Web2.0 : opportunités pour l'entreprise

Gestion de données dans les réseaux sociaux

par Tatel ABDESSALEM et Pierre SENELLART

Le Web devient de plus en plus social : des individus interagissent entre eux ou autour de contenus qu'ils partagent, commentent ou produisent. La structure de ces réseaux sociaux peut être utilisée pour découvrir des informations intéressantes sur ces individus

La dimension sociale est un des aspects les plus intéressants des nouvelles formes de contenus du Web 2.0 : un grand nombre de sites Web à succès intègrent un mécanisme social permettant une interaction entre utilisateurs, la production et le partage de contenus par ces utilisateurs, la possibilité de découvrir de nouveaux utilisateurs partageant les mêmes intérêts, etc. Ainsi, parmi les vingt sites attirant le plus de trafic en France au mois d'octobre 2009 d'après Alexa, huit d'entre eux sont construits autour de la notion de réseau social (par ordre de trafic décroissant, Facebook, YouTube, SkyRock, Wikipédia, Dailymotion, eBay, Comment ça marche, Blogger) et la plupart des autres intègrent une forme d'interaction sociale, par exemple au travers de forums de discussion.

Les sites de réseaux sociaux existant sur le Web sont de natures très variées. Certains sont plutôt orientés contenu, d'autres plutôt orientés utilisateurs. Dans la première catégorie, on retrouve des sites permettant de cataloguer et de donner des avis sur des produits culturels et autres (par exemple, LibraryThing pour les livres, *Del.icio.us* pour les sites Web, ou Yahoo! Movies pour les films). D'autres sites permettent le partage de documents multimédias (Flickr, YouTube, Dailymotion…), tandis que d'autres encore comme Wikipédia ou everything2 se concentrent sur l'édition. Dans cette première catégorie, on peut aussi mentionner les sites de vente en ligne tels eBay et PriceMinister et les sites de discussion autour d'un sujet donné comme Yahoo! Answers. La caractéristique de tous ces sites est que le contenu et l'interaction entre contenu et utilisateurs sont plus importants que l'interaction entre utilisateurs. Ainsi certains de ces sites (Wikipédia ou eBay, par exemple) n'explicitent pas la notion d'amitié ou de connexion entre utilisateurs, même si des liens implicites sont formés quand des utilisateurs de Wikipédia éditent les mêmes pages ou des utilisateurs d'eBay enchérissent sur les mêmes produits. À l'inverse, des liens explicites entre utilisateurs sont à la base des sites orientés utilisateurs, que ce soient des sites de réseaux sociaux purs comme Facebook ou LinkedIn, des communautés de blogs comme SkyBlog, des micro-blogs comme Twitter, ou des sites de rencontres comme Meetic. Pour chacun de ces sites, la structure du réseau social

(qui connaît qui, qui est ami avec qui) est aussi importante que le contenu lui-même (profil, messages). Enfin, le Web lui-même pourrait être vu comme un immense réseau social de pages, si on considère les hyperliens entre pages comme des liens de réseaux sociaux.

Cette omniprésence des réseaux sociaux dans les données du Web pose de nouveaux problèmes de gestion et d'extraction de ces données. Nous nous attachons ici à deux problèmes en particulier : comment extraire de l'information utile de la structure de ces réseaux, en particulier comment découvrir des communautés d'utilisateurs ? Comment évaluer la confiance envers un membre du réseau et, indirectement, décider du crédit à accorder au contenu qu'il met à la disposition des autres utilisateurs ?

Fouille de réseaux sociaux

Une manière naturelle de concevoir un réseau social est de le modéliser par un graphe au sens mathématique du terme : un ensemble de nœuds, représentant les entités du réseau (utilisateurs, groupes d'utilisateurs, contenus, etc.) reliés entre eux par des arêtes, ou liens, qui représentent les relations entre ces entités (appartenance à un groupe, possession ou intérêt pour le contenu, amitié entre utilisateurs, etc.). On peut donc utiliser les nombreux outils de la théorie des graphes pour traiter et interpréter ces données. On peut par exemple parcourir les composantes connexes du graphe du Web pour savoir quelle page peut être atteinte à partir de quelle autre en suivant des liens d'une page à l'autre.

Mais les graphes constituant les réseaux sociaux ne sont pas des graphes arbitraires. L'expérience des six degrés de séparation a popularisé une caractéristique partagée par la plupart d'entre eux : ce sont des graphes **petit-monde**. Le sociologue Stanley Milgram, dans les années 1960, a demandé à un ensemble de personnes aux États-Unis d'envoyer des courriers à des personnes qu'ils ne connaissaient pas, en autorisant ces personnes à transmettre uniquement ces courriers à des connaissances susceptibles de connaître plus facilement les destinataires finaux. Les destinataires intermédiaires pouvaient à nouveau retransmettre les courriers de proche en proche, jusqu'à atteindre les destinataires finaux. La moyenne du nombre d'intermédiaires quand le courrier arrivait à destination (ça n'était pas toujours le cas !) était d'environ cinq, formant une chaîne de six degrés de séparation entre deux individus. Ce chiffre six n'est pas à prendre au pied de la lettre, mais ce qui est important est que la distance typique entre deux individus dans un réseau social est en général nettement inférieure à ce chiffre.

Cette **faible distance typique** est une des quatre caractéristiques que l'on retrouve dans la plupart des réseaux sociaux, et en fait dans la plupart des grands graphes du monde réel. Les trois autres caractéristiques sont les suivantes :

- **Graphes creux.** Ces graphes ont très peu d'arêtes par rapport au nombre maximal possible : un individu donné n'est ami qu'avec une toute petite proportion des individus du monde.

- **Haute transitivité.** Quand un nœud A est relié à un nœud B et un nœud C, alors il est assez probable (mais pas du tout certain) que B soit relié à C : les amis de mes amis sont plus probablement mes amis qu'un individu arbitraire. Ceci permet concrètement à des réseaux sociaux comme Facebook, LinkedIn ou Viadeo de suggérer des contacts potentiels parmi les amis de ses amis.

- **Distribution des degrés en loi en puissance.** La distribution du nombre de nœuds connectés à un nœud donné (le degré) suit une loi de probabilité $P(k)=k^{-b}$ où k est le degré et b un nombre réel, souvent compris entre 2 et 3.

Ces caractéristiques sont importantes pour bien comprendre les particularités des réseaux sociaux et le type d'opérations, d'algorithmes, que l'on peut exécuter efficacement dans ces réseaux. Des chercheurs s'attachent à trouver des modèles mathématiques qui expliquent ces particularités et permettent de produire automatiquement des graphes ressemblant aux graphes du monde réel.

Une fois un réseau social modélisé par un graphe, et les caractéristiques de ce graphe comprises, on peut chercher à appliquer des techniques de fouille de graphe pour extraire de l'information intéressante de ce graphe : déterminer les entités ou individus les plus « importants » dans le graphe, trouver des individus similaires à un individu donné, ou encore diviser le réseau en communautés homogènes d'utilisateurs. Attardons-nous sur ce dernier problème, qui peut par exemple être utilisé dans un réseau social pour identifier des groupes d'utilisateurs ayant des intérêts communs. De nombreux algorithmes de partition de graphe ont été proposés pour identifier des communautés d'utilisateurs, adaptés à divers contextes. Considérons ainsi l'algorithme suivant, proposé par Newman et Girvan en 2004. Étant donné un graphe, le but est de le séparer en communautés de plus en plus petites et de plus en plus homogènes. On va pour cela déterminer les arêtes du graphe qui sont les plus « au milieu », en comptant le nombre de chemins les plus courts entre deux nœuds quelconques du graphe qui passent par cette arête. L'arête en question est alors ôtée du graphe, et on recommence l'opération en recalculant les chemins les plus courts entre nœuds du graphe. On s'arrête quand le graphe a été découpé en un nombre suffisant de composantes. Cet algorithme, pas particulièrement efficace (il nécessite un temps d'exécution cubique en nombre de nœuds), est toutefois intéressant par sa simplicité. D'autres méthodes, plus complexes, permettent des traitements plus rapides et sont applicables à des réseaux sociaux de plusieurs millions d'individus.

Confiance dans les réseaux sociaux

Les réseaux sociaux qui se développent sur le Web regroupent un nombre très large d'utilisateurs, sont souvent difficiles à contrôler et dans ces réseaux n'importe qui peut apporter une contribution. Se pose alors le problème de la fiabilité des informations partagées au sein d'un réseau social : quel crédit accorder à chaque contribution ? Une solution à ce problème pourrait venir des réseaux de confiance (*Web of Trust*). Dans un réseau de confiance

Web2.0 : opportunités pour l'entreprise

chaque utilisateur a la possibilité de maintenir des données sur la confiance qu'il accorde à d'autres membres du réseau. Ainsi, il peut aider d'autres utilisateurs à décider de la confiance qu'ils peuvent accorder à un utilisateur avec lequel ils n'ont jamais interagi, et indirectement décider du crédit qu'ils peuvent accorder aux contributions de cet utilisateur.

Le mécanisme qui permet d'inférer de nouveaux liens de confiance à partir des liens de confiance explicitement exprimés par les utilisateurs du réseau s'appelle la *propagation de la confiance*. Cette propagation peut s'effectuer selon des règles de transition qui ont fait l'objet de plusieurs études, notamment par Ramanathan V. Guha et al. en 2004, et qui peuvent se résumer comme suit :

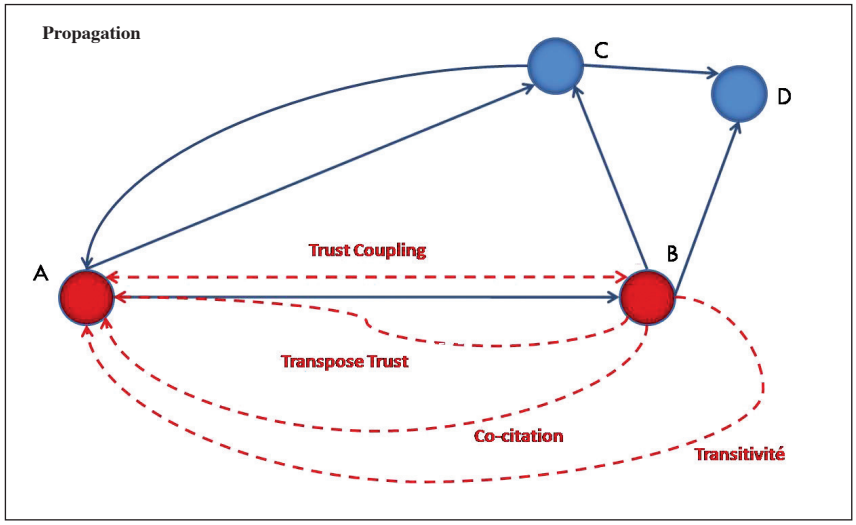
Considérons l'exemple de réseau de confiance donné ci-dessous. Les nœuds (A, B, C et D) représentent des utilisateurs et les arcs représentent des liens de confiances. Les arcs pleins représentent des liens de confiance explicite (AàC veut dire que A fait confiance à C) et les arcs en pointillés représentent des liens de confiance inférés.

La **Transitivité** est le mode de propagation le plus intuitif. Si l'utilisateur B fait confiance à un utilisateur C, et C à son tour fait confiance à A, alors B pourrait en déduire une valeur de confiance envers A.

La **Co-citation** repose sur l'idée que si deux utilisateurs A et C ont des jugements similaires sur la fiabilité d'une tierce personne D, alors ils pourraient également être d'accord sur la fiabilité d'autres personnes. Autrement dit, si B et C font confiance à D et que C fait confiance à A, alors B pourrait faire confiance à A également.

Le ***Trust Coupling*** repose sur le fait que si deux utilisateurs font confiance aux mêmes personnes, alors ils peuvent en déduire un lien de confiance réciproque entre eux. Dans notre exemple, si A et B font confiance à C, alors A pourrait avoir confiance en B et inversement.

Transpose Trust exprime un degré de réciprocité dans les relations de confiance.



Si un utilisateur A fait confiance à B, alors B pourrait développer un sentiment de confiance envers A.

Dans les réseaux sociaux, la confiance est liée à une utilité : Paul a confiance en Pierre pour réparer sa voiture, et peut ne pas lui faire confiance pour d'autres utilités. La propagation de la confiance n'a donc de sens que dans des réseaux de confiance liée à une utilité précise.

Dans le cas de la transitivité, il faut tenir compte de la confiance dans les recommandations des utilisateurs intermédiaires : le fait que Paul fasse confiance à Pierre pour réparer sa voiture et que Pierre à son tour fasse confiance à Jean pour la même utilité, ne suffit pas à décider Paul à faire confiance en Jean comme réparateur de voiture. Paul doit avoir confiance dans les capacités de Pierre à juger de la fiabilité des autres utilisateurs. Cette confiance est dite *confiance de recommandation* et se distingue de la *confiance fonctionnelle* utilisée habituellement. La propagation par transitivité s'effectue sur les liens de confiance de recommandation et ne tient compte de la confiance fonctionnelle que lors de la dernière étape de la propagation : si C fait confiance à A et que B a confiance en les recommandations de C, alors B pourrait aussi faire confiance à A.

La propagation de la confiance permet donc d'inférer de nouveaux liens de confiance entre les utilisateurs d'un réseau social, en s'appuyant sur les confiances exprimées et en appliquant les règles de transitions mentionnées ci-dessus. Ceci doit se faire en tenant compte de la nature subjective de la confiance, qui fait que le poids accordé à chaque type de transition peut varier d'un réseau à l'autre et d'un utilisateur à un autre.

Conclusion

Le succès et la popularité des sites de réseaux sociaux soulèvent de nouveaux défis auxquels la communauté scientifique tente de répondre. Nous avons décrit dans cet article deux problématiques particulières, la fouille dans les réseaux sociaux et la gestion de la confiance, sur lesquelles nous menons actuellement des recherches à Télécom ParisTech. ■

Talel ABDESSALEM *et* ***Pierre SENELLART*** *sont maîtres de conférences en informatique à Télécom ParisTech. Talel a obtenu son doctorat de l'Université Paris-Dauphine en 1997 et Pierre de l'Université Paris-Sud en 2007. Au sein de l'équipe DBWeb (http://dbweb.enst.fr/), ils s'intéressent aux aspects théoriques et appliqués des systèmes d'information modernes, en particulier sur le Web et dans des contextes collaboratifs comme les réseaux pair-à-pair.*