum local de modularité.

Du point de vue général, les étapes clés d'une approche de détection de communautés sont selon (Vakali *et al.,* 2012) : la définition de mesures pour le calcul des relations entre les utilisateurs afin de détecter de communautés implicites ; l'utilisation d'approches algorithmiques pour la détection de réseaux complexes tels les réseaux du Web ; le choix et utilisation d'une mesure pour l'évaluation des communautés.

2.2. *Tags et données sociales pour le profil utilisateur*

La construction d'un profil unique par utilisateur est une étape importante dans la construction des communautés. Le profil utilisateur est constitué d'un ensemble d'informations le concernant comme son nom, son âge, sa ville. Souvent, le profil contient aussi des informations sur ses centres d'intérêts et les évaluations qu'il donne à des ressources (Golbeck, 2009). Les intérêts peuvent être renseignés directement par l'utilisateur de manière explicite, ou de manière implicite en analysant son comportement.

Dans les approches avec profil explicite, (Hung *et al.,* 2008) définissent un profil comme un ensemble de tags (mots clés) et de poids. (Cattuto *et al.,* 2008) proposent un algorithme de recommandation basé sur les tags. Dans ce travail, l'utilisation des tags est analysée sur le site de musique `last.fm`, où les pistes musicales sont filtrées en fonction des classements (votes) personnels de l'utilisateur. Cette méthode se heurte au problème de l'initialisation du profil des nouveaux utilisateurs (*cold start*) qui reçoivent d'abord des recommandations peu pertinentes. (Firan *et al.,* 2007) utilisent les tags pour construire des profils pour `last.fm` en utilisant la musique déjà présente sur l'ordinateur de l'utilisateur pour contourner le problème du *cold start*. Cette amélioration peut s'apparenter à l'utilisation d'une connaissance locale pour améliorer le comportement de l'algorithme. Les auteurs ont montré que l'utilisation des tags dans les profils aboutissait à de meilleures recommandations par rapport à une utilisation de profils basés sur les chansons écoutées par les utilisateurs, mais remarquent un besoin de désambiguïsation des tags.

Les approches intégrant une composante implicite dans le profil utilisateur proposent de l'enrichir avec les usages et les comportements de l'utilisateur. En effet, le profil d'un utilisateur est également défini par son environnement ou contexte. (Dey *et al.,* 2001) décrivent le contexte comme l'ensemble des informations qui peuvent être utilisées pour caractériser la situation d'une entité, comme par exemple le réseau d'amis et les ressources annotées par l'utilisateur.

Les données issues des médias sociaux peuvent alimenter le contenu d'un profil utilisateur dans sa composante aussi bien explicite qu'implicite. (Abel *et al.,* 2011a) proposent un framework de modélisation du profil utilisateur en se basant sur ses

activités sur différents médias sociaux tels que *Flickr*[5], *Twitter* et *Delicious*[6]. Ils remarquent que cette méthode permet d'améliorer la qualité des recommandations dans un système n'ayant que peu d'information sur ses utilisateurs. Dans (Abel *et al.,* 2011b), les mêmes auteurs utilisent des tweets[7] pour modéliser les intérêts d'un utilisateur. Ils analysent le contenu posté par l'utilisateur sur Twitter : les tweets et hashtags[8], ainsi que les liens inclus dans les tweets pour caractériser leur contexte.

Afin de pouvoir exploiter les traces des interactions des utilisateurs avec les applications Web, le format standardisé, Activity Streams[9], associe des métadonnées aux actions réalisées par un utilisateur afin de les différencier les unes des autres et leur donner plus de sens. Il est basé sur le schéma **acteur verbe objet cible**. Par exemple, *un utilisateur publie une ressource*, *un utilisateur associe un tag à une ressource*, ou *un utilisateur est en contact avec un autre utilisateur*.

Plusieurs travaux ont étudié les posts et messages issus de médias sociaux, dans le cadre le Social CRM. (Ajmera *et al.,* 2013) définissent un système prenant en compte différents paramètres comme l'intention d'un post ou la nature de son auteur pour identifier les posts pertinents pour une entreprise. (Wu *et al.,* 2009) proposent un nouveau framework pour le CRM transformant les méthodes traditionnelles des CRM, qui se basent sur les individus, en méthodes se basant sur des groupes d'utilisateurs émergeant de l'analyse des réseaux sociaux.

Les méthodes de détection de communautés utilisent principalement les pages Web et les documents pour construire les communautés. Le comportement d'un utilisateur, ses actions et centres d'intérêts ne sont pas pris en compte dans la construction des communautés. La construction des profils avec les tags est généralement basée sur des tags définis par les utilisateurs et donc non contrôlés par le système, ce qui pose des problèmes, au niveau de l'homogénéité de ces tags et de leur variabilité sémantique.

3. Modèle du profil et construction de communautés

Nous présentons dans cette section notre solution pour l'amélioration de la relation client en étendant la notion de réseau social d'entreprise aux clients en prenant en compte les intérêts et usages des utilisateurs ainsi que leur propre réseau de contacts. Notre approche est constituée de trois parties : 1) la spécification d'une architecture générale du système ; 2) la modélisation et la méthode de construction du profil utilisateur et 3) l'utilisation d'algorithmes pour la détection de communautés de clients. Les utilisateurs de DisCoCRM sont les *community manager*. Les clients interagissant avec le SI de l'entreprise mais aussi avec les réseaux sociaux grand public constituent les sources des données de DiscoCRM.

---

5. `https://secure.flickr.com/`
6. `https://delicious.com/`
7. Message envoyé via Twitter
8. Mot clé, symbolisé par un # sur Twitter et Facebook, utilisé pour catégoriser un message
9. `http://activitystrea.ms/`

3.1. *L'approche DisCoCRM*

L'architecture générale d'un *Social CRM* doit prendre en compte les interconnexions entre une entreprise, ses ressources et les utilisateurs. Par conséquent, il est nécessaire de modéliser les ressources et les interactions des utilisateurs, entre eux et avec les ressources, au travers de la notion de profil utilisateur. Ce profil sera ensuite exploité par un mécanisme de détection de communautés. Les figures 1 et 2 présentent l'architecture générale de DisCoCRM, composée de trois parties distinctes :

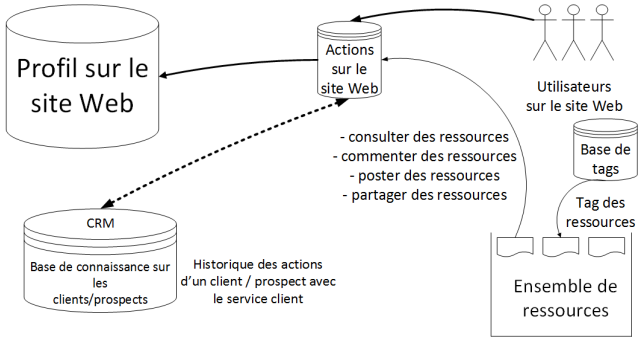

Figure 1. *Construction du profil utilisateur sur le site Web d'une entreprise*

1) *Le site Web* dédié à l'entreprise (figure 1) contient un ensemble de ressources, catégorisées via une base de tags. Chaque ressource est associée à ou plusieurs tags qui représentent des catégories. L'historique des actions d'un utilisateur sur le site Web permet d'étudier son comportement. Il peut consulter, partager, commenter des ressources existantes, mais aussi en poster de nouvelles. Les actions des utilisateurs sur le site Web permettent de construire le profil utilisateur sur le site Web.

2) *Les média sociaux grand public*, tels que Facebook et Twitter (figure 2), sont utilisés pour affiner le profil utilisateur avec des informations qui ne sont pas disponibles à l'intérieur du SI de l'entreprise. Sur ces réseaux, on s'intéresse aux interactions utilisateur suivantes : un utilisateur peut poster, partager ou mettre en favori des ressources, il peut également créer des ressources taguées (par exemple avec un hashtag sur Twitter) ou poster des liens vers d'autres sites Web, et posséder une liste de contacts au sein des réseaux sociaux. On suppose qu'il y a une intersection entre l'ensemble des tags utilisés dans les média sociaux et l'ensemble des tags utilisé dans le site Web de l'entreprise. Ainsi, les actions des utilisateurs sur les réseaux sociaux effectuées sur des ressources taguées avec des tags du SI de l'entreprise sont utilisées pour construire le profil utilisateur sur les médias sociaux.

3) *Le CRM de l'entreprise*, contenant une base de connaissances sur les clients et prospects de l'entreprise (en bas à gauche sur la figure 1), est aussi utilisé pour compléter les profils des utilisateurs connus et identifiés dans le SI de l'entreprise, et pour faire le lien entre les différentes identités d'un même utilisateur, soit en utilisant des services tels que *Facebook connect* ou *Google+ sign-in* ou de fédérations

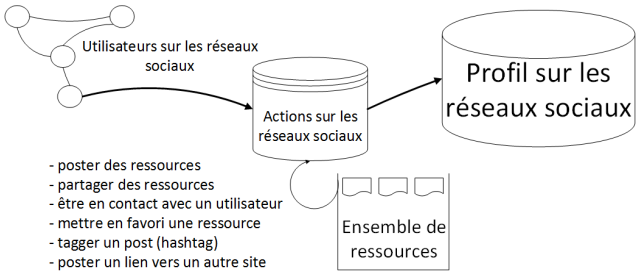

Figure 2. *Construction du profil utilisateur sur les réseaux sociaux*

d'identités tels que *OpenID*, ou encore des algorithmes de mise en correspondance (Li *et al.,* 2011). Dans l'expérimentation décrite dans la section 4, le prototype de DisCoCRM contient une déclaration explicite des correspondances d'identité entre un compte de l'application web et un compte Twitter. Nous supposons que cette base de connaissances contient des informations générales comme une adresse, un numéro de téléphone, une liste de produits achetés ou de problèmes déjà rencontrés par l'utilisateur.

Après avoir défini l'ensemble des ressources à notre disposition, nous allons détailler les différentes étapes de la construction de ce profil utilisateur.

3.2. *Construction du profil utilisateur*

La construction du profil utilisateur est basée sur les intérêts d'un utilisateur vis-à-vis des ressources du système, décrits manière explicite et/ou implicite. Nous utilisons pour cela (i) les évaluations des ressources par un utilisateur, sous forme de notes, (ii) l'intérêt d'un utilisateur pour une ressource par son dépôt et sa consultation et (iii) le réseau de contacts de l'utilisateur.

3.2.1. *Définitions*

On considère un ensemble d'utilisateurs $U = \{u_1, \ldots, u_n\}$ et un ensemble de ressources R qui peuvent être à l'intérieur du système dédié à l'entreprise R_{int} ou à l'extérieur du système R_{ext}. Nous supposons que les utilisateurs évaluent par une note $n \in \mathbb{N}$ les ressources du système $R_{int} \subseteq R$. Les notes sont stockées dans une matrice $M : U \times R_{int}$ pour un utilisateur $u_i \in U$ et une ressource $r_j \in R_{int}$, $M(u_i, r_j) = n_{ij}$. Les ressources R_{int} sont annotées par des tags, qui sont des termes issus du thésaurus du système. On note $T = \{t_1, \ldots, t_m\}$ l'ensemble des tags, chaque ressource étant annotée avec un sous-ensemble de T. Les ressources R_{ext} sont déjà taguées lorsqu'elles sont récupérées, et on ne conserve que les tags qui sont inclus

dans T. Les tags associés aux ressources sont stockés dans une matrice $MT : R \times T$ définie comme suit, pour une ressource $r_j \in R$ et un tag $t_k \in T$:

$$MT(r_j, t_k) = \begin{cases} 1 & \textit{si } t_k \textit{ est associé à } r_j, \\ 0 & \text{sinon.} \end{cases} \quad [1]$$

L'approche DisCoCRM incluant une prise en compte des interactions entre les utilisateurs, nous utilisons les réseaux sociaux grand public comme source de données et par exemple, la liste des *amis* d'un utilisateur sur Facebook ou les liens *follower / following* de Twitter. Pour un utilisateur u_i, on distingue donc : les *followers*, qui sont les utilisateurs qui ont déclaré explicitement vouloir suivre u_i, c'est-à-dire être en contact avec u_i sans qu'il n'ait besoin de donner son accord ; des *following*, qui sont les utilisateurs que u_i déclare explicitement vouloir suivre. Dans le cadre des exemples que nous développerons dans la section suivante, nous ne nous intéressons qu'au second type d'interaction, le premier étant indépendant des actions de u_i, alors que le deuxième correspond aux personnes à qui u_i s'intéresse.

On suppose que le site Web de l'entreprise permet aux utilisateurs de définir une liste de contacts au sein même du site, c'est-à-dire un sous-ensemble de U. Les différents liens entre les utilisateurs forment alors le réseau social interne du site Web. Les liens de contacts, qu'ils soient issus des réseaux sociaux externes à l'entreprise ou de son site Web, sont stockés dans une matrice symétrique $A : U \times U$ définie comme suit, pour deux utilisateurs $u_i, u_j \in U$:

$$A(u_i, u_j) = \begin{cases} 1 & \textit{si } u_i \textit{ est en contact avec } u_j, \\ 0 & \text{sinon.} \end{cases} \quad [2]$$

3.2.2. *Construction du profil*

L'objectif de l'approche proposée est de regrouper les utilisateurs en communautés thématiquement proches, en se basant sur les ressources qu'ils apprécient. Pour cela nous construisons un profil unique à chaque utilisateur en trois étapes.

Étape 1 : définition du profil explicite.

Nous calculons le degré d'appartenance da_{ij} d'un utilisateur u_i à un tag t_j :

$$da_{ij} = \frac{|R(u_i, t_j)|}{|R|} \times \frac{\sum n_{ij}}{n_{max} \times |n_{ij}|} \quad [3]$$

– $R(u_i, t_j)$ est l'ensemble des ressources notées par l'utilisateur u_i où le tag t_j apparaît et $|R(u_i, t_j)|$ la cardinalité de cet ensemble ;

– la seconde fraction de la formule est la moyenne des notes données par l'utilisateur u_i aux ressources de $R(u_i, t_j)$ divisée par n_{max}, la note maximale donnée aux ressources par les utilisateurs, afin d'obtenir une valeur comprise entre 0 et 1.

Étape 2 : prise en compte du comportement.

Comme nous voulons prendre en compte le comportement de l'utilisateur dans le système, que ce soit sur le site Web de l'entreprise ou sur les réseaux sociaux, nous affinons l'expression de da_{ij} avec le comportement de u_i en incluant :

– son intérêt pour une ressource, c'est-à-dire : le dépôt ou la consultation d'une ressource, le tweet, re-tweet, la mise en favori sur Twitter ou le post ou partage de ressources sur Facebook ;

– son réseau social : en prenant en compte les notes données par ses contacts.

Afin d'intégrer ces différents éléments, nous définissons un degré d'appartenance d'_{ij}, utilisant l'historique de consultation de u_i sur le site Web, ainsi que ses tweets, re-tweets et favoris sur Twitter et les ressources postées et partagées sur Facebook. La formule de calcul de d'_{ij} est définie comme suit, avec pour exemple de réseau social Twitter :

$$d'_{ij} = a \times \frac{|R_{consult}(u_i, t_j)|}{|R_{consult}|} + b \times \frac{|R_{tweet}(u_i, t_j)|}{|R_{tweet}|} + c \times \frac{|R_{re-tweet}(u_i, t_j)|}{|R_{re-tweet}|} + d \times \frac{|R_{bookmark}(u_i, t_j)|}{|R_{bookmark}|} \quad [4]$$

avec

– $R_{consult}$ l'ensemble des ressources R_{int} consultées par u_i

– R_{tweet}, $R_{re-tweet}$ et $R_{bookmark}$ sont respectivement les ensembles des tweets, des re-tweets et des favoris de u_i sur Twitter

– a, b, c et d des pondérations avec $a + b + c + d = 1$

Les pondérations peuvent être utilisées pour donner un poids plus ou moins important à une partie du degré d'appartenance, en fonction du comportement des utilisateurs au sein du système. Par exemple, il est possible de mettre en avant les tweets et les ressources consultées si les utilisateurs ne re-tweetent pas et n'ont pas beaucoup de favoris sur Twitter.

Étape 3 : prise en compte des contacts.

Afin de prendre en compte les informations provenant des utilisateurs en contact avec u_i, nous utilisons le premier degré da pour chaque utilisateur u_k étant en contact avec u_i. Ainsi, pour un utilisateur u_i et un tag t_j, nous définissons dc_{ij} comme suit :

$$dc_{ij} = \frac{\sum_{k=1}^{m} da_{kj}}{m} \quad [5]$$

– $A(u_i) \subseteq U$ l'ensemble des contacts de u_i and $m = |A(u_i)|$

– da_{kj} le degré d'appartenance de l'utilisateur u_k au tag t_j, avec $u_k \in A(u_i)$

La combinaison des trois paramètres pris en compte dans notre approche (notes, consultations et actions sur les réseaux sociaux, contacts) nous permet de définir le degré d'appartenance d_{ij} d'un utilisateur u_i à un tag t_j de la manière suivante :

$$d_{ij} = \alpha \times da_{ij} + \beta \times dc_{ij} + \gamma \times d'_{ij} \quad [6]$$

– α et β et γ des pondérations avec $\alpha + \beta + \gamma = 1$

Le profil de l'utilisateur u_i, noté X_i, est le vecteur de ses degrés d'appartenance à chaque tag : $X_i = (d_{i1}, d_{i2}, ..., d_{ij})$.

3.3. *Algorithmes de détection de communautés*

Une fois les profils des utilisateurs construits, les communautés d'utilisateurs peuvent être calculées. Dans la plateforme que nous développons nous avons retenu deux types d'approches : celles qui considèrent un profil utilisateur comme un point dans un espace multi-dimensionnel et celles qui considèrent le profil comme un sommet d'un graphe pondéré. L'objectif est de fournir au *community manager* un ensemble d'outils pour affiner sa connaissance sur les communautés et pouvoir décider des actions à mener (campagne marketing ciblée par exemple). Nous nous placerons dans trois cas d'utilisations. Dans le premier cas, nous supposons que le *community manager* n'a pas de connaissance sur la population qu'il étudie, cependant il peut définir l'ensemble des mots clés qui sont pertinents dans son domaine et donner un intervalle ou bien le nombre de communautés estimé. Dans le second cas, le *community manager*, de par son expérience, a une idée des communautés : il utilise l'algorithme pour confronter sa connaissance empirique aux données recueillies mais il a besoin d'une caractérisation des communautés qui seront calculées (par exemple sous la forme de tags représentatifs). Enfin, dans le troisième cas d'utilisation, le *community manager* à une connaissance précise des termes utilisés dans son domaine, qu'il peut organiser dans une hiérarchie, qui peut être utilisée pour piloter l'algorithme. Pour ces deux types d'approches et les trois cas d'utilisations, nous avons sélectionné deux algorithmes représentatifs : l'algorithme K-means et la méthode de Louvain (Blondel *et al.,* 2008).

La classification par l'algorithme K-means est une des techniques de classification non supervisées les plus utilisées. Nous l'avons retenue car elle converge rapidement après quelques itérations. Cela permet d'effectuer plusieurs simulations, avec un nombre de classes différent, et de laisser le choix au *community manager* d'interpréter les résultats en fonction du contexte. L'espace d'utilisation du K-means est à m-dimensions, m étant le nombre de tags présents dans le thésaurus. Chaque axe possède une échelle entre 0 et 1 correspondant au degré d'appartenance de l'utilisateur au tag associé à cet axe. Le barycentre de la communauté peut être utilisé comme utilisateur "type" de cette communauté. Les barycentres des communautés seront utilisés lorsque de nouveaux utilisateurs seront ajoutés au système afin de les classer dans la communauté la plus proche de leur profil.

La méthode de Louvain, en utilisant une représentation sous forme de graphe, permet d'introduire les éléments sémantiques nécessaires au troisième cas d'utilisation. Elle a l'avantage d'être très rapide, mais ne donne pas forcément les partitions optimales du graphe. Elle permet de faire ressortir les tags d'une communauté.

L'introduction des pondérations (formules 4 et 6) permet la prise en compte des différents aspects du comportement de l'utilisateur et d'en mettre un ou plusieurs en avant par rapport à d'autres. Elles nous permettent aussi d'obtenir un outil générique capable de s'adapter au contexte d'utilisation et aux souhaits du *community manager*. Cependant, trouver les "bonnes" pondérations n'est pas simple. Il est possible de les fixer de manière arbitraire, ou d'utiliser des *templates* prédéfinis ayant été adaptés au domaine métier. Toutefois, le *community manager* peut avoir une connaissance *a priori* des membres de son réseau et de ses communautés. Cette connaissance peut être prise en compte en utilisant des méthodes d'apprentissage améliorant les pondérations en fonction du contexte. Ainsi, le *community manager* pourra choisir une communauté en précisant qu'elle ne lui convient pas, pour plusieurs raisons : ses utilisateurs ne sont pas assez représentatifs, il y a des utilisateurs qui, pour lui, ne doivent pas être dans cette communauté. La méthode d'apprentissage se chargera de recalculer les pondérations et les communautés en s'adaptant aux informations données par le *community manager*, qui validera les nouvelles communautés. Les pondérations seront sauvegardées comme étant celles s'appliquant le mieux au contexte. Elles pourront continuer d'évoluer par la suite en fonction du retour d'information de l'utilisateur.

4. Expérimentations

Nous avons testé notre approche sur un jeu d'essai restreint. Celui-ci prend la forme d'une étude du comportement d'utilisateurs sur application de type base de connaissances. Celle-ci permet de stocker un ensemble de connaissances spécifiques à un domaine donné, dans notre cas les thématiques du goût, de la nutrition et de la santé. Ces connaissances sont organisées sous la forme d'un thésaurus. Notre jeu d'essai contient environ 20 utilisateurs, 30 tags et 50 ressources. La société gérant cette base a développé une plateforme qui s'apparente à un CRM dans le sens où les utilisateurs de l'application, ou clients, sont des industriels, des laboratoires de recherche et des experts du domaine de l'agro-alimentaire. L'application peut être considérée comme une plateforme d'échanges, où les utilisateurs ont la possibilité de chercher, poster, noter, annoter (de manière privée) et commenter (de manière publique) les ressources et se créer un réseau de contacts au sein de la plateforme. Nous nous intéressons uniquement aux actions de poster, consulter et noter une ressource. Ces ressources sont des articles de recherche, des synthèses d'études, des comptes-rendus de réunion, au format PDF. Au niveau du réseau social d'un utilisateur, nous ne prenons en compte que la partie interne de ce réseau. Le comportement des utilisateurs est assez variable, certains ne parlent que de domaines très pointus, alors d'autres restent très généraux et ne s'intéressent qu'aux branches hautes du thésaurus. La plupart des utilisateurs s'intéressent à plusieurs thèmes, mais consultent des ressources dont le thème n'est

pas obligatoirement lié à leur intérêt. Les ressources sont évaluées par les utilisateurs via un système de notation. Les notes possibles sont comprises entre 1 et 5. Si un utilisateur n'a pas noté une ressource, la note est de 0. Chaque ressource est annotée avec un ensemble de tags issus d'un thésaurus, dont un extrait est représenté dans la figure 3.

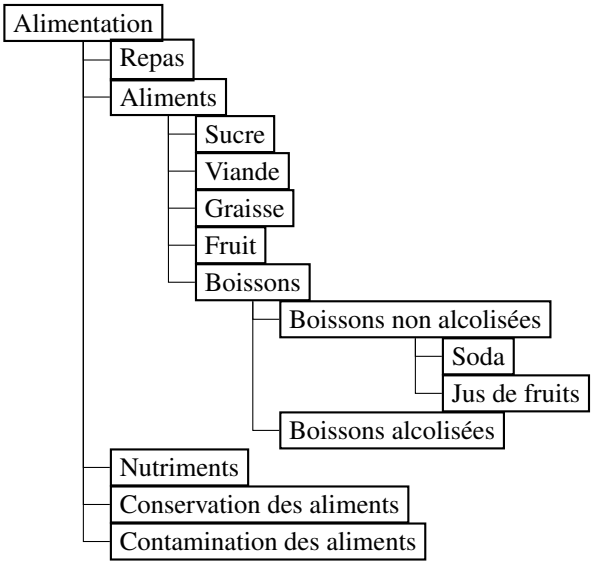

Figure 3. *Extrait de thésaurus dans le domaine alimentaire*

L'expérimentation comporte deux étapes de construction : 1) celle du profil utilisateur en calculant ses degrés d'appartenance et 2) celle des communautés d'utilisateurs.

4.1. *Construction du profil utilisateur*

Nous avons construit une matrice M des notes attribuées par chaque utilisateur pour chaque ressource. Le tableau 2 montre un extrait de cette matrice pour 4 utilisateurs et 5 ressources. Chaque ressource étant annotée par plusieurs tags, nous avons la liste des ressources avec les tags qui lui sont associés. Le tableau 1 montre un extrait de ces annotations. Nous avons ensuite construit la matrice M_d des degrés d'appartenance des utilisateurs aux différents tags, en nous basant uniquement sur le profil explicite des utilisateurs, c'est-à-dire uniquement sur les notes qu'ils ont attribuées aux ressources. Le tableau 3 montre un extrait de ces résultats pour 4 utilisateurs et 4 tags. Le profil d'un utilisateur est le vecteur de ses degrés d'appartenance à chaque tag. Puis nous avons affiné le calcul du profil des utilisateurs en prenant en compte leur comportement implicite, c'est-à-dire les consultations des ressources et leur réseau de contacts en interne. Le tableau 4 illustre le degré d'appartenance qui servira de profil

en prenant en compte ces deux paramètres. Nous avons expérimenté avec deux pondérations différentes, une première privilégiant les notes données par les utilisateurs, et une deuxième donnant plus de poids à son comportement. Pour la première expérimentation, nous avons donc fixé α = 0.6, β = 0.3 et γ = 0.1, pour la seconde α = 0.5, β = 0.3, γ = 0.2. Le tableau 4 donne un exemple de profil pour la première expérimentation. Pour rappel, la pondération α correspond aux notes données par l'utilisateur, β à son comportement, et γ à son réseau social.

Ressources	Tags
R_1	Repas, Viande, Graisse
R_2	Nutriments
R_3	Contamination des aliments, Viande, Fruit
R_4	Conservation des aliments, Viande, Fruit
R_5	Sucre, Fruit, Soda, Jus de fruits

Tableau 1. *Exemple d'annotation de ressources*

	R_1	R_2	R_3	R_4	R_5
u_1	4	0	5	4	5
u_3	5	4	4	3	4
u_{10}	1	0	0	0	0
u_{11}	5	3	4	5	3

Tableau 2. *Exemple de notes de ressources*

	Repas	Nutriments	Viande	Graisse
u_1	0.0008	0.0001	0.0251	0.0130
u_3	0.0143	0.0057	0.0076	0.0011
u_{10}	0.0005	0.0002	0.0034	0
u_{11}	0.0074	0.0032	0.0156	0.0025

Tableau 3. *Profil utilisateur explicite*

	Repas	Nutriments	Viande	Graisse
u_1	0.0215	0.0019	0.0415	0.0242
u_3	0.0319	0.0112	0.0275	0.0199
u_{10}	0.0003	0.0002	0.0096	0
u_{11}	0.0096	0.0173	0.0341	0.0054

Tableau 4. *Profil utilisateur explicite et implicite*

Nous pouvons déjà remarquer par exemple que le d_{ij} de l'utilisateur u_1 pour le tag $Repas$ a augmenté. En effet, l'existence d'un lien indirect entre les utilisateurs u_1 et u_{10} et le degré d'appartenance de l'utilisateur u_{10} au tag $Repas$; et u_1 consulte régulièrement des documents annotés par ce tag.

4.2. *Détection de communautés et bilan de l'expérimentation*

Pour tous les tableaux représentant les communautés, la première colonne, notée C#, correspond aux différentes communautés avec leur numéro. Les communautés affinées illustrées dans les tableaux 5, 6, 7, 8 et 9 ont pour différence une pondération plus forte pour l'historique de consultation, pour la colonne *Affinées 2* et une pondération plus forte pour les ressources notées, pour la colonne *Affinées*. Cela influe sur les degrés et donc sur les communautés qui en résultent.

4.2.1. *Méthode des K-Means*

Pour la construction de communautés d'utilisateurs, nous avons utilisé le logiciel libre de data-mining WEKA (Witten *et al.,* 2005) qui nous a donné les résultats présentés dans le tableau 5. Les représentants des communautés sont en italique. Après plusieurs essais, de l'algorithme avec un nombre différent de communauté, la valeur de 5 communautés est apparue raisonnable pour les trois profils. Dans la suite de la section, nous interprétons les résultats (tableau 5).

u_1, u_3 et u_{11} notent des ressources sur des thèmes proches au niveau de la structure du thésaurus, mais différents : Viande, Graisse, Obésité pour u_1 ; Repas, Nutriments, Jus de Fruits pour u_3 ; et u_{11} s'intéresse à la partie générale Alimentation du thésaurus sans se spécialiser dans un ou plusieurs domaines spécifiques.

u_1 a consulté 14 ressources et u_3 13 ressources. Parmi les ressources consultées par u_1, 4 ont été annotées avec le tag Repas, et pour u_3 2 ont été annotées avec le tag Repas, ce qui fait augmenter leur degré d'appartenance respectifs à ce tag. u_1 est en contact avec u_3 et u_{10}, qui ont tous les deux un degré d'appartenance non nul au tag Repas, ce qui fait augmenter le degré d'appartenance de u_1 au tag Repas. u_3 est intéressé par le tag Nutriments, alors que u_1 ne l'est pas. Comme u_1 et u_3 sont en contact et que u_1 a consulté quelques ressources associées au tag Nutriments, le degré d'appartenance de u_3 à ce tag augmente en affinant les degrés. u_{11} est lui en contact avec u_2, dont les intérêts sont différents : Infection, Bactérie, Parasite, etc.

En observant le comportement des utilisateurs, on remarque que u_1, u_3 et u_{11} partagent les mêmes intérêts, principalement les thèmes *Repas*, *Graisse*, *Obésité* et *Viande*. Il semble donc naturel que ces trois utilisateurs soient regroupés dans la même communauté. En utilisant les degrés non affinés, ces utilisateurs sont dans des communautés séparées. Cependant, en construisant les communautés avec les degrés affinés, ces utilisateurs se retrouvent dans la même communauté.

u_{14} se retrouve seul dans une communauté avec les degrés affinés alors que ce n'était pas le cas avec les degrés non affinés. Il ne consulte pas beaucoup de ressources, et pas les mêmes que u_3 et $_{20}$ qui étaient dans sa communauté, ce qui fait diminuer les degrés d'activité des tags qu'ils ont en commun. De plus, il est en contact avec des utilisateurs qui ont des intérêts complètement différents des siens.

u_8, u_{17}, u_{18} ne sont pas dans la même communauté que u_1 dans la colonne *Affinées*. Cependant, ils consultent beaucoup de ressources consultées aussi par les uti-

lisateurs de la communauté de u_1, et sur des thématiques proches des intérêts de la communauté de u_1. Ils sont donc passés dans la communauté de u_1.

L'utilisation des contacts d'un utilisateur et de son historique de navigation permet d'affiner le profil et de regrouper des utilisateurs au comportement similaire dans la même communauté, ce qui n'était pas toujours le cas avec un profil prenant en compte uniquement les notes données par un utilisateur à des ressources.

C#	Profil explicite	α = 0.6 et β = 0.3 et γ = 0.1 (Affinées)	α = 0.5 et β = 0.3 et γ = 0.2 (Affinées 2)
1	*U1*, U4, U12	*U1*, **U3**, U4, **U11**, U12, U13, U20	*U1*, U3, U4, **U8**, U11, U12, U13, **U17**, **U18**, U20
2	*U2*, U6, U7, U9, U10, U13, U15, U18	*U2*, U6, U7, U8, U9, U10, U15, U18	*U2*, U6, U7, U9, U10, U15
3	*U3*, U14, U20	*U5*, U16, U17	*U5*, U16
4	*U5*, U8, U11, U16, U17	*U14*	*U14*
5	*U19*	*U19*	*U19*

Tableau 5. *Communautés non affinées et affinées avec K-Means*

4.2.2. *Méthode de Louvain*

Nous avons ensuite utilisé la méthode de Louvain, ne nécessitant pas de définir le nombre de communautés *a priori* et visant à maximiser la modularité. On peut remarquer que le nombre de communautés varie en fonction des profils utilisés, passant de 5 à 6. Le tableau 6 présente le résultat de la méthode de Louvain pour les communautés d'utilisateurs et le tableau 7 présente la liste des tags dont *parle* une communauté, obtenue par l'algorithme.

Des similarités existent entre les communautés trouvées par le k-means et celles trouvées par la méthode de Louvain. En effet, pour le profil explicite, u_1 et u_{12} sont dans la même communauté, u_2, u_6, u_9 et u_{10} aussi ainsi que u_5 et u_{17}. Dans le cas des communautés *Affinées*, u_2, u_6, u_8, u_9 et u_{18} sont dans la même communauté, ainsi

C#	Profil explicite	α = 0.6 et β = 0.3 et γ = 0.1 (Affinées)	α = 0.5 et β = 0.3 et γ = 0.2 (Affinées 2)
1	U2, U6, U8, U9, U10	U7, U15	U4, U12, U14
2	U5, U17	U1, U2, U6, U8, U9, U18, U20	U2, U5, U6, U8, U9, U18
3	U4, U18, U20	U3, U4, U11, U12, U13, U14, U17	U3, U11, U13
4	U1, U11, U12	U10, U16, U19	U1, U20
5	U3, U14, U16, U19	U5	U16, U17, U19
6	U7, U13, U15		U7, U10, U15

Tableau 6. *Communautés non affinées et affinées avec Louvain*

que u_3, u_4 et u_{11} . Dans les cas des communautés *Affinées 2*, u_2, u_6, u_9 sont dans la même communauté.

Cependant, u_5 et u_{17} n'ont qu'un seul intérêt en commun (Maladie de la nutrition), mais sont dans la même communauté lorsque l'on utilise les profils explicites, que ce soit avec la méthode de Louvain ou avec le k-means. Avec la méthode de Louvain, ces deux utilisateurs ne sont plus dans la même communauté lorsque l'on utilise les profils explicites et implicites, alors qu'avec le k-means, dans le cas des communautés *Affinées* ils restent toujours dans la même communauté.

Il y a des groupes d'utilisateurs qui sont toujours dans la même communauté quelles que soient les pondérations : par exemple u_2, u_6, u_8 et u_9, intéressés par les tags Contamination, Conservation, Infection. Ces tags sont bien ressortis par la méthode de Louvain dans le tableau 7.

Dans le cas des utilisateurs u_1, u_3 et u_{11}, u_1 et u_{11} sont dans la même communauté lorsque seul le profil explicite est pris en compte. Cependant, lorsque l'on introduit la prise en compte du comportement, u_3 et u_{11} sont dans la même communauté alors que u_1 est dans une communauté différente.

4.2.3. *Méthode de Louvain pilotée par une connaissance du domaine*

Nous avons ensuite utilisé la hiérarchie de termes du thésaurus pour piloter la méthode de Louvain en traduisant le thésaurus en graphe pondéré de manière inverse à la profondeur. Ainsi, pour notre domaine, un utilisateur qui annote un document avec un tag précis cumule aussi les pondérations des termes plus généraux. Le tableau 8 présente le résultat de la méthode de Louvain pilotée, au niveau des communautés d'utilisateurs, et le tableau 9 présente la liste des tags dont *parle* une communauté. On remarque que le nombre de communautés a été réduit à 4. Les communautés sont plus denses et les tags des communautés reprennent globalement la hiérarchie du thésaurus.

u_6, u_8 et u_9 notent des ressources avec des tags en commun, par exemple Contamination des aliments. Cependant, la majorité de leurs actions se fait sur des ressources différentes. u_6 consulte principalement des ressources dont les tags sont Conservation des aliments, Fruit, Légumes, Soda et Jus de Fruits. u_8 consulte beaucoup de ressources avec le tag Contamination des aliments. u_9 consulte des ressources sur l'infection, les bactéries, la grippe et la rubéole entre autres. u_6 est aussi en contact avec u_4, qui lui note des ressources sur les thèmes Soda et Jus de Fruits.

u_6, u_8 et u_9 sont dans la même communauté en utilisant les profils explicites. Lorsque l'on introduit les profils implicites, leur différence de comportement fait qu'ils sont les trois dans des communautés séparées. u_6 passe donc dans la communauté 2 qui s'intéresse aux thèmes Soda, Jus de Fruits, etc. u_8 reste dans la communauté s'intéressant à la Contamination des aliments.

On peut aussi remarquer que u_5 et u_{17} ne sont plus dans la même communauté avec le profil explicite. Dans ce cas, u_{17} est dans la même communauté que u_{13}, et partagent des thèmes communs : hémopathie et immunopathologie.

C#	Profil explicite	α = 0.6 et β = 0.3 et γ = 0.1 (Affinées)	α = 0.5 et β = 0.3 et γ = 0.2 (Affinées 2)
1	Contamination-des-aliments, Bactérie, Parasite, Virus, Grippe, Rubéole	Alimentation, Aliments, Allergie, Facteur-allergène	Sucre, Boissons-non-alcoolisées, Soda, Jus-de-fruits
2	Nutriments, Maladie, Maladie-de-la-nutrition, Infection	Contamination-des-aliments, Conservation-des-aliments, Viande, Bactérie, Parasite, Virus, Grippe, Rubéole	Nutriments, Contamination-des-aliments, Conservation-des-aliments, Boissons-alcoolisées, Maladie-de-la-nutrition, Infection, Bactérie, Virus, Grippe
3	Conservation-des-aliments, Boissons, Boissons-non-alcoolisées, Soda, Boissons-alcoolisées, Immunopathologie	Repas, Sucre, Graisse, Boissons, Boissons-non-alcoolisées, Soda, Jus-de-fruits, Boissons-alcoolisées, Maladie, Obésité	Repas, Graisse, Obésité, Hémopathie
4	Aliments, Sucre, Viande, Graisse	Légumes, Carence-alimentaire	Viande, Parasite
5	Fruit, Légumes, Jus-de-fruits, Carence-alimentaire, Obésité	Nutriments, Maladie-de-la-nutrition, Infection	Fruit, Légumes, Boissons, Maladie, Carence-alimentaire, Immunopathologie
6	Alimentation, Repas, Hémopathie, Allergie, Facteur-allergène		Alimentation, Aliments, Allergie, Facteur-allergène, Rubéole

Tableau 7. *Tags des communautés non affinées et affinées avec Louvain*

Dans le cas des utilisateurs u_1, u_3 et u_{11}, les 3 utilisateurs sont dans la même communauté lorsque seul le profil explicite est pris en compte et dans le cas des communautés *Affinées*. Cependant, pour les communautés *Affinées 2*, u_{11} n'est plus dans la même communauté qu'u_1 et u_3. Les 3 utilisateurs notent des ressources qui se situent dans la même partie du thésaurus. Le pilotage de la méthode de Louvain avec la hiérarchie du thésaurus permet de regrouper ces 3 utilisateurs dans la même communauté en utilisant uniquement le profil explicite, ce qui n'était pas le cas avant. Pour les communautés *Affinées 2*, u_1 et u_3 étant en contact ; et u_{11} étant en contact avec u_2 aux intérêts complètement différents, l'accent mis sur le réseau social laisse u_1 et u_3 dans la même communauté alors que u_{11} change de communauté.

4.2.4. *Bilan de l'expérimentation*

Notre approche permet de regrouper les utilisateurs dans des communautés plus pertinentes que si l'on utilisait un profil simple, sans analyse du comportement et du réseau de contacts d'un utilisateur. Les différents paramètres de pondération des

C#	Profil explicite	α = 0.6 et β = 0.3 et γ = 0.1 (Affinées)	α = 0.5 et β = 0.3 et γ = 0.2 (Affinées 2)
1	U5, U6, U8, U9, U20	U5, U8, U20	U5, U8, U11, U20
2	U1, U3, U4, U11, U12, U14, U15, U19, U18, U19	U1, U3, U4, U6, U11 , U12, U14, U16, U18, U19	U1, U3, U4, U6, U12, U14, U16, U18, U19
3	U7, U13, U17	U2, U17	U13, U17
4	U2, U10	U7, U19, U10, U13, U15	U2, U7, U19, U10, U15

Tableau 8. *Communautés non affinées et affinées avec Louvain pilotée*

C#	Profil explicite	α = 0.6 et β = 0.3 et γ = 0.1 (Affinées)	α = 0.5 et β = 0.3 et γ = 0.2 (Affinées 2)
1	Alimentation, Repas, Nutriments, Contamination-des-aliments	Alimentation, Repas, Nutriments, Contamination-des-aliments	Alimentation, Repas, Nutriments, Contamination-des-aliments
2	Conservation-des-aliments, Aliments, Sucre, Viande, Graisse, Fruit, Légumes, Boissons, Boissons-non-alcoolisées, Soda, Jus-de-fruits	Conservation-des-aliments, Aliments, Sucre, Viande, Graisse, Fruit, Légumes, Boissons, Boissons-non-alcoolisées, Soda, Jus-de-fruits	Conservation-des-aliments, Aliments, Sucre, Viande, Graisse, Fruit, Légumes, Boissons, Boissons-non-alcoolisées, Soda, Jus-de-fruits
3	Boissons-alcoolisées, Maladie, Maladie-de-la-nutrition, Carence-alimentaire, Obésité, Hémopathie, Immunopathologie, Allergie, Facteur-allergène, Infection, Bactérie	Maladie, Maladie-de-la-nutrition, Carence-alimentaire	Boissons-alcoolisées, Maladie, Maladie-de-la-nutrition, Carence-alimentaire, Obésité, Hémopathie, Immunopathologie, Allergie
4	Parasite, Virus, Grippe, Rubéole	Boissons-alcoolisées, Obésité, Hémopathie, Immunopathologie, Allergie, Facteur-allergène, Infection, Bactérie, Parasite, Virus, Grippe, Rubéole	Facteur-allergène, Infection, Bactérie, Parasite, Virus, Grippe, Rubéole

Tableau 9. *Tags des communautés non affinées et affinées avec Louvain pilotée*

profils affinés permettent de découvrir des communautés à la volée en fonction des critères que l'on souhaite mettre en avant : basées plus sur les notes des utilisateurs, sur leur liste de contacts ou sur les consultations et dépôts. Cela permet de privilégier plus ou moins un aspect du comportement de l'utilisateur en fonction du contexte et

du type de communauté souhaité. Il est aussi possible de faire varier les communautés en utilisant différentes valeurs de classes pour l'algorithme du k-means.

La méthode de Louvain a l'avantage d'être très rapide dans son exécution, mais ne garantit pas de trouver la solution optimale. Les communautés trouvées grâce à cette méthode varient de celles trouvées avec le k-means.

Le pilotage de la méthode de Louvain par la hiérarchie de termes du thésaurus permet de réduire le nombre de communautés et de regrouper les utilisateurs en fonction des concepts généraux du thésaurus. Cela permet de limiter le problème des utilisateurs ne s'intéressant qu'aux feuilles du thésaurus pouvant se retrouver dans une mauvaise communauté.

Afin d'automatiser les traitements que peut demander le *community manager*, nous avons développé DisCoCRM sous la forme d'une application Web, dont les écrans illustrés sur les figures 4, 5, présentent respectivement les grandes fonctionnalités accessibles depuis la page d'accueil de l'application ainsi que la définition des pondérations et le choix des algorithmes.

Figure 4. *Grandes fonctionnalités du prototype de DisCoCRM*

5. Conclusion et perspectives

Dans cet article, nous avons présenté une approche pour la détection de communautés et une plateforme destinée aux *community managers* dans le contexte de la gestion de la relation client. La détection des communautés exploite des profils utilisateur basés sur les usages, les comportements et les contacts. Les données du profil sont

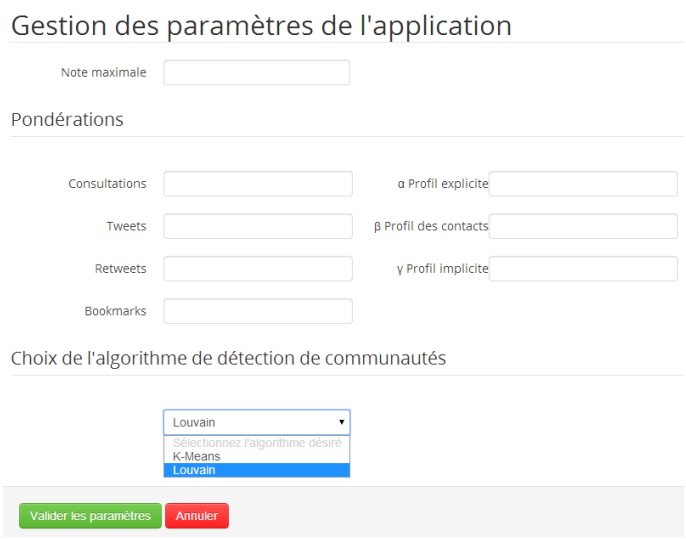

Figure 5. *Gestion des paramètres pour la constitution du profil*

collectées à partir des applications Web de l'entreprise et à partir des réseaux sociaux. Les composantes du profil sont modulables au moyen de pondérations et aboutissent à la notion de profils affinés, permettant ainsi de détecter des communautés en fonction des critères que l'on souhaite voir renforcés. Nous avons également inclus dans la plateforme deux catégories d'algorithmes dont la complexité permet d'envisager le traitement du volume de données nécessaire au CRM : l'un opérant une classification dans un espace multi-dimensionnel (k-means) ; l'autre (la méthode de Louvain) travaillant sur des graphes pondérés permettant de faire émerger les caractéristiques des communautés et de piloter la détection par la connaissance du domaine. Une série de trois expérimentations sur des jeux de données test a été présentée.

Nos perspectives se déclinent en trois niveaux : modélisation, algorithmique, analyse des échanges. Nous envisageons donc l'utilisation de graphes hétérogènes, analysés via une algèbre de chemin prenant en compte la richesse sémantique des relations ; ainsi que l'utilisation d'algorithmes pour détecter des communautés recouvrantes. En effet, dans le cadre du CRM, les clients d'une entreprise s'intéressent souvent à différents produits ou services et peuvent facilement appartenir à plusieurs communautés. L'analyse des échanges au sein d'une communauté ou d'un réseau social est essentielle pour le CRM car elle permet de savoir qui sont les utilisateurs influents et sur qui leur avis influe. Cette analyse peut utiliser les propriétés locales d'un nœud pour évaluer sa connectivité intra et inter-communauté et donner une indication de son influence.

Du point de vue du prototype, nous pensons étendre la collecte de données à partir d'autres réseaux sociaux et à partir de forums. L'étude d'un autre réseau de micro-

blogging comme *Tumblr*[10] peut permettre d'accroitre la connaissance sur les intérêts d'un utilisateur. En effet, *Tumblr* propose à un utilisateur de suivre d'autres utilisateurs, partager du contenu, *reblogger* du contenu qu'il trouve intéressant, mettre en avant du contenu d'autres utilisateurs.

Remerciements

Ce travail est réalisé dans le cadre d'une bourse CIFRE numéro 2012 / 0261, financé par l'entreprise eb-Lab.

6. Bibliographie

Abel F., Araújo S., Gao Q., Houben G.-J., « Analyzing cross-system user modeling on the social web », *Web Engineering*, Springer, p. 28-43, 2011a.

Abel F., Gao Q., Houben G.-J., Tao K., « Semantic enrichment of twitter posts for user profile construction on the social web », *The Semanic Web : Research and Applications*, Springer, p. 375-389, 2011b.

Ajmera J., Ahn H.-i., Nagarajan M., Verma A., Contractor D., Dill S., Denesuk M., « A CRM system for social media : challenges and experiences », *Proc. of the 22nd international conference on World Wide Web*, International World Wide Web Conferences Steering Committee, p. 49-58, 2013.

Blondel V. D., Guillaume J.-L., Lambiotte R., Lefebvre E., « Fast unfolding of communities in large networks », *Journal of Statistical Mechanics : Theory and Experiment*, vol. 2008, n° 10, p. P10008, 2008.

Cattuto C., Baldassarri A., Servedio V., Loreto V., « Emergent community structure in social tagging systems », *Advances in Complex Systems*, vol. 11, n° 04, p. 597-608, 2008.

Cohen S., Kimelfeld B., Koutrika G., « A Survey on Proximity Measures for Social Networks », *Search Computing*, vol. 7538, p. 191-206, 2012.

Cordina P., Fayon D., *Community management : fédérer des communautés sur les médias sociaux*, Pearson Education France, 2013.

Dey A. K., Abowd G. D., Salber D., « A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications », *Hum.-Comput. Interact.*, vol. 16, n° 2, p. 97-166, December, 2001.

Firan C. S., Nejdl W., Paiu R., « The benefit of using tag-based profiles », *Web Conference, 2007. LA-WEB 2007. Latin American*, IEEE, p. 32-41, 2007.

Fortunato S., « Community detection in graphs », *Physics Reports*, vol. 486, n° 3, p. 75-174, 2010.

Girvan M., Newman M. E., « Community structure in social and biological networks », *Proc. of the National Academy of Sciences*, vol. 99, n° 12, p. 7821-7826, 2002.

Golbeck J., « Trust and nuanced profile similarity in online social networks », *ACM Transactions on the Web (TWEB)*, vol. 3, n° 4, p. 12, 2009.

10. `https://www.tumblr.com/`

Hung C.-C., Huang Y.-C., Hsu J. Y.-j., Wu D. K.-C., « Tag-based user profiling for social media recommendation », *Workshop on Intelligent Techniques for Web Personalization and Recommender Systems at AAAI*, 2008.

Li J., Wang G. A., Chen H., « Identity matching using personal and social identity features », *Information Systems Frontiers*, vol. 13, n° 1, p. 101-113, 2011.

Melville P., Sindhwani V., Lawrence R. D., « Social Media Analytics : Channeling the Power of the Blogosphere for Marketing Insight », 2009.

Mohan S., Choi E., Min D., « Conceptual modeling of enterprise application system using social networking and web 2.0 ?social CRM system ? », *Convergence and Hybrid Information Technology, 2008. ICHIT'08. Int. Conference on*, IEEE, p. 237-244, 2008.

Newman M. E., « Modularity and community structure in networks », *Proc. of the National Academy of Sciences*, vol. 103, n° 23, p. 8577-8582, 2006.

Newman M. E., Girvan M., « Finding and evaluating community structure in networks », *Physical review E*, vol. 69, n° 2, p. 026113, 2004.

Palla G., Derényi I., Farkas I., Vicsek T., « Uncovering the overlapping community structure of complex networks in nature and society », *Nature*, vol. 435, n° 7043, p. 814-818, 2005.

Papadopoulos S., Zigkolis C., Kompatsiaris Y., Vakali A., « Cluster-based landmark and event detection on tagged photo collections », *IEEE Multimedia*, vol. 18, n° 1, p. 52-63, 2010.

Perrin C., Dynamique identitaire et partitions sociales : le cas de l'identité'raciale'des noirs en france, PhD thesis, Université de Bourgogne, 2011.

Plantié M., Crampes M., « Survey on Social Community Detection », *Social Media Retrieval*, Springer, p. 65-85, 2013.

Porter M. A., Onnela J.-P., Mucha P. J., « Communities in networks », *Notices of the AMS*, vol. 56, n° 9, p. 1082-1097, 2009.

Quan H., *Online Social Networks & Social Network Services : A Technical Survey*, CRC Press, 2011.

Rignault L., Bonneton L., *Manuel Du Social Media Marketing*, BoD-Books on Demand France, 2012.

Rome J., Haralick R., « Towards a formal concept analysis approach to exploring communities on the world wide web », *Formal Concept Analysis*, vol. 3403, p. 33-48, 2005.

Vakali A., Kafetsios K., « Emotion aware clustering analysis as a tool for Web 2.0 communities detection : Implications for curriculum development », 2012.

Witten I. H., Frank E., *Data Mining : Practical Machine Learning Tools and Techniques*, 2 edn, Morgan Kaufmann, 2005. (Ercument-2011-11-01).

Wu B., Ye Q., Yang S., Wang B., « Group CRM : a new telecom CRM framework from social network perspective », *Proc. of the 1st ACM international workshop on Complex networks meet information & knowledge management*, CNIKM '09, ACM, New York, NY, USA, p. 3-10, 2009.

Yang B., Liu D., Liu J., « Discovering Communities from Social Networks : Methodologies and Applications Handbook of Social Network Technologies and Applications », *in* , B. Furht (ed.), *Handbook of Social Network Technologies and Applications*, Springer US, Boston, MA, chapter 16, p. 331-346, 2010.