Dieudonné Tchuente
Nadine Baptiste-Jessel
Marie-Francoise Canut
Institut de Recherche en Informatique de Toulouse
UMR 5505, équipe SIG/D2S2

ACCÈS À L'INFORMATION DANS LES RÉSEAUX SOCIONUMÉRIQUES

L'analyse des réseaux sociaux est menée dans les sciences sociales depuis les années 1930 (Breslin et Decker, 2007). Cette analyse vise, d'une part, à identifier les structures sociales distinctes dans les réseaux et, d'autre part, à expliquer le comportement des individus au sein de ces structures sociales, au moyen d'approches issues de la sociométrie, avec des outils tels que les matrices ou les graphes. Sous un autre angle, les méthodes ethnographiques (monographies, questionnaires, entretiens individuels ou collectifs, entretiens fermés ou ouverts, *focus group*...) sont également utilisés pour la collecte de données et l'analyse de réseaux sociaux. Cependant, les impacts de l'usage de ces dernières méthodes au niveau des analyses peuvent être limitatifs à plusieurs niveaux, tels que le manque de significativité des résultats (faible taille d'échantillons) (Stutzman, 2006), le biais dans un recueil trop dirigé de données dû au fait que les hypothèses sont préalables aux analyses (il est impossible de trouver des résultats complètement inattendus), le biais dans la qualité de l'information fournie par les utilisateurs (l'utilisateur n'est pas forcément sincère dans ses réponses), etc.[1].

L'avènement du Web social ou Web 2.0 a énormément favorisé le développement des réseaux socionumériques (en 2010, près de trois quarts des internautes en Europe consultent ces nouvelles plateformes quotidiennement[2]). Très visités et comportant des applications diversifiées (*mails*, *chats*, photos, *tags*, groupes, événements, pages…), les réseaux socionumériques sont devenus de véritables systèmes d'exploitation. Des masses de données, riches par leur diversité, et importantes par leur quantité, sont désormais disponibles sur la toile. Plusieurs nouvelles méthodes permettent de collecter automatiquement ces données et proposent des alternatives à l'ethnographie. Cependant, les méthodes de collecte automatique de données issues des utilisateurs dans les réseaux socionumériques peuvent accéder de manière transparente à certaines d'entre elles. Ceci

est d'autant plus crucial lorsque plusieurs travaux montrent que les internautes ne sont pas conscients des nombreuses possibilités d'atteinte à leur vie privée sur les réseaux socionumériques (Dwyer *et al.*, 2008 ; Stutzman, 2006 ; Nagle et Singh, 2009).

Dans cet article, nous présentons sommairement les principales méthodes actuelles de collectes de données utilisateurs sur les réseaux socionumériques. Nous insistons particulièrement (avec une expérimentation à l'appui) sur les méthodes basées sur les *Application Programming Interface* (API) des réseaux socionumériques. En effet, on sait par exemple que *Facebook* dispose aujourd'hui de plus d'un million de développeurs (des tiers) repartis dans plus de 180 pays dans le monde. Aujourd'hui *Facebook* compte plus de 550 000 applications tierces[3] développées, et plus de 70 % des utilisateurs interagissent avec ces applications.

Méthodes d'accès à l'information dans les réseaux socionumériques

Nous regroupons les méthodes d'accès à l'information dans les réseaux socionumériques en cinq principales catégories : les méthodes ethnographiques, les méthodes de *Web mining*, les méthodes de fils de discussion, les méthodes du Web sémantique, et les méthodes basées sur les API des réseaux socionumériques. Dans les paragraphes qui suivent, nous présentons sommairement chaque méthode, leurs avantages et leurs inconvénients.

Les méthodes ethnographiques sont issues des sciences humaines et sociales et ne sont pas spécifiques au Web. Elles s'appuient sur des techniques telles que l'observation des activités, les entretiens (individuels, collectifs, ouverts, fermés…), les monographies… Ces méthodes ont ainsi été parmi les premières à être utilisées pour l'analyse des réseaux socionumériques (Boyd, 2007 ; Stutzman, 2006 ; Dwyer *et al.*, 2008 ; Coutant et Stenger, 2010). Ces techniques privilégient le contact humain avec les utilisateurs qui donne une dimension très réaliste des analyses qui sont réalisées par la suite. Toutefois, elles peuvent avoir des limites telles que celles énoncées dans l'introduction de cet article (faibles échantillons, recueil trop dirigé de l'information, biais dans les informations recueillies…).

Les méthodes de *Web mining* appliquées aux réseaux sociaux consistent à analyser les contenues des pages Web afin d'identifier divers types de relations (cooccurrences de termes, citation de liens hypertextes, citations de co-auteurs, etc.) (Mika, 2005 ; Matsuo *et al.*, 2006 ; Jin *et al.*, 2007). Dans le cas très précis des réseaux socionumériques, Alim *et al.* (2009) réalisent un recueil de données sur des profils publics[4] des utilisateurs *via* la fouille des fichiers HTML des profils utilisateurs. Si cette méthode d'extraction fonctionne sur certains sites de réseaux socionumériques dont les profils publics des utilisateurs contiennent assez d'information (*MySpace* par exemple), la plupart des sites de réseaux socionumériques donnent l'accès à très peu d'information (très souvent, seuls le nom, prénom et éventuellement la liste d'amis) dans le profil public de leurs utilisateurs (*Facebook*, *FriendSter*...).

Les méthodes de fils de discussion permettent d'extraire et d'analyser les contenus publiés par des utilisateurs sur des *chats*, forums, ou groupes de discussion (Reffay et Lancieri, 2006 ; Sidir *et al.*, 2006 ; Dimitracopoulou et Bruillard, 2006). Ces techniques disposent d'assez d'outils pour extraire des contenus très pertinents et réalistes à partir des échanges des utilisateurs. Cependant, elles sont limitées à des environnements fermés dont l'accès est généralement restreint exclusivement aux propriétaires des plateformes. Dans le cas des réseaux socionumériques, en général, l'accès aux données de ce type d'applications n'est pas possible pour des développeurs tiers.

Les méthodes du Web sémantique[5] consistent à représenter les ressources disponibles sur le Web, de telle sorte qu'elles soient interprétables automatiquement par les machines. Ces méthodes consistent d'une part à créer des vocabulaires[6] de représentation des informations d'un domaine spécifique, et d'autre part à écrire des règles[7] qui vont permettre aux machines de déduire des informations à partir des documents (données) représentés dans les vocabulaires définis. Ainsi, plusieurs vocabulaires du Web sémantique ont été définis. Ils peuvent représenter des données issues des plateformes de réseaux socionumériques : FOAF[8] (description des personnes, de leurs liens, et de leurs activités), SIOC[9] (description des échanges entre personnes, *post* de blogs, forums, etc.), SKOS[10] (description des concepts associés à d'autres vocabulaires), etc. Bien que très peu de réseaux socionumériques offrent le moyen d'exporter les données de leurs utilisateurs sous ces formats, ces vocabulaires sont de plus en plus utilisés sur le Web. FOAF, par exemple, compte parmi les vocabulaires du Web sémantique les plus utilisés. La plateforme *Livejournal* permet d'extraire les informations de ses utilisateurs sous le format FOAF. Golbeck et Rothstein (2008) les utilisent pour la détection des identités multiples sur le Web, ou la détection de communautés basées sur les centres d'intérêts. La plateforme *Wordpress* permet l'extraction des données utilisateurs au format SIOC. Certains auteurs fusionnent tous ces vocabulaires afin de réaliser des analyses plus approfondies et plus riches (Breslin et Decker, 2007 ; Uldis, 2008).

Les méthodes basées sur les API de réseaux socionumériques s'appuient sur les API fournies par les réseaux socionumériques aux développeurs tiers (API *Facebook* et *OpenSocial* de *Google* notamment) pour développer des applications web qui seront embarquées dans les réseaux socionumériques. L'API *Facebook* est spécifique à la plateforme *Facebook*, alors qu'*OpenSocial* se veut interopérable et est utilisé dans plusieurs plateformes (*Orkut*, *LinkedIn*, *MySpace*, *Viadeo*, etc.). Toutefois, *Facebook* étant le précurseur de ce type d'API, il dispose jusqu'à aujourd'hui (et de loin) de la plateforme hébergeant le plus d'applications tierces[11]. D'autres modèles d'applications s'appuyant sur ces API permettent de développer des applications sur les graphes sociaux des réseaux socionumériques, mais sont plutôt embarquées dans les sites des développeurs ou des entreprises (*Facebook Connect*, *Google Friend Connect*, *MySpace Data Availability*, etc.). Les applications développées par ces API fournissent le moyen d'accéder à plusieurs types d'informations chez l'utilisateur : son profil explicite (informations d'identité, professionnelles, académiques, centres d'intérêts, etc.), ses activités (statuts, liens, photos, *tags*, commentaires, vidéos, groupes, événements, pages, flux d'activités, etc.), et son graphe social. Par rapport aux méthodes présentées précédemment, la diversité et la quantité des informations pouvant être extraites à partir de ces API peuvent enrichir considérablement les informations nécessaires à des analyses.

L'exécution des applications tierces (qui peuvent permettre l'extraction des données des profils) étant transparente aux utilisateurs, ces derniers peuvent facilement être la cible de diverses attaques ou d'atteintes aux données personnelles. Il peut s'agir d'attaques telles que la recomposition d'un réseau à partir des fragments de données accessibles (Tianjun et Hsinchun, 2008), les attaques sur la machine physique de l'utilisateur par scan des ports et exécution de scripts malicieux (Patakis *et al.*, 2009), la collecte massive des données des profils utilisateurs pour des usages douteux (Bonneau et Danezis, 2009)… Afin de comprendre et de présenter les enjeux réels de l'usage des API de réseaux socionumériques pour accéder aux données des utilisateurs, nous avons mené une expérimentation sur la plateforme *Facebook*.

Expérimentation de l'accès aux données *via* les API (cas de *Facebook*)

Afin de vérifier la possibilité d'accéder aux données des utilisateurs de réseaux socionumériques dans *Facebook*, nous avons étudié les fonctionnalités offertes par son API[12], et par la suite développé une application tierce afin de vérifier concrètement[13] la possibilité d'accéder aux données. Cette application est destinée uniquement à des utilisateurs qui ce sont portés volontaires pour participer à notre étude. Une charte sur les conditions d'extraction, de sauvegarde et de destruction des données est présentée en page d'accueil de cette application (diagramme 1).

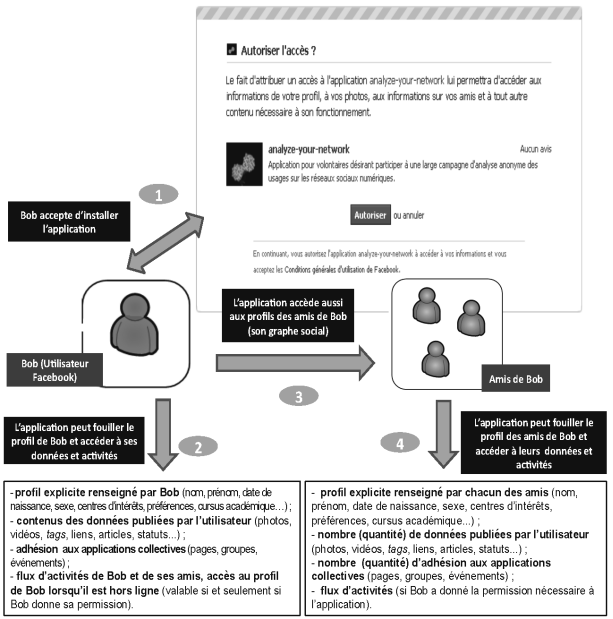

Diagramme 1 : Fonctionnement de l'application développée.

Lorsque l'utilisateur installe l'application, cette dernière accède par défaut (de manière transparente) à un certain nombre d'informations dans son profil, mais également à des informations dans le profil de tous ses amis. L'application développée a été installée par 85 utilisateurs volontaires. Cependant, du fait de l'accès aux profils des amis, au total 7 081 profils ont été consultés. Globalement, nous regroupons les informations accessibles par l'application dans quatre catégories :

– Informations accessibles par défaut dans le profil de celui qui installe l'application : *le profil explicite*[14] (nom et prénom, date de naissance, sexe, centres d'intérêts, préférences, établissement fréquentés, emplois occupés…) ; *le graphe social* (les amis de l'utilisateur, les relations entre ses amis[15], les listes d'amis créées par l'utilisateur…) ; *les informations publiées par l'utilisateur*[16] (mur, statuts, photos, albums, vidéos, liens, articles, *tags*, échantillon de commentaires…) ; *les adhésions à des applications collectives*[17] (groupes, pages, événements).

– Informations accessibles mais qui requièrent des autorisations explicites de l'utilisateur : *les flux d'activités*[18] (par lesquels on peut accéder aux activités des amis de l'utilisateur) ; *la permission d'accès hors ligne au profil* par l'application, *les permissions d'écriture sur le profil par l'application* (mise à jour du statut de l'utilisateur, envoie d'*e-mail* ou de SMS au nom de l'utilisateur, création d'événements ou d'articles au nom de l'utilisateur…).

– Informations accessibles par défaut dans le profil des amis de l'utilisateur de l'application : leur profil explicite, et les statistiques quantitatives sur les informations qu'ils ont publiées et sur les applications collectives (nombre de messages sur le mur, nombre de statuts, nombre de photos, nombre d'albums, nombre de vidéos, nombre de liens, nombre d'articles, nombre de *tags*, nombre de groupes, nombre de pages, nombre d'événements…).

– Informations jamais accessibles que ce soit sur le profil de l'utilisateur ou chez ses amis : il s'agit des échanges à caractère très privé tels que les *mails* ou la messagerie

instantanée, ou les contenus de messages postés sur les forums, les pages et les événements.

De manière générale, bien que *Facebook* définisse une politique de confidentialité vis-à-vis des développeurs (ou entreprises) utilisant son API, rien n'empêche techniquement à une application d'accéder (de manière transparente) à la quasi-totalité du profil et du graphe social de l'utilisateur qui l'installe. Bien que le réseau entier ne soit pas accessible, de nombreuses informations (graphe social de l'utilisateur, profil explicite de l'utilisateur et de ses amis, traces d'activités de l'utilisateur et de ses amis) peuvent faciliter la reconstitution automatique de réseaux égocentriques (réseaux de relations d'amitiés entre les amis d'un utilisateur). Le format de cet article ne permet pas de rentrer davantage dans les détails, mais des résultats d'analyses sur les données extraites peuvent être consultés dans Tchuente *et al.* (2011).

Dans cet article, nous avons présenté les principales méthodes d'extraction automatique de données publiées par les internautes sur le Web et dans les réseaux socionumériques en particulier. Si le passage du Web au Web 2.0 a énormément simplifié la publication et la gestion collaborative des contenus sur le Web par les internautes, l'accessibilité de plus en plus ouverte à ces contenus est problématique. Avec les API de réseaux socionumériques, par exemple, il devient possible pour tout utilisateur d'extraire et de manipuler (même sans autorisation préalable) un bon nombre de données personnelles et d'activités d'internautes. Au vu des usages de plus en plus importants des sites tels que les réseaux socionumériques, il semble important de sensibiliser leurs utilisateurs sur les risques d'accès non contrôlés au contenu de leurs profils.

NOTES

1. Ce travail de recherche est issu du projet de recherche « Réseaux sociaux numériques » mené durant une période de 24 mois en 2008-2009, financé par le groupe La Poste (Direction de l'Innovation et des E-services – DIDES et Mission Recherche et Prospective). Nous tenons ici à remercier celui-ci ainsi que les collègues des laboratoires CEREGE ayant participé à ce projet.
2. <http://www.comscore.com/press/release.asp?press=2725>.
3. Une application tierce est une application disponible sur un réseau socionumérique, mais développée par n'importe qui (programmeur autonome, entreprises, organisations, etc.).
4. Le profil public est la partie du profil accessible même aux internautes non inscrits sur le réseau socionumérique.
5. Ce que certains auteurs appellent Web 3.0.
6. Sous forme de description RDF (*Resource Description Framework*).
7. Avec des langages comme OWL (*Web Ontology Language*).
8. *Friend Of A Friend* <http://www.foaf-project.org/>.
9. *Semantically-Interlinked Online Communities* <http://sioc-project.org/>.
10. *Simple Knowledge Organization-System* <http://www.w3.org/2004/02/skos/>.
11. Statistiques *Facebook* <http://www.facebook.com/press/info.php?statistics>.
12. <http://developers.facebook.com/docs/>.
13. <http://apps.facebook.com/analyze_network>.
14. Informations que l'utilisateur renseigne explicitement (en général, lorsqu'il crée son profil).

15. Ces relations entre amis de l'utilisateur définissent son réseau égocentrique.
16. Avec les détails associés à chaque information (date et lieu de publication, contenu, etc.).
17. Dates d'adhésion et caractéristiques de l'application collective (nom, type, description, nombre de fans...).
18. Les activités des amis de l'utilisateur qui apparaissent dans sa page d'accueil.

RÉFÉRENCES BIBLIOGRAPHIQUES

ALIM, S., ABDUL-RAHMAN, R., NEAGU, D. et MICK, R., «Data Retrieval from Online Social Network Engineering Applications» *in Technology and Secured Transactions*, ICITST, 2009, p. 1-5.

BONNEAU, J., ANDERSON, J. et DANEZIS, G., «Prying Data out of a Social Network» *in Advances in Social Network Analysis and Mining*, ASONAM, 2009, p. 249-254.

BOYD, D. et ELLISON, N., «Social Network Sites: Definition, History, and Scholarship», *Journal of Computer-Mediated Communication*, vol. 13, 2007.

BRESLIN, J. et STEFAN, D., «The Future of Social Networks on the Internet», *IEEE Internet Computing*, déc. 2007, p. 86-90.

COUTANT, A. et STENGER, T., «Processus identitaire et ordre de l'interaction sur les réseaux socionumériques», *Les Enjeux de l'Information et de la Communication*, août 2010. En ligne sur <http://w3.u-grenoble3.fr/les_enjeux/2010/Coutant-Stenger/index.html>, consulté le 19/01/2011.

DIMATRACOUPOULOU, A. et BRUILLARD, E., «Enrichir les interfaces de forums par la visualisation d'analyses automatiques des interactions et du contenu», *Revue Sticef*, vol. 13, 2006.

DWYER, C., HILTZ, R. et GEORGE, W., «Understanding Development and Usage of Social Networking Sites: The Social Software Performance Model», *Proceedings of the 41st Hawaii International Conference on System Sciences*, 2008.

GOLBECK, J. et ROTHSTEIN, M., «Linking Social Networks on the Web with FOAF: A Semantic Web Case Study», *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, 2008.

JIN, Y., MATSUO, Y. et ISHIZUKA, M., «Extracting a Social Network among Entities by Web Mining», *ESWC*, 2007.

MIKA, P., «Flink: Semantic Web Technology for the Extraction and Analysis of Social Networks», *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 3, n° 2-3, 2005, p. 211-223.

MATSUO, Y., HAMASAKI, M., TAKEDA, H., NISHIMURA, T., HASIDA, K. et ISHIZUKA, M. «POLYPHONET: An Advanced Social Network Extraction System», *Proceedings Word Wide Web Conference,* 2006.

NAGLE, F. et SINGH, L., «Can Friends be trusted? Exploring Privacy in Online Social Networks» *in Advances in Social Network Analysis and Mining*, ASONAM, 2009, p. 312-315.

PATSAKIS, C., ASTHENIDIS, A. et CHATZIDIMITRIOU, A., «Social Networks as an Attack Platform: *Facebook* Case Study», *Networks*, ICN, 2009, p. 245-247.

REFFAY, C. et LANCIERI, L., «Quand l'analyse quantitative fait parler les forums de discussion», *Revue Sticef*, vol. 13, 2006.

SIDIR, M., LUCAS, N. et GIGUET, E., «De l'analyse des discours à l'analyse structurale des réseaux sociaux : une étude diachronique d'un forum éducatif», *Revue Sticef*, vol. 13, 2006.

STUTZMAN, F., «An Evaluation of Identity-sharing Behavior in Social Networks Communities», *iDMAa Journal*, vol. 3, n° 1, 2006. En ligne sur <http://www.ibiblio.org/fred/pubs/stutzman_pub4.pdf>, consulté le 19/01/2011.

TCHUENTE, D., CANUT, M.-F., BAPTISTE-JESSEL, N., COUTANT, A., STENGER, T. et RAMPNOUX, O., «Pour une approche interdisciplinaire des TIC : le cas des réseaux socionumériques», *Document Numérique*, mars 2011 [à paraître].

TIANJUN, F. et HSINCHUN, C., «Analysis of cyberactivism: A Case Study of Online Free Tibet Activities», *Intelligence and Security Informatics*, 2008.

ULDIS, B., PASSANT, A., CYGANIAK, R. et BRESLIN, J., «Weaving SIOC into the Web of Linked Data», Beijing, LDOW, 2008.