Apports des réseaux sociaux dans les SI

Une application à la gestion de la relation client

Ian Basaille, Lylia Abrouk, Nadine Cullot, Éric Leclercq

Laboratoire LE2I - UMR CNRS 6306
Université de Bourgogne
9, Avenue Alain Savary
F-21078 Dijon
Prenom.Nom@u-bourgogne.fr

RÉSUMÉ. *Depuis quelques années, le Web s'est transformé en une plateforme d'échange. Le métier de gestion de relation client doit évoluer, se connecter aux réseaux sociaux et mettre l'entreprise au cœur des échanges. Nous proposons dans cet article une approche de détection de communautés, de clients d'une entreprise, basée sur le comportement explicite et implicite des clients. Pour cela, nous définissons une mesure de similarité, entre un utilisateur et un tag, qui prend en compte la notation et la consultation des ressources ainsi que les contacts des utilisateurs. Nous validons cette approche sur une base exemple.*

ABSTRACT. *In recent years, the Web has evolved into an exchange platform. Customer relationship management must evolve and connect to social networks and place the company at the heart of communication. We propose in this paper an approach to community detection of customers of a company based on explicit and implicit behavior of customers. For this, we define a similarity measure, between a user and a tag, that takes into account the rating and consulting resources and user contacts. We validate this approach against a small database.*

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 révolution, inspirée par l'arrivée du Web 2.0, des réseaux sociaux et de leur popularité croissante. En effet, depuis quelques années, le Web s'est transformé en une plateforme d'échanges 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 les mécanismes d'annotation et de partage de ressources.

Les entreprises commerciales utilisent traditionnellement le Web comme média à des fins marketing à travers des sites dédiés, ou au moyen de publicités incluses dynamiquement sur d'autres sites. 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 ou des prospects est un challenge aux enjeux primordiaux pour la compétitivité d'une entreprise.

Les consommateurs sont amenés à interagir avec les sites marchands non seulement pour connaître les produits ou réaliser des achats, mais aussi pour correspondre avec les entreprises à travers les services après-vente en ligne. Les avatars et les conseillers clients au téléphone permettent de répondre aux attentes et aux questions des consommateurs. Ces derniers deviennent également acteurs pour la marque ou la société en donnant des notes, des avis et commentaires sur les informations mises à leur disposition, voire même en enrichissant ces informations, soit sur les sites dédiés de la marque, soit sur les réseaux sociaux.

Du point de vue de l'entreprise, l'offre d'outils de CRM est importante. Ces outils visent à proposer des environnements riches permettant d'améliorer la gestion des services commerciaux, marketing ou après-vente et de disposer de méthodes d'analyse, notamment statistiques. Ils 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. Le terme *Social CRM* est associé à l'utilisation des médias sociaux dans le cadre de la relation client.

On distingue généralement trois catégories d'outils CRM :

1. Les outils *généralistes*. Ce sont des logiciels « sur étagère ». On peut citer le logiciel *SugarCRM*, qui propose des fonctionnalités classiques des CRM (gestion de la relation commerciale, marketing, service client, outils d'analyse) mais aussi des outils collaboratifs ; et le logiciel *SalesForce* qui offre des fonctionnalités similaires.

2. Les outils *intégrables*. Ce sont des modules logiciels destinés à s'interconnecter avec d'autres applications du système d'information 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 logiciel *Smarter Analytics d'IBM* qui permet l'analyse de données à des fins décisionnelles au sein d'une entreprise.

3. Les outils *génériques*. Ce sont des logiciels développés pour être paramétrables et adaptables aux besoins des sociétés. L'offre peut être modulable et toucher tous les

domaines de la gestion de la relation client. Ces outils peuvent plus facilement être étendus pour évoluer vers du *Social CRM*.

D'une façon générale, la multiplication des canaux de communication et la dispersion des conversations sur le Web à propos d'un produit ou d'une entreprise ont rendu la gestion de la relation client complexe. Les logiciels CRM doivent évoluer et prendre en compte les *clients internautes* qui sont au centre de l'activité de l'entreprise, qui participent à la *e-réputation* d'une entreprise et qui agissent donc directement sur sa croissance ou sa compétitivité. Ils doivent évoluer pour s'intégrer dans les architectures de SI des entreprises connectées aux réseaux sociaux grand public. Le *Social CRM* doit répondre aux besoins de l'entreprise, mais aussi des clients de l'entreprise, et chaque acteur de la relation client au sein de l'entreprise devra disposer d'outils simples et interactifs permettant de faire face aux exigences des entreprises, mais surtout de réactivité attendue par les clients / prospects et de leur besoin d'immédiateté sur les médias numériques. Ainsi, les applications dites *Social CRM* permettront de tirer parti du potentiel des réseaux sociaux en termes de connaissances sur les clients et les consommateurs. Un besoin émerge donc pour créer des plates-formes permettant aux CRM d'être connectés aux réseaux sociaux et de replacer l'entreprise au sein des échanges ayant lieu sur le Web et auxquelles elle ne participe pas encore.

Les outils de détection de communautés sont des éléments essentiels pour répondre aux besoins du *Social CRM*. Les communautés peuvent se trouver aussi bien dans les réseaux sociaux publics que dans le réseau social professionnel d'une entreprise. Les communautés d'utilisateurs existent de manière implicite, leur découverte profite aussi bien à l'entreprise qu'à ses clients. L'amélioration de la communication avec les clients nécessite la prise en compte de leurs intérêts et comportements.

La problématique scientifique associée est très large : elle concerne le traitement de gros volumes de données permettant l'analyse des *traces* des interactions des internautes avec le Web, que ce soit à travers des sites dédiés, des forums, des annotations ou des commentaires pour mieux connaître les clients internautes et améliorer leurs relations avec les entreprises. Elle inclut également le développement de méthodes, d'algorithmes permettant d'identifier des internautes pour les regrouper en communautés et pour interagir de façon plus ciblée avec ces communautés.

Nous proposons dans cet article une approche de détection de communautés de clients basée sur les usages et comportements de l'utilisateur, avec des applications aussi bien à l'intérieur qu'à l'extérieur du système (activités des internautes sur un site dédié ou sur des réseaux sociaux publics).

L'article est organisé de la façon suivante : la section 2 présente des définitions et travaux sur la détection de communautés, la section 3 présente notre approche, qui est composée de trois parties : l'architecture générale d'un système de type *Social CRM*, la construction du profil client et la détection de communautés. Cette approche a été testée sur un jeu de données réduit mais représentatif, les résultats sont présentés dans la section 4. Enfin, nous concluons dans la section 5 et présentons plusieurs perspectives.

2. État de l'art

Depuis les débuts du Web jusqu'à aujourd'hui, la recherche de communautés implicites a fortement évolué. Elle a d'abord concerné l'étude des liens entre documents pour aboutir, depuis quelques années, à l'étude des liens entre individus en fonction de leurs inter-actions (au moyen d'applications Web). (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 et des SI, la recherche d'identité, la propriété des données et la sécurité. De nombreux travaux spécifiques se concentrent sur la notion de communautés (Kumar *et al.*, 1999). (Cohen *et al.*, 2012) présentent un état de l'art des mesures de proximité dans les réseaux sociaux, permettant de quantifier le degré de similarité entre deux utilisateurs. Nous donnons dans un premier temps des définitions liées au contexte de ce travail et nous présentons ensuite plusieurs travaux connexes sur la découverte des communautés et l'utilisation de tags dans le profil utilisateur.

2.1. Les réseaux sociaux et les réseaux sociaux d'entreprise

Un réseau social peut être modélisé sous la forme d'un graphe. Les nœuds sont les éléments du réseau, tels que les utilisateurs ou ressources ; et les arêtes, ou arcs si le réseau est orienté, sont les relations entre ces éléments. Les relations peuvent décrire des liens d'affinité entre les utilisateurs, des similarités thématiques entre des ressources, etc. Le terme réseau social est depuis quelques années associé au Web 2.0 où l'utilisateur est devenu un acteur principal : il peut ajouter, annoter, interagir et diffuser ou partager du contenu. Ce Web "collaboratif" favorise ainsi les interactions sociales en ligne. L'Activity Stream [1] est un format standardisé permettant d'associer des métadonnées aux actions réalisées par un utilisateur sur les différents réseaux sociaux sur lesquels il est inscrit afin de les différencier les unes des autres et de leur donner plus de sens. Il est basé sur le schéma ***acteur verbe objet*** *cible*, par exemple *un utilisateur associe un tag à une ressource* ou *un utilisateur est en contact avec un autre utilisateur*

Les réseaux sociaux sont également apparus comme un outil essentiel au sein du SI des entreprises. Ils sont maintenant utilisés quotidiennement par les salariés ; et les entreprises développent de plus en plus leur propre réseau social afin de partager ou d'élaborer, en interne, du contenu ; ou afin de créer des liens entre les usagers du SI. Il existe plusieurs outils spécifiques qualifiés de réseaux sociaux d'entreprise, comme par exemple **Yammer** [2] qui permet de travailler en réseau avec ses collègues ou **Bluekiwi** [3], une plateforme de collaboration et de dialogue pour les échanges internes et externes de l'entreprise.

1. http://activitystrea.ms/
2. www.yammer.com
3. http://www.bluekiwi-software.com

2.2. Découverte des communautés Web

Une communauté Web est un ensemble de pages Web créées par des personnes ou organisations ayant un intérêt commun sur un sujet précis. La communauté Web est généralement un ensemble dense connexe de pages Web de même contenu thématique (Rome, Haralick, 2005). Dans le Web social, les communautés sont généralement un ensemble de ressources, d'utilisateurs ou de tags (Papadopoulos *et al.*, 2010). (Vakali, Kafetsios, 2012) identifient et définissent trois étapes dans la détection de communautés :

1. définition de mesures pour le calcul des relations entre les utilisateurs afin de détecter de communautés implicites

2. utilisation d'approches algorithmiques pour la détection de réseaux complexes tels les réseaux du Web

3. choix et utilisation d'une mesure pour l'évaluation des communautés

Les premiers travaux sur la construction de communautés plaçait le lien hypertexte comme base de calcul ((Kleinberg, 1999) (Imafuji, Kitsuregawa, 2002)) en structurant généralement des communautés thématiques. Aujourd'hui, le concept de communauté concerne principalement des utilisateurs aux rôles multiples : consommateur, producteur de ressources, émetteur de messages, évaluateur etc. On remplace ainsi les liens par la notion de profil.

2.3. Utilisation de tags pour le profil utilisateur

Le profil utilisateur est constitué d'un ensemble d'informations concernant un utilisateur, comme son nom, son âge, sa ville. Il contient aussi généralement des informations sur ses centres d'intérêts et les notes qu'il donne à des ressources (Golbeck, 2009). Ces intérêts peuvent être renseignés directement par l'utilisateur de manière explicite, ou de manière implicite en analysant son comportement. Le profil peut être représenté par un ensemble de mots clés (tags). (Cattuto *et al.*, 2008) proposent un algorithme de recommandation basé sur les tags des utilisateurs. 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 (*cold start*), les nouveaux utilisateurs recevant d'abord des recommandations peu pertinentes. Une solution hybride (basée sur l'aspect collaboratif, mais aussi sur le contenu) proposée par (Yoshii *et al.*, 2006) utilise un modèle probabiliste pour intégrer les votes des utilisateurs et le contenu des données en utilisant un réseau bayésien pour améliorer les méthodes classiques.

Le profil dépend des usages et comportements de l'utilisateur. Prendre en compte le profil explicite ou les thématiques n'est pas suffisant, 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.

2.4. Intérêts et limites

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 ses 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 des tags et de leur variabilité sémantique. Par exemple, plusieurs orthographes d'un même tag peuvent être disponibles ce qui rend la recherche et l'utilisation des tags difficiles. Deux utilisateurs peuvent aussi utiliser deux mots différents pour définir le même concept.

Les réseaux sociaux d'entreprise permettent de favoriser la collaboration au sein des entreprises, mais il leur manque encore des outils de découverte de communautés pouvant améliorer leur efficacité. Ils sont aussi, par définition, restreints au périmètre interne de l'entreprise.

L'approche que nous présentons part des hypothèses suivantes : nous travaillons avec le réseau social interne à un entreprise, et nous les étendons en incluant les clients de l'entreprise. A l'intérieur de cet espace de travail, nous construisons un profil, toujours avec des tags mais issus non plus de l'utilisateur mais d'un vocabulaire défini et contrôlé. Nous analysons aussi les actions des utilisateurs afin d'améliorer le profil.

3. Modèle du profil et construction de communautés

La compréhension et la bonne communication entre une entreprise et ses clients sont importantes pour la propagation de l'information, la réputation de l'entreprise et les ressources qu'elle peut offrir à ses clients.

Les réseaux sociaux d'entreprise sont centrés sur les membres de l'entreprise. Ils sont généralement utilisés pour la communication entre ces membres et pour le partage d'informations. 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 composée de trois parties : 1) une modélisation de l'architecture générale du système ; 2) la construction du profil utilisateur et 3) l'utilisation d'un algorithme pour la détection de communautés d'utilisateurs.

3.1. Architecture générale

Les utilisateurs (qui sont les clients de l'entreprise) interagissent avec le système d'information de l'entreprise mais aussi avec les réseaux sociaux grand public. 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. La figure 1 présente l'architecture générale d'un système utilisant un *Social CRM*, elle est composée de trois parties distinctes :

1. *Le système d'information du site Web* dédié de l'entreprise contient des ressources et un thésaurus. Les ressources sont taguées avec les termes du thésaurus. Le comportement de l'utilisateur est enregistré via l'historique des ressources qu'il a consultées et les notes qu'il attribue aux ressources. Les utilisateurs peuvent déclarer des liens entre eux, formant ainsi le réseau social interne du site Web.

2. *Le CRM de l'entreprise* contient les informations relatives aux échanges entre un client identifié ou un prospect et le service client de la marque, comme le nom, le prénom, la date de naissance, l'adresse, la liste des produits achetés et des problèmes que le client a pu avoir.

3. *Les réseaux sociaux publics* tels que Facebook ou Twitter permettront d'affiner les profils des utilisateurs avec des informations qui ne sont pas disponibles au sein de l'entreprise comme par exemple une liste de ressources qui intéresse un utilisateur et qu'il n'a pas encore visitées à l'intérieur du système d'information de l'entreprise.

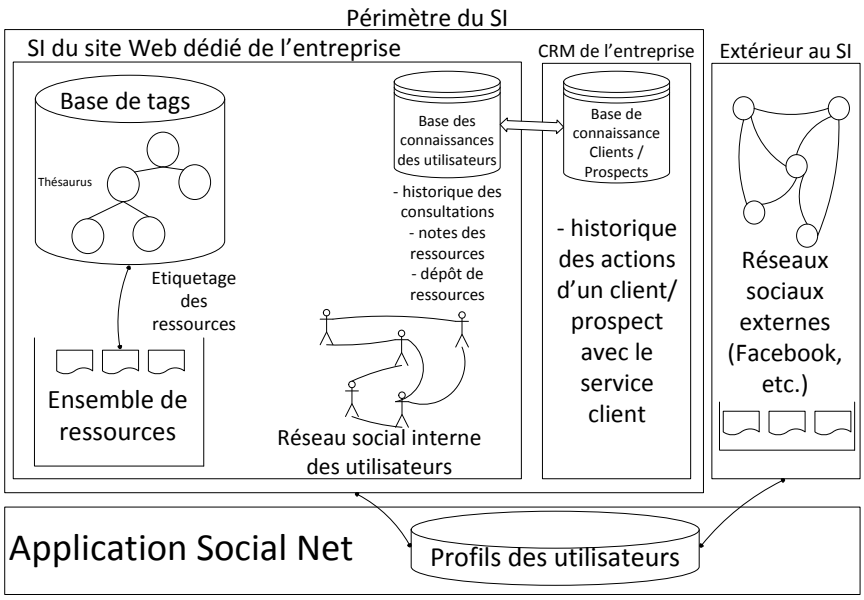

Figure 1. Architecture générale du système d'information

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 soit de manière explicite soit de manière 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 le dépôt et la consultation de cette dernière 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 R_{int} dédié à l'entreprise ou à l'extérieur du système R_{ext}. Nous supposons que les utilisateurs émettent une note $n \in \mathbb{N}$ sur 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 $r_j \in R_{int}$, $M(u_i, r_j) = n_{ij}$. Cette matrice est mise à jour dynamiquement lorsque de nouveaux utilisateurs, de nouvelles ressources apparaissent sur le site.

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 celles dont les tags 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 $t_k \in T$:

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

On suppose que le système permet aux utilisateurs de définir une liste de contacts, 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 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. Nous calculons le degré d'appartenance da_{ij} d'un utilisateur u_i à un tag t_j :

$$da_{ij} = \frac{|R(u_i, t_j)|}{|R|} \times m_j \quad (3)$$

avec

$$m_j = \frac{\sum n_{ij}}{n_{max} \times |n_{ij}|} \tag{4}$$

– $R(u_i, t_j)$ 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 $R(u_i, t_j)$

– m_j la moyenne des notes données par l'utilisateur u_i aux ressources $R(u_i, t_j)$ divisée par n_{max} afin d'obtenir une valeur comprise entre 0 et 1

– n_{max} est la note maximale donnée aux ressources par les utilisateurs

Comme l'objectif de notre approche est de prendre en compte le comportement de l'utilisateur dans le système, nous affinons l'expression de da_{ij} avec le comportement de u_i, soit avec son intérêt pour une ressource par le dépôt ou la consultation de cette dernière, soit avec son réseau de contacts en prenant en compte les ressources notées par ses contacts. Le degré d'appartenance de u_i à un tag t_j est modifié si ce dernier est associé à une ressource consultée dans l'ensemble des ressources R par u_i, et en fonction de son degré d'appartenance pour les contacts de u_i. Pour cela nous définissons, sur une session s d'un utilisateur, un degré d'appartenance de session d'_{ijs} en fonction des documents consultés $R_{consult}$ annotés avec le tag t_j.

$$d'_{ijs} = \frac{|R_{consult}(u_i, t_j)|}{|R_{consult}|} \tag{5}$$

Pour une nouvelle session de recherche, le degré d'appartenance de session est la moyenne du degré de la session courante et de la session précédente. Cela est motivé par le fait qu'un utilisateur change d'intérêts pour les documents et thématiques avec le temps.

$$d'_{ij} = \frac{d'_{ijs} + d'_{ijs-1}}{2} \tag{6}$$

– d'_{ijs} est le degré d'appartenance en fonction des documents consultés annotés avec le tag t_j pour la session courante s de u_i

– d'_{ijs-1}est le degré d'appartenance en fonction des documents consultés annotés avec le tag t_j pour la session précédente $s-1$ de u_i

A partir des trois paramètres pris en compte dans notre approche (notes, consultations et contact), le degré d'appartenance d_{ij} d'un utilisateur u_i à un tag t_j est défini comme suit :

$$d_{ij} = \alpha \times da_{ij} + \beta \times \frac{\sum_{k=1}^{m} da_{kj}}{m} + \gamma \times d'_{ij} \tag{7}$$

– α et β et γ sont des pondérations à paramétrer avec $\alpha + \beta + \gamma = 1$
– $A(u_i) \subseteq U$ est l'ensemble des contacts de l'utilisateur u_i et $m = |A(u_i)|$
– da_{kj} est le degré d'appartenance d'un utilisateur u_k au tag t_j, avec $u_k \in A(u_i)$

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

3.3. Détection de communautés

Une fois le profil de l'utilisateur construit, on calcule les communautés d'utilisateurs. Afin de constituer les groupes d'utilisateurs, nous utilisons l'algorithme K-means (k-moyennes). La classification K-means est une des techniques de classification non supervisées les plus utilisées. Nous l'avons choisie car elle converge rapidement au bout de quelques itérations. Cela permet de simuler plusieurs cas, avec un nombre de classes différent à chaque fois, et de laisser le choix à l'utilisateur (par exemple le community manager) d'interpréter les résultats en fonction du contexte.

Étant donné un entier K, l'algorithme K-means vise à départager l'ensemble des données à classifier en K classes les plus homogènes possible. Les utilisateurs (objets) X_j $(1 \leq j \leq N)$ sont représentés sous forme de vecteurs. L'algorithme se déroule de la façon suivante :

1. choisir K objets au hasard parmi les objets de la collection. Soient $(X_1, ..., X_k)$ les objets obtenus. $(X_1, ..., X_k)$ sont les représentants de K classes $(C_1, ..., C_k)$

2. affecter chaque objet X_j de la collection à l'une des classes en fonction du représentant le plus proche :

$$argmin_{k,1 \leq k \leq K} d(X_j, X_k) \tag{8}$$

3. calculer de nouveaux représentants pour les classes. Ils correspondent à la moyenne des objets de la classe :

$$\forall k, 1 \leq k \leq K, X_k = \frac{1}{|C_k|} \sum_{j, x_j \in C_k} x_j \tag{9}$$

4. retourner en 2 tant que la différence entre les anciens et les nouveaux représentants est supérieure à un seuil fixé

4. Expérimentations

Nous avons testé notre approche sur un jeu d'essai restreint. Celui-ci prend la forme d'une base de connaissances, permettant 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. Ce jeu d'essai contient environ 20 utilisateurs et 50 ressources.

Les utilisateurs de cette base sont des chercheurs, des industriels et des experts du domaine de l'agro-alimentaire. Cette base 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. Nous nous intéressons uniquement aux actions de poster, de consulter et de noter une ressource. Ces ressources sont des articles de recherche, des synthèses d'études, des comptes-rendus de réunion, au format PDF.

La société gérant cette base a développé une plateforme qui s'apparente à un CRM dans le sens où les administrateurs de la base ont pour client des industriels, des laboratoires de recherche et des experts.

L'expérimentation se compose de deux étapes :

1. la construction du profil utilisateur avec le calcul de ses degrés d'appartenance
2. la construction des communautés d'utilisateurs

Les ressources sont évaluées par une note 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, représenté dans la figure 2.

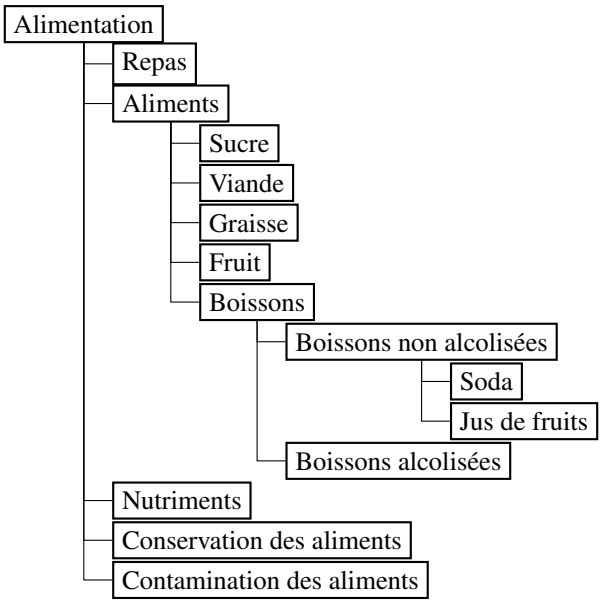

Figure 2. Extrait de thésaurus dans le domaine alimentaire

4.1. Construction du profil utilisateur

Nous avons construit la matrice M avec l'ensemble des utilisateurs U et les ressources R avec les notes des utilisateurs. Nous avons également la liste des ressources avec les tags associés, le tableau 2 illustre quelques exemples de cette annotation et le tableau 4 quelques notes données par les utilisateurs .

Tableau 1.

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 2. Exemple d'annotation de ressources

Tableau 3.

	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 4. Exemple de notes de ressources

Enfin, nous avons construit la matrice M_d et calculé les degrés d'appartenance des utilisateurs aux différents tags. Nous présentons dans le tableau 6 un extrait de ces résultats pour quelques utilisateurs et quelques tags.

Tableau 5.

	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 6. Profil utilisateur explicite

Nous affinons notre approche par le comportement implicite des utilisateurs : nous prenons en compte les consultations des ressources et le réseau de contacts des utili-

sateurs. Le tableau 8 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 plus les notes données par les utilisateurs, et une deuxième donnant un peu plus de poids au comportement de l'utilisateur. Nous avons donc fixé α = 0.6 et β = 0.3 et γ = 0.1 pour la première expérimentation, et α = 0.5 et β = 0.3 et γ = 0.2 pour la deuxième. Le tableau 8 donne un exemple de profil pour la première expérimentation.

Nous pouvons déjà remarquer par exemple que le d_{ij} de l'utilisateur u_1 pour le tag $Repas$ a augmenté. Ceci s'explique par le lien entre les utilisateurs u_1 et u_{10} et le degré d'appartenance de l'utilisateur u_{10} au tag $Repas$. On note aussi que l'utilisateur u_1 consulte régulièrement des documents annotés par ce tag.

Tableau 7.

	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 8. Profil utilisateur explicite et implicite

4.2. Détection de communautés

Pour la construction de communautés d'utilisateurs, nous avons utilisé le logiciel libre de data-mining WEKA (Witten, Frank, 2005) qui nous a donné les résultats présentés dans le tableau 10. Les représentants des communautés sont en italique.

Après avoir étudié le comportement des utilisateurs et remarqué que certains se spécialisaient dans quelques branches du thésaurus, que d'autres le parcouraient de façon plus globale, nous avons choisi de prendre 5 communautés d'utilisateurs, correspondant aux 5 thèmes du thésaurus présenté en figure 2 (Repas, Aliments, Nutriments, Conservation des aliments et Contamination des aliments décrits dans la figure 2).

En observant le comportement des utilisateurs, on remarque que u_1, u_3 et u_{11} partagent les mêmes intérêts, principalement le thème *Aliments*. Il semble donc logique que ces trois utilisateurs soient regroupés dans la même communauté. En construisant les communautés avec 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_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, son degré d'appartenance à ce tag augmente en affinant les degrés.

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 le cas avec un profil prenant en compte uniquement les notes données par un utilisateur à des ressources.

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.

Les communautés affinées illustrées dans le tableau 10 ont pour différence une pondération plus forte pour l'historique de consultation pour la colonne *Affinées 2* et la pondération plus forte pour les ressources notées pour la colonne *Affinées*. Cela influe sur les degrés et donc sur les communautés qui en résultent. Par exemple, 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 utilisateurs 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.

Tableau 9.

Communautés	Pas de pondération	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)
Communauté 1	*U1*, U4, U12	*U1*, **U3**, U4, **U11**, U12, U13, U20	*U1*, U3, U4, **U8**, U11, U12, U13, **U17**, **U18**, U20
Communauté 2	*U2*, U6, U7, U9, U10, U13, U15, U18	*U2*, U6, U7, U8, U9, U10, U15, U18	*U2*, U6, U7, U9, U10, U15
Communauté 3	*U3*, U14, U20	*U5*, U16, U17	*U5*, U16
Communauté 4	*U5*, U8, U11, U16, U17	*U14*	*U14*
Communauté 5	*U19*	*U19*	*U19*

Tableau 10. Communautés non affinées et affinées

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

5. Conclusion et perspectives

Dans cet article, nous avons présenté une méthode de détection de communautés basée sur les usages, les comportements et le réseau social des utilisateurs. L'utilisation des données comportementales et sociales permet d'affiner les profils des utilisateurs et de créer des communautés basées non seulement sur leurs intérêts, mais aussi sur les thématiques qui les intéressent à un instant donné et sur les intérêts de leurs contacts. Cela permet de former des communautés dynamiques évoluant en fonction des comportements des utilisateurs du système et offre des possibilités en terme de recommandation d'utilisateurs et de ressources. Nous avons constaté que les communautés obtenues reflètent ce qui intéresse les utilisateurs à un instant donné. Notre approche offre aussi la possibilité de découvrir des communautés en fonction des paramètres de pondération des profils affinés, permettant ainsi l'évolution des communautés en fonction des critères que l'on souhaite voir renforcés.

Dans le futur, nous allons développer un système de recommandation de ressources à consulter, de thématiques qui peuvent être intéressantes et d'utilisateurs dont les intérêts sont proches afin de tirer parti des communautés et du réseau social d'un utilisateur. Nous allons aussi poursuivre les expérimentations afin d'affiner les valeurs de pondération des degrés pour les différents types de communautés souhaités et travailler sur des bases d'utilisateurs, de ressources et de tags plus grandes. Les utilisateurs seuls dans une communauté devront aussi être pris en compte pour leur permettre de recevoir des recommandations plus pertinentes. Nous allons aussi nous orienter vers l'utilisation d'algorithmes de détection de communautés ne nécessitant pas la spécification du nombre de classes a priori. Nous allons aussi étendre le domaine de travail aux réseaux sociaux sur le Web (Facebook, Twitter, etc.) et sur des forums liés au contexte de travail afin de regrouper plus d'informations sur les utilisateurs que nous pourrons identifier sur ces réseaux. Pour cela, nous allons étudier le format Activity Stream.

Remerciements

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

Bibliographie

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

Cohen S., Kimelfeld B., Koutrika G. (2012). A survey on proximity measures for social networks. *Search Computing*, vol. 7538, p. 191–206.

Dey A. K., Abowd G. D., Salber D. (2001, décembre). 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. http://dx.doi.org/10.1207/S15327051HCI16234_02

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

Imafuji N., Kitsuregawa M. (2002). Effects of maximum flow algorithm on identifying web community. In *Proceedings of the 4th international workshop on web information and data management*, p. 43–48.

Kleinberg J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, vol. 46, n° 5, p. 604–632.

Kumar R., Raghavan P., Rajagopalan S., Tomkins A. (1999). Trawling the web for emerging cyber-communities. *Computer networks*, vol. 31, n° 11, p. 1481–1493.

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

Quan H. (2011). *Online social networks & social network services: A technical survey*. CRC Press.

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

Vakali A., Kafetsios K. (2012). Emotion aware clustering analysis as a tool for web 2.0 communities detection: Implications for curriculum development.

Witten I. H., Frank E. (2005). *Data mining: Practical machine learning tools and techniques* (2e éd.). Morgan Kaufmann. /bib/witten/Witten2005/DataMiningPracticalMachineLearningToolsandTechniques2ded-MorganKaufmann.pdf ((Ercument-2011-11-01))

Yoshii K., Goto M., Komatani K., Ogata T., Okuno H. (2006). Hybrid collaborative and content-based music recommendation using probabilistic model with latent user preferences. In *Proceedings of the international conference on music information retrieval.*