Introduction à l'analyse des réseaux sociaux *

Maria Malek
LARIS-EISTI

3 novembre 2009

1 Introduction

L'analyse des réseaux sociaux est définie comme étant l'étude des entités sociales (les personnes dans les organisations qu'on appelle acteurs) ainsi que leurs interactions et leurs relations. Ces interactions et relations peuvent être représentées par un graphe ou un réseau, dans lequel chaque nœud représente un acteur et chaque lien est une relation. Nous pouvons étudier les propriétés de la structure et sa rôle ainsi que la position et le prestige de chaque acteur social [11]. Nous pouvons rechercher aussi les différents types de sous-graphes comme par exemple les communautés formées par des groupes d'acteurs ayant des intérêts commun, en isolant le groupe d'individus ayant une densité élevée. Les réseau social peut être aussi une source permettant l'élaboration de recommandations : trouver un expert dans un domaine donné, suggérer des produits à vendre, proposer un ami, etc. Cette élaboration peut être fondée sur des algorithmes d'exploration de chemins, d'analyse de degrés,etc[1][4].

Les réseaux sociaux peuvent être modéliser par des graphes [6, 2] ayant des propriétés spécifiques, nous citons :

Le diamètre du réseau : (le plus long chemin entre deux nœuds) est de longueur bornée. Cette propriété est appelée la propriété du petit monde et issue des études sociologiques annonçant qu'il existe 6 degrés de séparation selon l'expérience du petit monde de Stanley Milgram 1967 [1].

*Rapport technique - LARIS-EISTI, octobre 2009.

1. 2 personnes choisies aléatoirement sont reliées au plus par 6 personnes intermédiaires.

La transitivité dans le réseau : qui est basée sur le postulat : *l'ami de mon ami est mon ami.* Cette propriété se traduit par le nombre de triangles dans le réseau d'où la mesure du taux de clustering suivant :

$$C = \frac{3 * \texttt{nombre de triangles dans le réseau}}{\texttt{nombre de triples connectés de nœuds}}$$

La distribution des degrés : (nombre de nœuds connectés à un nœud donné) obéit à la distribution de degrés en loi de puissance.

Nous avons évoqué la modélisation d'un réseau social par un graphe, or la description du réseau peut nécessiter d'éléments décrivant le contenu sémantique des acteurs et/ou des liens. Il est possible donc d'envisager une modélisation sémantique qui consiste à :

- stocker au niveau des acteurs (nœuds) les information concernant le profil utilisateur ;
- stocker au niveau de lien une plusieurs informations : types de communication (email, etc), nombre ou fréquence d'interaction, la durée, nombre de participants, etc.

Les réseaux sociaux sont utilisés principalement pour décrire les interactions entre les entités sociales. Les réseaux sociaux sont encore utilisés pour modéliser d'autres types d'interaction comme les liens entre un ensemble de pages Web ou bien les citations bibliographiques dans une corpus de documents.

L'analyse des réseaux sociaux est utile pour la toile parce que le Web est une société virtuelle et donc un réseau social virtuel, où chaque page est considérée comme un acteur social et chaque hyperlien est une relation. Un certain nombre de résultats issus des études des réseaux sociaux peuvent être adaptés et étendus dans le contexte des applications Web. Les idées développées dans le domaine de l'analyse des réseaux sociaux ont largement contribué au succès des moteurs de recherche[3].

De même, l'analyse des citations exprimées dans les références bibliographique permet de découvrir des similarités entre ces documents outre les mesure fondées sur le contenu, et de définir des centres d'affinités entre les auteurs, etc.

Par ailleurs, nous citons le terme *réseautage social* qui signifie la communication basée sur le partage et les commentaires des contenus. La problématique qui se pose ici sera l'extraction d'un réseau social dans un espace de travail collaboratif ou en analysant les interactions entre les internautes.

Dans les sections suivantes, nous détaillons les mesures utilisés dans des réseaux sociaux. Les mesures locales sont les indices permettant de donner des informations locales sur un acteur donné comme la centralité ou le prestige.

Les mesures globales sont porteuses d'éléments d'information globale sur le réseau.

Nous terminons ce rapport par une section décrivant les outils et les plateformes des réseaux sociaux à explorer.

2 Les mesures locales

Dans cette section, nous introduisons deux types de notions pour l'analyse des réseaux sociaux [5] : la centralité et le prestige qui sont bien liés à l'analyse des hyperliens et la recherche dans le Web. La centralité ainsi que le prestige sont des mesures de degrés de la pertinence d'un acteur dans un réseau social.

2.1 La centralité

les acteurs importants sont ceux qui sont liés et impliqués avec les autres acteurs d'une façon extensive. Dans le cadre d'une organisation, une personne qui a beaucoup de contacts et qui communiquent bien avec les autre personnes est considérée plus importante qu'une autre personne ayant moins de contacts. Ces contacts sont modélisés par des liens. Un acteur central est un acteur qui est impliqué dans plusieurs liens. La figure 2.1 montre un exemple simple qui utilise un graphe non dirigé. Chaque nœud dans le réseau social est un acteur et chaque lien indique que les deux acteurs aux extrémités communiquent ensemble. Intuitivement, nous remarquons que l'acteur 1 est l'acteur le plus central parcequ'il/elle communique avec la majorité des autres acteurs. Il y a différents types de liens ou d'implications entre acteurs. Par conséquence, plusieurs types de centralités seront définies sur les graphes dirigés et non dirigés. nous détaillons ces mesures dans la suite.

Degré de Centralité Les acteurs centraux sont les acteurs les plus actifs et qui ont le plus de liens avec les autres acteurs. Soit n le nombre total d'acteurs dans le réseaux.

Graphe non dirigé Dans un graphe non dirigé, le degré de centralité d'un acteur i noté $C_D(i)$ est le degré du nœud acteur (le nombre d'arêtes) noté $d(i)$ normalisé par le degré maximal, $n-1$.

$$C_D(i) = \frac{d(i)}{n-1}$$

La valeur de cette mesure varie entre 0 et 1 puisque $n-1$ est la valeur maximale de $d(i)$

Graphe dirigé Dans ce cas, nous distinguons entre les liens entrants d'un acteur i (les liens pointant vers i) et les liens sortants (les liens pointant à partir de i). Le degré de centralité est définie à partir des liens sortants (le nombre des liens sortants :$d_o(i)$).

$$C'_D(i) = \frac{d_o(i)}{n-1}$$

Centralité de proximité Cette approche définit la centralité en utilisant la notion de proximité ou de distance. X_i est un acteur central s'il peut interagir facilement avec les autres acteurs. Par conséquence, sa distance avec les autres doit être courte. Nous utilisons donc la distance la plus courte pour calculer cette mesure. Soit $d(i,j)$ la distance la plus courte à partir de l'acteur i vers l'acteur j (mesurée par le nombre de liens via le chemin le plus court)

Graphe non dirigé La centralité $C_C(i)$ de proximité d'un acteur i est définie par :

$$C_C(i) = \frac{n-1}{\sum_{j=1}^{n} d(i,j)}$$

La valeur de cette mesure varie entre 0 et 1 comme $n-1$ est la valeur minimale du dénominateur, qui correspond à la somme de la distance la plus courte à partir de l'acteur i vers les autres acteurs. Noter bien que cette equation est valable seulement dans un graphe connecté.

Graphe dirigé La même equation est utilisée pour les graphes dirigés. Le calcul de la distance doit intégrer les directions des liens ou les arcs.

Centralité d'intermédiarité Si deux acteurs non adjacents j et k veulent communiquer et si l'acteur i se localise sur le chemin entre j et k, alors i a un certain contrôle sur leur interaction. L'intermédiarité mesure ce contrôle de i sur les deux autres acteurs. Par conséquent, si i se localise sur le chemin de plusieurs interactions alors i est un acteur important.

Graphe non dirigé Soit p_{jk} le nombre des chemins les plus courts entre les deux acteurs j et k. L'intermédiarité d'un acteur i est définie par le nombre des chemins les plus courts entre j et k passant par i notés par $p_{jk}(i)$ (avec $j \neq i$ et $k \neq i$),w normalisé par le nombre total des chemins les plus courts entre tous les pairs d'acteurs qui n'incluent pas i :

$$C_B(i) = \sum_{j<k} \frac{p_{jk}(i)}{p_{jk}}$$

Noter bien qu'il peut exister plusieurs chemins plus courts entre l'acteur j et l'acteur k. Quelques-uns passent par i et d'autres non. Nous supposons que tous les chemins ont la même chance d'être utilisés. $C_B(i)$ a un minimum de 0, atteint quand i ne figure sur aucun des chemins les plus courts. Le maximum atteint est de $\frac{(n-1)(n-2)}{2}$ qui correspond au nombre de pairs d'acteurs qui ne contiennent pas i.

Dans le réseau (figure 2.1) l'acteur 1 est l'acteur le plus central). Il figure sur tous les 16 plus cours chemins qui lient les 6 autres acteurs. $C_B(1)$ atteint la valeur maximale de 15 tandis que $C_B(2) = C_B(3) = C_B(4) = C_B(5) = C_B(6) = C_B(7) = 0$.
Remarquer bien que la mesure de co-citation est symétrique.

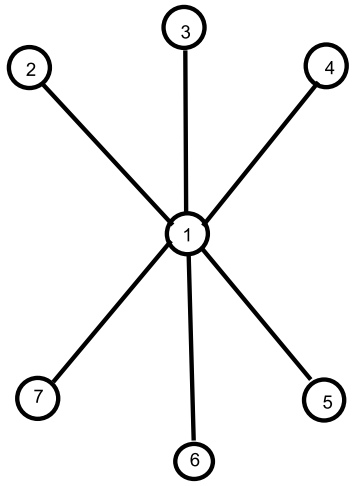

FIGURE 1 – Exemple d'un réseau social centré sur un acteur

Si nous souhaitons que la valeur de cette mesure varie entre 0 et 1, nous pouvons normaliser par $\frac{(n-1)(n-2)}{2}$, qui correspond à la valeur maximale de $C_B(i)$. L'intermédiarité standard de l'acteur i est définie par :

$$C'_B(i) = \frac{2 \sum_{j<k} \frac{p_{jk}(i)}{p_{jk}}}{(n-1)(n-2)}$$

A l'opposé de la mesure de proximité, l'intermédiarité peut être calculée même si le graphe n'est pas fortement connecté.

Graphe dirigé La même relation peut être utilisée mais doit être multipliée par 2 car nous considérons dans ce cas $(n-1)(n-2)$ pairs puisqu'un

chemin de j vers k est different du chemin inverse allant de k vers j. De même, p_{jk} considèrent les chemins dans les deux directions.

2.2 Le prestige

Le prestige est une mesure plus raffinée que la centralité. Nous distinguons entre les liens sortants et ceux entrants. Un acteur prestigieux est un acteur ayant beaucoup de liens entrants.

Le prestige d'un acteur est mesuré par le nombre des liens entrants. Nous citons Les mesures suivantes :

le degré de prestige donné par la relation :

$$P_D(i) = \frac{d_I(i)}{n-1}$$

avec $P_I(i)$ le degré entrant du nœud i, n étant toujours le nombre d'acteurs (nœuds) dans le réseau

le prestige de proximité :

Soit I_i l'ensemble d'acteurs qui atteignent l'acteur i : soit j un acteur appartenant à I_i ($j \in I_i$), et soit $d(j,i)$ le chemin le plus court de j vers i ; la distance moyenne est $\frac{\sum_{j \in I_i} d(j,i)}{|I_i|}$ où $|I_i|$ étant la norme (le nombre d'éléments) de I_i. La proximité est définie comme étant le rapprochement des auteurs acteur de l'acteur i. Le prestige de proximité est défini comme étant :

$$P_p(i) = \frac{\frac{|I_i|}{n-1}}{\frac{\sum_{j \in I_i} d(j,i)}{|I_i|}}$$

Remarquer bien que $P_p(i) \in [0,1]$ et que plus les autres acteurs sont prêts de l'acteur i plus cette mesure est élevée.

le prestige de tri (rank prestige) : Les mesures proposées jusqu'au là sont fondées sur les liens entrants et sortants d'un acteur donné. La mesure du prestige du tri considère la réputation et l'importance des acteurs choisissant l'acteur i ! Ce prestige est calculé selon la formule :

$$P_R(i) = A_{1i}P_R(1) + A_{2i}P_R(2) + .. + A_{ni}P_R(n)$$

où A étant la matrice d'adjacence ; et $P_R(i)$ étant la mesure de prestige de tri lié à l'acteur i. Remarquer bien que le prestige de tri de l'acteur i dépend des prestiges de tri des acteurs j ayant des liens sortants vers l'acteur i (ji).

Pratiquement, pour calculer le prestige de tri on procède ainsi :

- A_{ij} vaut 1 si i pointe vers j,sinon 0
- On pose P le vecteur contenant les valeurs de tri des n acteurs :

$$P = (P_R(1), P_R(2), .., P_R(n))^T$$

- A est la matrice exprimant la propriété adjacente, on a donc :

$$P = A^T P$$

- Recherche de valeurs propres de A^T

Noter bien que cet algorithme est utilisé pour trié les résultats des moteurs de rechercher comme Google (algorithme PageRank).

2.3 Les mesures de citations bibliographiques

Nous traitons dans cette partie les mesures de citation bibliographiques. Une publication sous n'importe quelle forme contient une partie de citations d'autres publications. Quand un papier i cite une autre papier j, un lien est crée entre les deux papiers dans le sens i vers j. Ce lien peut donner une indication de relations entre auteurs, papiers, pays,etc. Nous présentons deux mesures de citations : la co-citation et le couplage bibliographie.

2.3.1 La co-citation

La co-citation est une mesure de similarités entre deux documents qui exprime le fait que si les papiers i et j sont cités par le papier k alors ils sont liés. De même, si les papiers i et j sont cités par plusieurs papiers alors ils sont *similaires*.

Les citations sont modélisées par une matrice L appelée la matrice de citation ; le terme L_{ij} vaut 1 si i cite j, sinon il vaut 0. La co-citation est une mesure qui est définie par :

$$C_{ij} = \sum_{k=1}^{n} L_{ki} L_{kj}$$

Remarquer bien que la mesure de co-citation est symétrique.

2.3.2 Le couplage bibliographique

La couplage bibliographique est une mesure de similarités entre deux documents qui exprime le fait que si les papiers i et j citent le papier k alors ils sont liés. De même, si les papiers i et j citent plusieurs papiers alors ils sont *similaires*. Si les papiers i et j citent plusieurs papiers alors ils

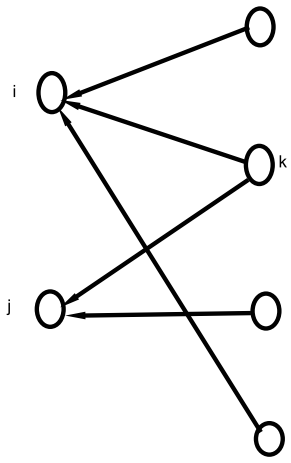

FIGURE 2 – Exemple de co-citation entre i et j

sont similaires. Les citations sont modélisées par une matrice L appelée la matrice de citation ; le terme L_{ij} vaut 1 si i cite j, sinon il vaut 0. Le couplage bibliographique est une mesure qui est définie par :

$$B_{ij} = \sum_{k=1}^{n} L_{ik} L_{jk}$$

Remarquer bien que la mesure du couplage bibliographique est symétrique.

3 Les mesures globales

Il existe certaines mesures globales qui permettent d'avoir une idée sur la structure globale du réseau, nous citons :

– La densité du graphe : permet de décrire la connectivité à l'intérieur du graphe représentant le réseau :

$$D = \frac{m}{\frac{n*(n-1)}{2}}$$

où m étant le nombre de liens dans le graphe, n étant le nombre de nœuds, remarquer bien que le nombre maximal de liens dans un graphe est $\frac{n*(n-1)}{2}$.

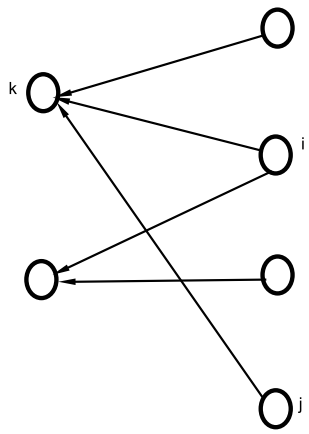

FIGURE 3 – Exemple de couplage bibliographique entre i et j

- La distance géodésique entre deux nœuds est le plus court chemin entre les deux nœuds.
- La distance moyenne d'un graphe connecté est égale à la moyenne des distances géodésiques entre toutes les pairs d'acteurs.
- Le diamètre d'un graphe connecté est égal à la distance géodésique maximale au sein d'un groupe.

4 Standards, Outils & Plateformes

Nous citons un ensemble de standards et d'Ontologies permettant la modélisation d'un réseau social :

FOAF (Friend Of A Friend) qui est fondé sur RDF, il permet de modéliser les profils acteurs-utilisateurs, ainsi que les interactions entre les utilisateurs. Le profil utilisateur est décrit et stocké du côté acteur [10].

XFN : permet de commenter les liens afin d'indiquer les relations personnelles.

SIOC (Semantically Interlinked Online Communities) est un standard pour exprimer l'information contenue dans les forums de discussion, blogs, mailing liste, etc.

Ainsi, nous présentons une liste non exhaustive d'outils permettant la modélisation sémantique d'un réseau social :

Jena est une plateforme Java pour la construction des applications Web avec RDF, RDFs, OWL, SparQL, elle inclut un moteur basé sur les règles d'inference.

JUNG est un outil de programmation pour l'analyse des réseaux sociaux (librairie Java) contenant des algorithmes de traitement de graphes, de fouille de données et d'analyse de réseaux sociaux.

Prefuse est un outil de visualisation des données des réseaux sociaux.

socialAction est un outil permettant la visualisation et statistique pour le processus d'analyse

Banac est une collection d'opérateurs qui fonctionne en utilisant les graphes RDF, inférer étendre et émerger des graphes.

Distiller Banach operator permet d'extraire des sous graphes selon des propriétés données.

Ucinet [3]est un logiciel d'analyse de réseaux sociaux.

Références

[1] Ecole d'été web intelligence. In *WI09*. Université de Lyon, 2009.

[2] C. Berge. *Graphes*. Gauther-Villars, 1983.

[3] S. P. Borgatti, M. G. Everett, and L. C. Freeman. *Ucinet for Windows : Software for Social Network Analysis*, 2002.

[4] G. Ereteo, F Gando, M. Buffa, and P. Grohan. Analyse des réseaux sociaux et web sémantique : un état de l'art. Technical report, Projet ISICIL (ANR), 2009.

[5] M. G. Everett and S. P. Borgatti. The centrality of groups and classes. *Journal of Mathematical Sociology*, 23(3) :181–201, 1999.

[6] J-C. Fournier. *Théorie de Graphes et applications*. Lavoisier, 2006.

[7] Bing Liu. *Web Data Mining : Exploring Hyperlinks, Contents, and Usage Data*. Data-Centric Systems and Applications. Springer, 2007.

[8] P. Mika. Bootstrapping the foaf-web : An experiment in social network mining, 2004.

[9] Peter Mika. Flink : Semantic web technology for the extraction and analysis of social networks. *Web Semantics : Science, Services and Agents on the World Wide Web*, 3(2-3) :211–223, October 2005.

[10] Peter Mika. Ontologies are us : A unified model of social networks and semantics. In *International Semantic Web Conference*, LNCS, pages 522–536. Springer, 2005.

[11] M. E. J. Newman. The structure and function of complex networks. Mar 2003.