Atelier 5

Le Web social

Hakim Hacid et Cécile Favre

Le Web Social 2010

En conjonction avec
10ème Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC 2010).

Organisé par :

Hakim Hacid
Alcatel-Lucent Bell Labs
France

Cécile Favre
Laboratoire ERIC
Université Lyon 2
France

Hammamet, Tunisie, 26 Janvier 2010

Préface Atelier Web Social - EGC 2010

Préambule

Cette première édition de l'atelier sur le Web social se déroulant dans le cadre d'EGC 2010 vise à permettre la rencontre des chercheurs et des jeunes chercheurs, à la fois du monde académique et industriel, autour des problématiques liées au Web social en général et à l'extraction de connaissances à partir du Web social en particulier. Il s'agit entre autres de confronter les idées afin d'avoir une vision plus claire des éléments qui entourent ce nouveau phénomène, faire un état des lieux des avancements dans les différentes pistes composant ce nouveau Web et enfin tenter de ressortir les verrous scientifiques et industriels à court, moyen et long termes, à lever autour du Web social.

En effet, avec l'avènement du Web 2.0, l'utilisateur est au centre des préoccupations des différentes technologies composant ce nouveau modèle comme les mashups, les environnements collaboratifs, les réseaux sociaux, etc. Le principal ingrédient rajouté est le social qui consiste à mettre en relation les utilisateurs, à leur faciliter l'interaction et à la rendre plus riche et plus productive. Le Web social devient ainsi de plus en plus la partie la plus intéressante de tout le Web, au point de défier de grands acteurs bien établis sur le Web traditionnel comme le moteur de recherche Google. Ceci constitue une énorme avancée d'un point de vue utilisateur et ouvre aussi de grandes perspectives de recherche dans un environnement qui devient de plus en plus complexe, moins structuré et plus hostile compte tenu de la grande masse d'information généralement cachée à l'utilisateur.

Les réseaux sociaux concentrent certainement la majeure partie des travaux qui sont faits autour du Web social. Les travaux dans ce domaine se focalisent principalement sur les propriétés structurelles, e.g. la force des liens sociaux, le key player, etc. Au-delà des réseaux sociaux, le social se manifeste sous d'autres formes et dans d'autres endroits sur le Web : les médias sociaux tels que Youtube ou Flickr, les news sociales telles que Twitter ou Digg, le bookmarking social comme Delicious (del.icio.us). Toutes ces parties constituent un énorme réservoir d'informations sociales qui renferme des connaissances pouvant être utiles à l'utilisateur. Ceci peut se manifester éventuellement par la mise en place de nouveaux services à valeur ajoutée exploitant cette connaissance qui est très faiblement exploitée par les utilisateurs et les fournisseurs de services actuellement.

Ainsi, les thématiques d'intérêt pour cette atelier incluent : les réseaux sociaux ; les phénomènes sociologiques du Web social ; les fournisseurs de services et Web social ; les réseaux sociaux en entreprise ; le Web sémantique et Web social ; l'analyse de données sociales et l'extraction de connaissances ; l'exploitation de l'analyse sociale ; la personnalisation de contenu et de services ; les modèles de monétisation du Web social ; la recherche « sociale » d'information ; l'extraction et la structuration de données à partir de plateformes sociales ; la modélisation des données sociales ; l'interrogation de données sociales ; le Web social et la mobilité ; l'extraction et l'analyse de communautés ; la vie privée dans le Web social ; la qualité de l'information dans le Web social ; les méthodes de filtrage de l'information sociale ; la portabilité de l'information dans le Web social ; les techniques de veille numérique. Les soumissions reçues ont abordé un large éventail de ces thématiques mais les plus dominantes sont sans doute les aspects sémantiques et communautaires dans les environnements sociaux.

Nous espérons que cette édition soit le début d'une longue série dans le futur afin d'inscrire dans le temps la possibilité de réunir tous les acteurs qui contribuent à cette thématique. Enfin, il nous semble pertinent de vouloir pousser cet événement sur la scène internationale qui, nous le souhaitons, sera l'avenir de cet atelier.

Remerciements

Les responsables de l'atelier souhaitent remercier vivement toutes les personnes ayant contribué à la tenue de cet atelier. En particulier :

- les auteurs pour la qualité de leurs contributions constituant la base essentielle de discussions fructueuses ;
- les membres du comité de lecture dont le travail d'évaluation était crucial pour assurer la qualité de l'atelier ;
- Alexandre Passant pour avoir accepté de donner un exposé sur les thématiques du Web sémantique et des réseaux sociaux ;
- les organisateurs d'EGC qui ont mis en place l'environnement et les moyens pour la réussite des ateliers.

Comité de lecture

- Frédéric Amblard, Université Toulouse 1
- Michaël Aupetit, CEA LIST
- Zohra Bellahsene, Université Montpellier 2
- Sonia Ben Mokhtar, Université Lyon 1
- Amel Bouzeghoub, Télécom & Management SudParis
- Johann Daigremont, Alcatel-Lucent Bell Labs France
- Jérôme David, INRIA Grenoble Rhône-Alpes
- Jean-Gabriel Ganascia, Université Paris 6
- Fabien Gandon, INRIA Sophia Antipolis
- Samir Ghamri-Doudane, Alcatel-Lucent Bell Labs France
- Fabrice Guillet, Polytech'Nantes
- Rushed Kanawati, Université Paris 13
- Luigi Lancieri, Orange Lab
- Christine Largeron, Université Saint-Etienne
- Bénédicte Le Grand, Université Paris 6
- Nicolas Lumineau, Université Lyon 1
- Linas Maknavicius, Alcatel-Lucent Bell Labs France
- Pierre Maret, Université Saint-Etienne
- Alexandre Passant, DERI-National University of Ireland
- Nathalie Pernelle, Université Paris 11
- Mathieu Roche, Université Montpellier 2
- Fatiha Sais, Université Paris 11
- Yacine Sam, Université de Tours
- Vincent Toubiana, New York University
- Julien Velcin, Université Lyon 2
- Gilles Venturini, Université de Tours
- Emmanuel Viennet, Université Paris 13

Relecteurs additionnels

- Cédric Lopez, Université Montpellier 2
- Anna Stavrianou, Université Lyon 2

Les responsables de l'atelier Web Social – EGC 2010
Hakim Hacid, Alcatel-Lucent Bell Labs France
Cécile Favre, Laboratoire ERIC - Université Lyon 2

De l'intérêt du Web Sémantique pour le Web Social, et réciproquement

Alexandre Passant
Digital Enterprise Research Institute (DERI)
National University of Ireland, Galway
IDA Business Park, Lower Dangan
Galway, Ireland
alexandre.passant@deri.org

Résumé. Ces dernières années ont vu la montée en puissance de deux visions du Web: d'un coté le Web Sémantique, qui propose des formalismes de représentations unifiées dans une optique d'échange et de compréhension des données à l'échelle du Web; de l'autre le Web Social (ou Web 2.0) vision plus pragmatique qui met l'accent sur la place centrale de l'utilisateur au sein de la démarche de production de contenus. Si celles-ci ont souvent été opposées, nous montrerons dans cet exposé en quoi elles ne sont en réalité pas contradictoires mais au contraire complémentaires et de quelle manière elles peuvent bénéficier chacune des apports de l'autre.

Ainsi, nous présenterons d'une part l'intérêt des formalismes du Web Sémantique (RDF, RDFS, OWL, SPARQL) pour la représentation et l'interrogation de données issues d'applications Web 2.0. Plus spécifiquement, nous détaillerons le rôle joué par des ontologies comme FOAF - Friend Of A Friend - et SIOC - Semantically-Interlinked Online Communities - dans ce contexte, ainsi que différentes applications reposant sur les standards précédents. D'autre part, nous détaillerons en quoi les paradigmes mis en avant par le Web Social (partage, collaboration, ouverture) permettent la création à grande échelle de connaissances représentées selon les formalismes précédents, notamment au sein l'initiative Linking Open Data, que cela se fasse par l'intermédiaire d'exporteurs pour des applications Web 2.0 existantes ou via de nouveaux systèmes comme les wikis sémantiques.

De plus, tout au long de cet exposé, nous présenterons différentes applications actuellement déployées sur le Web ou en entreprise et mettant en avant cette complémentarité entre Web Social et Web Sémantique, conduisant a un Web optimisé a la fois pour les humains et les machines, au niveau des modes de publication pour le premier et de la modélisation des données pour le second.

Découverte de communautés par analyse des usages

Lylia Abrouk, David Gross-Amblard, Damien Leprovost

Laboratoire Le2i-CNRS
Université de Bourgogne, France
{prénom.nom}@u-bourgogne.fr
http://www.u-bourgogne.fr/LE2I

Résumé. Dans les sites Web collaboratifs actuels, un effort de saisie important est demandé aux utilisateurs afin d'identifier la communauté à laquelle ils appartiennent (description du profil personnel, du réseau social, etc.). Dans cet article, nous proposons une méthode de découverte de communautés basée sur les actions des utilisateurs. Elle repose sur une analyse en composantes principales des usages (ACP) et a été validée sur une base de données de préférences filmographiques de grande taille (MovieLens).

1 Introduction

Depuis quelques années, le Web s'est transformé en une plateforme d'échange générique, où tout utilisateur devient un fournisseur de contenu par le biais de technologies comme les commentaires, les blogs et les wikis. Ce nouveau Web collaboratif ou participatif (Web 2.0) comprend des sites populaires comme Myspace[1], Facebook[2] ou Flickr[3], permettant de construire des réseaux sociaux selon ses relations professionnelles ou ses intérêts. Cependant, ces sites exigent de chaque utilisateur une description explicite de son réseau social ou de son profil. De plus, seules les communautés ainsi explicitées sont identifiées.

Or un grand nombre de communautés d'utilisateurs existent de façon implicite dans de nombreux domaines. Par exemple, tout site de musique généraliste rassemble une communauté d'utilisateurs ayant des goûts musicaux variés. Mais cette communauté est en fait composée de sous-communautés potentiellement disjointes, toutes liées à la musique (la communauté des amateurs de musique pop, de musique punk, etc.). Découvrir et identifier précisément ces communautés implicites est un gain pour de nombreux acteurs : le propriétaire du site, les régies publicitaires en ligne et surtout, les utilisateurs du système.

Dans cet article, nous proposons une méthode de détection de communautés. La méthode est générique car elle ne s'appuie que sur un étiquetage des ressources et sur l'utilisation de ces ressources par les utilisateurs (par exemple, tel utilisateur consulte tel fichier musical, étiqueté `rock`). Le cœur de notre méthode est une analyse statistique en composantes principales (ACP (Falissard, 2005)) des étiquettes des ressources manipulées par les utilisateurs. Cette méthode permet de représenter les données originelles (utilisateurs et étiquettes manipulées) dans

[1] http ://www.myspace.com
[2] http ://www.facebook.com
[3] http ://www.flickr.com

un espace de dimension inférieure à celle de l'espace originel, tout en minimisant la perte d'information. La représentation des données dans cet espace de faible dimension en facilite considérablement l'analyse et permet ainsi de regrouper ou d'opposer des communautés.

L'article est organisé de la façon suivante. La section 2 présente notre approche de détection de communautés. Cette approche est validée expérimentalement en section 3 sur une bases de données de préférences filmographiques de grande taille (MovieLens). L'état de l'art est présenté en section 4. Conclusion et perspectives sont présentées en section 5.

2 Modèle

Premières définitions On considère un ensemble d'utilisateurs $U = \{u_1, \ldots, u_n\}$ et un ensemble de ressources R sur un site donné (par exemple des fichiers de musiques, des vidéos, des nouvelles). Nous supposons que les utilisateurs émettent un vote sur un sous-ensemble des ressources du site. Ce vote n'est pas nécessairement explicite et peut être obtenu en se basant sur les usages des utilisateurs (la musique qu'ils sélectionnent, les titres qu'ils achètent, les ressources qu'ils annotent ou recommandent). Les votes sont illustrés par une matrice M : $|U| \times |R|$ définie comme suit, pour un utilisateur $u_i \in U$ et une ressource $r_j \in R$:

$$M(u_i, r_j) = \begin{cases} 1 & \textit{si } u_i \textit{ a de l'intérêt pour } r_j, \\ 0 & \text{sinon.} \end{cases} \quad (1)$$

Cette matrice est mise à jour dynamiquement lorsque de nouveaux utilisateurs, de nouvelles ressources ou de nouveaux usages apparaissent sur le site. Nous supposons également qu'un ensemble de *tags* $T = \{t_1, \ldots, t_m\}$ est défini (par exemple, musique `pop`, `rock`, `punk`, etc.), et que chaque ressource est annotée avec un sous-ensemble de ces tags (sous-ensemble potentiellement vide). Ces annotations proviennent des fournisseurs de ressources, qui peuvent être les utilisateurs eux-même, et peuvent s'enrichir au fur et à mesure. Étant donnés les votes des utilisateurs et ces annotations, nous définissons l'ensemble $A(u_i) \subseteq R$ des ressources intéressant l'utilisateur $u_i \in U$ et l'ensemble $A(u_i, t_j) \subseteq R$, où $t_j \in T$, l'ensemble des ressources intéressant u_i et annotées par le tag t_j.

L'objectif principal de l'approche proposée est de scinder les utilisateurs en communautés distinctes, en se basant sur les groupes de tags qu'ils apprécient. Nous calculons le degré d'appartenance x_{ij} d'un utilisateur u_i à un tag t_j :

$$x_{ij} = \frac{|A(u_i, t_j)|}{|A(u_i)|}. \quad (2)$$

Plus un coefficient x_{ij} est proche de 1, plus l'utilisateur i manipule des tags de type j.

Communautés de tags On cherche ensuite à rassembler les tags similaires, de façon statistique. Pour cela, on utilise la technique de l'analyse en composantes principales (ACP). Dans cette section, nous donnons l'intuition de cette méthode, les détails étant explicités en section 3.

Dans la suite, l'usage d'une ressource portant un tag donné est vu comme la réalisation d'une variable aléatoire représentant ce tag. Les intérêts de chaque utilisateur sont alors autant de réalisations indépendantes des m variables représentant les m tags possibles. L'objectif de

l'ACP est de trouver des combinaisons linéaires des variables représentant les tags pour expliquer au mieux les intérêts des utilisateurs. Ainsi, à chaque utilisateur u_i, nous associons le vecteur de ses degrés d'appartenance à chaque tag, $X_i = (x_{i1}, x_{i2}, ..., x_{im})$. Ce vecteur représente le positionnement de l'utilisateur dans l'espace des tags, et l'ensemble des vecteurs X_i donne ainsi un nuage de points dans l'espace des tags. De la même manière, on peut associer à chaque tag t_j le vecteur V_j, correspondant à ses degrés d'appartenance chez les n utilisateurs : $V_j = (x_{1j}, x_{2j}, .., x_{ij}..., x_{nj})$. Ces nuages de points sont difficiles à analyser, à cause des dimensions considérées (nombre de tags, nombre d'utilisateurs) et de la variabilité des observations. L'analyse en composantes principales va alors :

1. Permettre une projection du nuage de points utilisateurs (initialement exprimés dans un espace de dimension k) sur des plans principaux (de dimension 2) qui reconstituent au mieux la variabilité entre les utilisateurs.
2. Permettre une représentation des variables initiales dans ces plans principaux, la contribution des variables dans la construction des axes principaux n'étant pas la même pour toutes les variables. Par exemple, la figure 1 donne une représentation compacte des rassemblements de tags selon leurs usages.

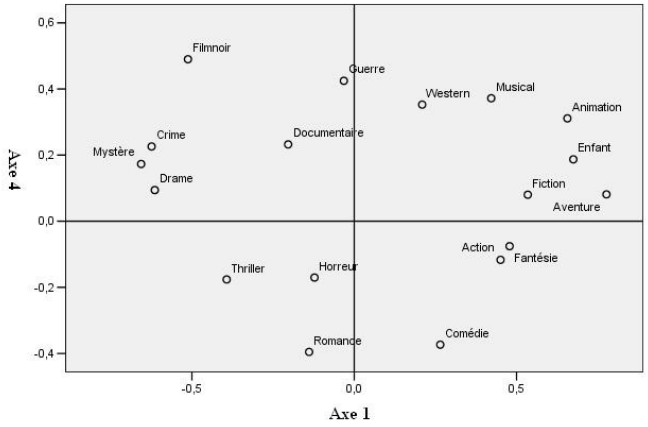

FIG. 1 – *Projection des variables sur deux axes*

Ainsi, des axes explicatifs sont identifiés, en minimisant la perte d'information effectuée lors de cette simplification. La figure 1 représente les variables originales de nos expérimentations sur deux axes significatifs, appelés *composantes principales* (dans cette figure, nommés axes 1 et 4). Cette figure présente la corrélation des variables d'origine avec les composantes principales (une variable est bien représentée sur l'axe si sa corrélation avec la composante principale correspondante est en valeur absolue proche de 1). Selon la composante 1 (Axe 1), on voit que les tags `Animation` et `Enfant` sont très corrélés (corrélation supérieure à 0,6). De même, la composante 4 oppose les tags `Filmnoir`, `Guerre` aux tags `Romance`, `Comédie`.

Notre méthode de rassemblement de tags est alors la suivante : l'ACP fournit les composantes principales pertinentes pour l'analyse des usages. Selon chacune de ces composantes, on ignore les tags situés dans la zone de faible corrélation (corrélation entre $-\alpha$ et $+\alpha$, pour un seuil $\alpha \in]0, 1]$ fixé). Les tags restants, situés dans les zones de forte corrélation (inférieure à $-\alpha$ ou supérieure à $+\alpha$), sont rassemblés dans une même communauté de tags. Par exemple, `Animation` et `Enfant` seront dans une même communauté. L'algorithme 1 résume la méthode.

Algorithme 1 : Découverte

entrées : Vecteurs V_j, seuil de décision α
sorties : Communautés de tags $G_1, \ldots, G_K$
début
  identifier les composantes principales $C = ((c_1, c_2), (c_3, c_4)...)$, expliquant la plus grande proportion de la variabilité des données
  tant que *(il reste des composantes principales* (c, c') *dans* C*)* **faire**
    ignorer les tags non corrélés (|coordonnées selon c et c'| $< \alpha$)
    rassembler dans une même communauté les tags corrélés selon c (|coordonnées selon c| $> \alpha$)
    rassembler dans une autre communauté les tags corrélés selon c' (|coordonnées selon c'| $> \alpha$)
    supprimer ces tags
  fin tant que
fin

Communautés d'utilisateurs Une fois l'ensemble des tags T décomposé en K communautés de tags $G_1, \ldots, G_K$, on en déduit les communautés d'utilisateurs. Pour cela, pour un utilisateur u_i donné, on calcule son degré d'appartenance x'_{ij} à chaque communauté de tag G_j :

$$x'_{ij} = \sum_{t_k \in G_j} x_{ik}.$$

Sa communauté $c(u_i)$ est alors sa communauté de tag majoritaire, c'est à dire l'indice j tel que x'_{ij} soit maximal. Chaque utilisateur est alors associé à ce groupe de tags. Ce groupe aura comme intitulé l'ensemble des tags qui le constituent.

3 Expérimentation

Contexte Nous avons testé la méthode sur la base de films MovieLens[4]. Cette base contient 100 000 votes pour 1 682 films appréciés par 943 utilisateurs. Les films sont évalués par une note entre 1 et 5. Nous avons remplacé ces notes par un vote binaire (les notes supérieures à 2 indiquant un intérêt pour le film). Nous avons construit la matrice M avec l'ensemble des

[4]http ://www.grouplens.org/node/73

utilisateurs U et l'ensemble des films R, et calculé le degré d'appartenance des utilisateurs aux différents tags. Nous présentons les résultats de notre approche sur un ensemble de 18 tags (1 : Aventure, 2 : Enfant, 3 : Animation, 4 : Mystère, 5 : Crime, 6 : Drame, 7 : Fiction, 8 : Filmnoir , 9 : Fantasy, 10 : Musical, 11 : Action, 12 : Thriller, 13 : Romance, 14 : Comédie, 15 : Horreur, 16 : Guerre, 17 : Documentaire, 18 : Western). Le seuil de décision α a été fixé à 0,6 de façon empirique (la sélection automatique de ce seuil n'a pas pu être abordée dans le cadre de ce premier travail.)

Matrice de corrélation La première étape de l'analyse est de vérifier que les données sont factorisables, c'est-à-dire qu'elles sont corrélées entre elles. Pour cela, on examine la matrice de corrélation :

- Si les coefficients de corrélation entre variables sont faibles, il est improbable d'identifier des facteurs communs. On peut éventuellement supprimer les variables qui ont une corrélation faible.
- Un autre paramètre pouvant aider au choix des variables est la qualité de la représentation (*Communalities*) ; QLT_j est le cosinus carré de l'angle formé entre la variable initiale x_j et l'axe principal c.

Le tableau de la figure 2 représente la matrice de corrélation entre une partie des variables initiales et les 6 premières composantes principales.

Tag	1	2	3	4	5	6
Aventure	**,777**	,349	-,272	,081	,037	-,056
Enfant	**,675**	-,231	*,465*	,187	-,145	-,147
Animation	**,657**	-,200	*,391*	,311	-,052	-,218
Mystère	**-,657**	,258	*,367*	,173	-,254	-,057
Crime	**-,624**	,265	,094	,226	,237	-,016
Drame	**-,614**	-,561	-,230	,094	,016	-,112
Fiction	,535	,531	-,252	,080	,249	-,152
Filmnoir	-,512	,066	,209	**,490**	,083	,158
Fantasy	,479	-,108	,197	-,076	,208	-,022
Musical	,422	-,409	,380	,372	-,193	,096
Action	,451	**,746**	-,262	-,117	-,128	,028
Thriller	-,393	**,704**	,314	-,176	-,221	,011
Romance	-,139	**-,685**	-,221	**-,395**	-,231	-,023
Comédie	,265	**-,592**	,161	**-,373**	,225	,242
Horreur	-,122	,424	*,360*	-,170	,369	,179
Guerre	-,032	-,037	**-,633**	**,425**	-,331	-,103
Documentaire	-,204	-,263	-,166	,232	**,639**	-,400
Western	,209	-,105	-,262	,353	,142	**,780**

FIG. 2 – *Corrélation entre les variables et les composantes*

La qualité de la représentation de la variable Action, par exemple, est obtenue en élevant au carré les coefficients de corrélation entre cette variable et les 6 axes principaux, puis en les

sommant :

$$QLT_{\texttt{Action}} = (0,451)^2+(0,746)^2+(0,262)^2+(0,117)^2+(0,128)^2+(0,028)^2 = 0,859.$$

Ainsi pour chaque variable initiale, nous obtenons la variance prise en compte par l'ensemble des facteurs extraits. Plus cette valeur est proche de 1, plus l'ensemble de l'information contenue dans la variable est prise en compte. Il serait par exemple possible de négliger la variable correspondant au tag `Fantasy` en raison de sa faible qualité de représentation (nous l'avons cependant conservée lors de nos expérimentations).

Sélection des composantes principales La deuxième étape consiste à déterminer le nombre de facteurs à retenir. On tient compte :

- des facteurs qui permettent d'extraire une quantité d'information (valeur propre) > 1. Quand on a beaucoup de variables, il y a un grand nombre de facteurs pour lesquels la valeur propre est supérieure à 1. Dans ce cas, on retient beaucoup de facteurs et l'interprétation devient difficile.
- de la distribution des valeurs propres : utilisation du graphique des valeurs propres.

La figure 3 représente la variance expliquée par chaque composante principale (valeur propre). Pour savoir combien de composantes principales utiliser, on recherche une rupture de pente sur le graphique. Cette rupture signifie que l'on passe d'un facteur représentant beaucoup d'information à un facteur en représentant moins. On s'arrête au facteur précédant cette rupture de pente. Dans notre expérimentation, on retient les 6 premières composantes dont la valeur propre est supérieure à 1. Le pourcentage de variance expliquée est de 70%.

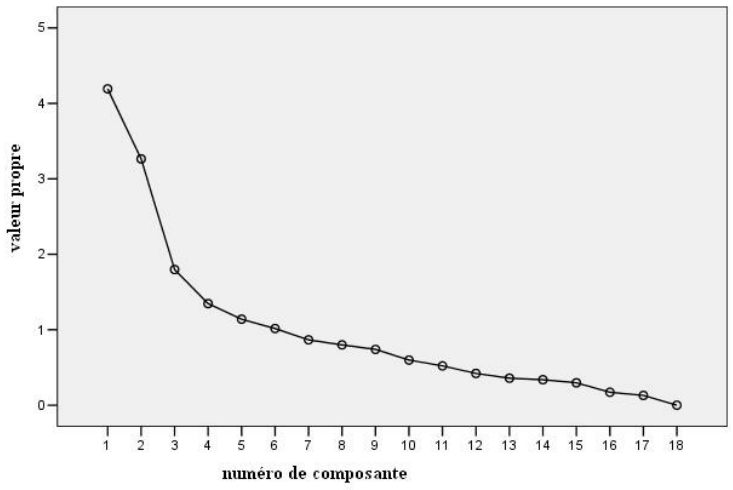

FIG. 3 – *Variance expliquée par chaque composante principale*

Les composantes obtenues ont la structure suivante :

- La 1re composante principale est la combinaison qui totalise la plus grande quantité de variance,
- La 2e composante principale est la combinaison qui totalise la 2ème plus grande quantité de variance. On peut déterminer autant de composantes principales qu'il existe de variables. La valeur propre de la 1re composante principale est 4,192 (soit 23,29% de la variance), celle de la 2e composante est 3,264 (soit 18,13% de la variance),etc. Les composantes principales sont indépendantes les unes des autres.

À partir de la matrice de corrélation, on voit que :

- La 1re composante principale représente essentiellement les variables `Aventure`, `Enfant`, `Animation`, `Mystère`, `Crime` et `Drame`.
- La 2e composante principale représente essentiellement les variables `Action`, `Thriller`, `Romance` et `Comédie`.
- La 3e composante principale représente essentiellement la variable `Guerre` et à un moindre degré les variables `Enfant`, `Animation`, `Mystère` et `Horreur`.
- La 4e composante principale représente essentiellement les variables `Filmnoir`,`Guerre` d'une part, et `Romance`, `Comédie` d'autre part.
- La 5e composante principale représente essentiellement la variable `Documentaire`.
- La 6e composante principale représente essentiellement la variable `Western`.

Interprétation des axes La dernière étape de l'expérimentation est l'interprétation des axes. on donne un sens à un axe à partir des coordonnées des variables. Ce sont les valeurs extrêmes qui concourent à l'élaboration des axes. Les facteurs avec de larges coefficients (en valeur absolue) pour une variable donnée indiquent que ces facteurs sont proches de cette variable. Nous rapprochons les tags par les degrés d'appartenance des utilisateurs à ces tags en nous basant sur les graphiques générés lors de cette étape :

- Le 1er axe (figure 4) oppose les tags `Animation`, `Enfant` et `Aventure` aux tags `Mystère`, `Crime` et `Drame`. Ceci correspond à une interprétation naturelle : les personnes qui aiment le premier groupe de films n'aimant en général pas le second. Deux communautés sont ainsi créées.
- Le 2e axe oppose les films de `Romance` et de `Comédie` aux films `Thriller` et `Action`, en créant ainsi deux nouvelles communautés.
- Le 3e axe (figure 5) oppose les films de `Guerre` aux films étiquetés `Enfant`, d'`Animation` ou de `Mystère`.

Les axes 4, 5 et 6 nous donnent les résultats suivants :

- Le 4e axe oppose les films `Filmnoir` et les films de `Guerre` aux films de `Romance` et de `Comédie`.
- Le 5e axe oppose les films `Documentaire` aux films de `Guerre`.
- Le 6e axe oppose les films `Western` aux films `Documentaire`.

Cette interprétation nous donne 7 groupes de tags, comme indiqué au tableau 1. Les groupes qui sont disjoints sont 1 et 2, 3 et 4, 4 et 6 et enfin 6 et 7. Les utilisateurs sont regroupés en fonction de ces communautés de tags. Les tags qui ne sont pas pris en compte par les axes sont expliqués par leur faible occurrence : par exemple le tag `Fantasy` n'est utilisé que 22 fois sur toute la collection des 1682 films.

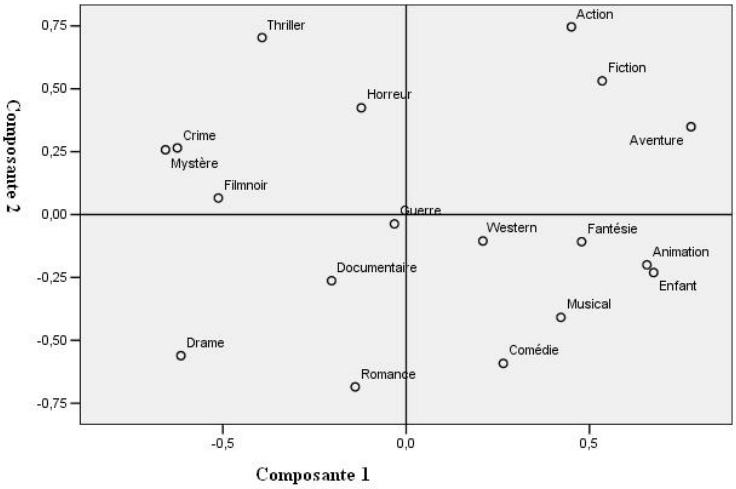

FIG. 4 – *Composantes 1 et 2*

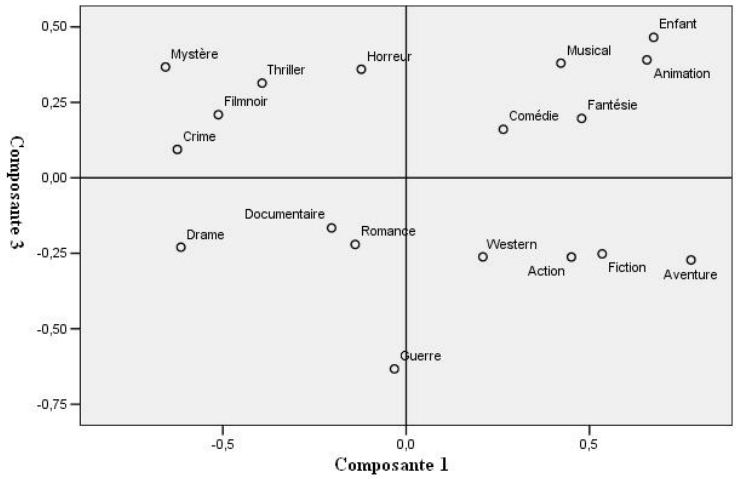

FIG. 5 – *Composantes 1 et 3*

communauté	tags associés
1	Aventure, Enfant, Animation
2	Mystère, Crime, Drame
3	Action, Thriller
4	Romance, Comédie
5	Western
6	Filmnoir, Guerre
7	Documentaire

TAB. 1 – *Communautés de tags*

4 État de l'art

Depuis les débuts du Web jusqu'à aujourd'hui, la recherche de communautés implicites a fortement évolué. De nombreux travaux, envisagent divers aspects des réseaux sociaux et des communautés, selon que l'on considère une communauté comme un ensemble de documents relatif à une thématique, ou comme un ensemble de personnes partageant un intérêt pour une thématique.

Découverte des communautés Web Dès les premiers travaux sur la reconnaissance des communautés sur le Web (par exemple Gibson et al. (1998)), le lien hypertexte est utilisé comme base de raisonnement. L'apport majeur en la matière est l'algorithme HITS de Kleinberg (1998), définissant les notions d'autorités et de *hubs*, structurant une communauté autour d'un sujet donné. Imafuji et Kitsuregawa (2002) concluent à l'appartenance d'une page à une communauté si cette page est plus majoritairement référencée depuis l'intérieur de la communauté que depuis son extérieur. Ils utilisent un algorithme de flot maximum afin d'isoler les noeuds faisant partie d'une même communauté, en se basant sur l'algorithme proposé par Flake et al. (2000). Dourisboure et al. (2007) identifient au sein d'un graphe du Web les communautés comme autant de sous-graphes denses et bipartis au sein de ce graphe. Le graphe biparti représente d'une part les centres d'intérêt de la communauté (les autorités selon HITS) et d'autre part ceux qui citent la communauté (les *hubs*). Cette méthode permet de mettre en évidence les éventuels partages des mêmes centres d'intérêt par plusieurs communautés d'acteurs, ou au contraire le partage de mêmes acteurs par plusieurs centres d'intérêt des communautés. Ces approches fournissent une analyse avancée des liaisons entre les différentes pages structurant une communauté thématique, mais ne permettent pas en revanche de rapprocher des utilisateurs de par leurs intérêts ou activités : le partage de lien hypertexte n'étant plus nécessairement la base de l'activité communautaire dans les échanges sociaux du Web collaboratif (évaluation de contenu par l'utilisateur, apposition de tags, ...).

Interprétations des tags utilisateurs Les systèmes de recommandations proposent à l'utilisateur un lot de ressources en corrélation avec son profil ou son activité. Firan et al. (2007) proposent un algorithme de recommandation basés sur les tags des utilisateurs. Ils prennent pour

exemple l'utilisation des tags sur le site de musique Last.fm[5], 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 utilisateurs et le contenu des données, en utilisant un réseau bayésien pour améliorer les méthodes classiques. Permettant un positionnement pertinent de l'utilisateur par rapport aux tags du système, ces solutions ne permettent pas de tenir compte des possibles similarités entre tags. La mise en lumière des tags similaires ou antagonistes que propose notre solution permet d'affiner ce positionnement de l'utilisateur.

Distances sémantiques Cattuto et al. (2008) présentent une autre approche statistique pour évaluer les distances sémantiques. Validée sur les données du site `del.icio.us`[6], site sur lequel il existe une structure communautaire, les auteurs utilisent l'annotation des données pour construire un réseau pondéré de ressources. Dans ce contexte, la similarité entre les ressources est proportionnelle au chevauchement de leurs jeux de tags. Pour prendre en compte la représentativité des tags, la méthode TF-IDF est utilisée. Les auteurs proposent de détecter les communautés d'utilisateurs par les similarités de leurs tags. Ils utilisent le coefficient de corrélation de Pearson comme mesure de similarité, puis appliquent des méthodes de partitionnement. À la différence de notre méthode, ils ne réduisent pas le nombre de tags manipulés, qui risque d'être extrêmement grand.

Systèmes de recommandation Le rapprochement de tag est également abordé dans les systèmes de recommandation. Dans leur définition du système *Socialranking*, Zanardi et Capra (2008) procèdent à un enrichissement de requête basé notamment sur la similarité des tags, fondée sur leurs apparitions communes sur des ressources différentes. Une autre approche, proposée par Hotho et al. (2006) sous le nom *FolkRank* et utilisant à nouveau de la théorie des graphes, consiste à utiliser *PageRank* pour modéliser les relations entres les ressources, les utilisateurs et les tags. Cette approche, qui permet d'exploiter d'avantage les relations éparses, est également explorée par Bertier et al. (2009) : dans le cadre de *Gossple*, les auteurs utilisent la probabilité de passer d'un tag à un autre comme indicateur de leur similarité.

5 Conclusion et perspectives

Dans cet article, nous avons présenté une méthode de découvertes de communautés d'utilisateurs par observation des usages, basée sur la technique de l'ACP. Une prochaine étape consiste en l'automatisation complète de la méthode, en particulier par l'estimation fine et automatique des seuils de sélection à utiliser, ainsi que la comparaison avec d'autres méthodes statistiques.

[5] http ://www.lastfm.com
[6] http ://delicious.com

Remerciements

Ce travail est partiellement financé par l'ANR Contenu & Interaction Neuma 2008-2011[7] et le projet CheckSem[8].

Références

Bertier, M., R. Guerraoui, V. Leroy, et A.-M. Kermarrec (2009). Toward personalized query expansion. In *SNS '09 : Proceedings of the Second ACM EuroSys Workshop on Social Network Systems*, New York, NY, USA, pp. 7–12. ACM.

Cattuto, C., A. Baldassarri, V. D. P. Servedio, et V. Loreto (2008). Emergent community structure in social tagging systems. *Advances in Complex Systems (ACS) 11*(04), 597–608.

Dourisboure, Y., F. Geraci, et M. Pellegrini (2007). Extraction and classification of dense communities in the Web. In *WWW'07 : Proceedings of the 16th international conference on World Wide Web*, New York, NY, USA, pp. 461–470. ACM.

Falissard, B. (2005). *Comprendre et utiliser les statistiques dans les sciences de la vie*. Masson, Paris.

Firan, C. S., W. Nejdl, et R. Paiu (2007). The benefit of using tag-based profiles. In *LA-WEB '07 : Proceedings of the 2007 Latin American Web Conference*, Washington, DC, USA, pp. 32–41. IEEE Computer Society.

Flake, G. W., S. Lawrence, et C. L. Giles (2000). Efficient identification of Web communities. In *KDD'00 : Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, pp. 150–160. ACM.

Gibson, D., J. Kleinberg, et P. Raghavan (1998). Inferring Web communities from link topology. In *HYPERTEXT'98 : Proceedings of the ninth ACM conference on Hypertext and hypermedia : links, objects, time and space—structure in hypermedia systems*, New York, NY, USA, pp. 225–234. ACM.

Hotho, A., R. Jäschke, C. Schmitz, et G. Stumme (2006). Folkrank : A ranking algorithm for folksonomies. In *Proc. FGIR 2006*.

Imafuji, N. et M. Kitsuregawa (2002). Effects of maximum flow algorithm on identifying Web community. In *WIDM'02 : Proceedings of the 4th international workshop on Web information and data management*, New York, NY, USA, pp. 43–48. ACM.

Kleinberg, J. M. (1998). Authoritative sources in a hyperlinked environment. In *SODA'98 : Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, Philadelphia, PA, USA, pp. 668–677. Society for Industrial and Applied Mathematics.

Yoshii, K., M. Goto, K. Komatani, T. Ogata, et H. G. Okuno (2006). Hybrid collaborative and content-based music recommendation using probabilistic model with latent user preferences. In *ISMIR'06 : 7th International Conference on Music Information Retrieval*, pp. 296–301.

[7] http ://neuma.irpmf-cnrs.fr
[8] http ://iutdijon.u-bourgogne.fr/checksem

Zanardi, V. et L. Capra (2008). Social ranking : uncovering relevant content using tag-based recommender systems. In *RecSys '08 : Proceedings of the 2008 ACM conference on Recommender systems*, New York, NY, USA, pp. 51–58. ACM.

Summary

Most of the existing social network systems require from their users an explicit statement of their friendship relations. In this paper we focus on implicit communities of users and present an approach to automatically detect communities of Web users, based on user's resource manipulations. Our proposal relies on the Principal component analysis (PCA) method and is assessed on a large movie data set.

Analyse statique et sémantique de réseaux sociaux d'entreprises et institutions : vers un modèle multidimensionnel convergent

Christophe Thovex*
Francky Trichet*

*LINA - Laboratoire d'Informatique de Nantes Atlantique (UMR-CNRS 6241)
Université de Nantes, équipe COD - Connaissance & Décisions
2 rue de la Houssinière BP 92208 - 44322 Nantes Cedex 03
christophe.thovex@univ-nantes.fr
francky.trichet@univ-nantes.fr,

Résumé. Les réseaux sociaux du Web 2.0 sont devenus planétaires, comme en témoignent FaceBook et MSN fédérant chacun 3.6% de la population mondiale. Dès 1989, L. C. FREEMAN publiait les premières métriques d'Analyse de Réseaux Sociaux (ARS), principalement basées sur des modèles de fouille de graphes. Nos travaux visent à faire converger ces modèles d'analyse statique, étendus par de multiples contributions, avec les aspects conceptuels de graphes sociaux d'entreprises et d'institutions. Ces aspects conceptuels constituent des ontologies découvertes dans les informations endogènes, connexes aux réseaux sociaux étudiés et orientées métiers. Cette approche originale et multidisciplinaire vise à découvrir de nouvelles mesures multidimensionnelles en ARS, pour de nouvelles fonctions décisionnelles en gestion de ressources humaines. Elle s'inscrit, en partenariat avec un éditeur logiciel leader de la gestion de capital humain et social, dans le cadre du projet SOCIOPRISE retenu par le Secrétariat d'Etat à la prospective et au développement de l'économie numérique.

1 Introduction

Nos tendances et besoins en communication appellent en permanence de nouvelles fonctions et applications sur les réseaux sociaux, comme en témoignent les éruptions constantes de nouveaux modes de socialisation tels que Tweeter pour le partage instantané d'informations brèves, Facebook pour le partage d'informations personnelles ou Diigo pour le partage de signets. Ces espaces d'échanges virtuels, à l'avantage des espaces d'échanges réels, facilitent l'analyse statistique et l'apparition de métriques et méthodes d'Analyse de Réseaux Sociaux (ARS, SNA - *Social Networks Analysis*). Les mesures de *centralité* introduites par L. C. FREEMAN sont régulièrement citées et reprises en matière d'ARS. Naturellement, l'ARS s'étend peu à peu aux entreprises - ARSE - pour fournir de nouveaux outils d'organisation du travail et de gestion des ressources humaines. La culture du travail collaboratif est de plus en

Atelier Web Social 2010 - En conjonction avec EGC 2010 (Hammamet - Tunisie)
H. Hacid (Alcatel-Lucent Bell Labs) et C. Favre (Lab. ERIC, Université de Lyon II)

plus couplée aux outils de type Web 2.0, caractérisant une forme d'entreprise "2.0" sensibilisée à la gestion du capital humain et social [Jean et Rallet (2001)].

Un réseau social peut être formalisé à l'aide d'un graphe, pouvant ou non être orienté, valué et/ou pondéré. À partir d'une telle structure, deux formes d'ARS peuvent être différenciées : l'*ARS statique* et l'*ARS sémantique*.

L'*ARS statique* étudie l'état *E* de graphes sociaux à un instant *t*. Elle est fondée sur des modèles et mesures de structures [1] et de flux [2], dans des graphes probabilistes dits aléatoires [Erdõs et Rényi (1959)], pseudo-aléatoires [Krivelevich et Sudakov (2002)], libres d'échelle [Barabasi et Albert (1999)] ou quelconques. L'*ARS statique* permet la classification de groupes d'individus ou communautés par le calcul de *degrés*, *connectivités*, *distances* et *flux* [3] et la découverte de relations implicites entre individus au sein du graphe social.

L'*ARS sémantique* étudie les aspects conceptuels des graphes sociaux. Elle est fondée sur les principes initiés par les travaux de J. SOWA sur les graphes conceptuels et les réseaux sémantiques [Sowa (2000)]. Elle se réfère également au Web sémantique, à l'ingénierie des ontologies [Gruber (1995)] et aux inférences logiques, en corrélation avec les sciences cognitives - *cf.* Manine (2009), Aimé et al. (2009), Gruber (2008) - ou langages du Web sémantique [4]. Avec la croissance exponentielle des réseaux sociaux et flux d'information, l'ARS sémantique devient cruciale pour la découverte et la gestion de connaissances, du contenu d'entreprise aux grandes communautés du Web. L'ARS sémantique peut notamment apporter de réels avantages en matière de gestion du capital humain et social ou d'optimisation des groupes et méthodes de travail au sein d'organisations professionnelles (sociétés, institutions).

À l'heure actuelle, très peu de travaux visent à intégrer les deux formes d'analyse différenciées. L'objectif de nos travaux consiste à répondre à ce manque en définissant un système convergent de modèles statistiques et conceptuels intégrant l'analyse statique et l'analyse sémantique de réseaux sociaux d'entreprises et d'institutions. L'approche adoptée est pluridisciplinaire car basée sur des principes électriques et des théories de sciences cognitives. Elle conduit à la définition d'un modèle multidimensionnel permettant le développement de nouveaux outils décisionnels pour l'optimisation du travail et de la gestion du capital humain et social. Dans l'état actuel de nos travaux, ce modèle inclut la définition de trois nouvelles mesures : (1) une mesure de *tension* d'un réseau social, (2) une extension de la mesure d'intermédiarité de L.C. FREEMAN baptisée *intermédiarité sémantique* et (3) une mesure de *réactance* d'un réseau social permettant l'évaluation du *stress* individuel des membres de ce dernier.

Ces travaux s'inscrivent dans le cadre du projet SOCIOPRISE retenu par le Secrétariat d'État à la prospective et au développement de l'économie numérique, dans le cadre de l'appel à projets "Web innovant" inscrit au plan de relance. SOCIOPRISE est mené en partenariat avec la société OpenPortal Software (http ://www.openportal.fr), éditeur logiciel de solutions pour la gestion du capital humain.

La suite de cet article est structurée comme suit. La section 1 introduit de façon synthétique les principes et méthodes respectivement utilisés pour l'ARS statique et l'ARS sémantique. La

1. Modèles et mesures structurels comme dans Freeman (1977), Burt (1995), Lazega (2001).
2. Modèles et mesures de flux comme dans Latora et Marchiori (2001), Thomassen (1990).
3. Le nombre d'arêtes connectées à un sommet est le *degré* du sommet. Le nombre d'autres sommets accessibles depuis un sommet donne sa *connectivité*. La *distance* entre deux sommets est le nombre d'arcs les séparant. Un *flux* élémentaire est caractérisé par un nombre d'unités circulant entre deux sommets - *e.g.* réseaux hydrauliques, électriques, routiers, etc.
4. Des langages basés XML, standards W3C - *i.e.* OWL, RIF, FOAF, SIOC, MOAT, etc..

section 2 présente en détails les contributions apportées à la problématique de convergence entre les deux types d'analyse. Ces contributions sont basées sur (1) un rapprochement entre principes électroniques et mesures d'analyse statique, puis (2) un rapprochement entre les nouvelles mesures d'analyse statique, définies en (1), et l'ingénierie des connaissances. Nos travaux sont dédiés à l'Analyse de Réseaux Sociaux d'Entreprises et d'Institutions - ARSEI.

2 ARS : état et approches unidimensionnelles

2.1 Analyse statique

L'analyse statique des réseaux sociaux étudie l'état *E* de graphes sociaux à un instant *t*, *E* étant défini par les structures et/ou les flux des graphes étudiés. Les premières notions d'ARS publiées dans Freeman et al. (1960) portaient sur le leadership dans les communautés. Ces notions se sont enrichies autour des mesures de centralité et d'intermédiarité [Freeman (1977)] dans les graphes sociaux, caractérisant les propriétés de réseaux sociaux en terme de *pouvoir*, *prestige*, *proximité* ou *confiance*.

Les mesures de centralité sont basées sur la comparaison du degré d'un sommet à ceux du graphe, voisins ou distants. Un sommet connecté (directement ou non) à un grand nombre de sommets du graphe porte une *centralité de pouvoir* importante. Un sommet connecté aux sommets à forts degrés du graphe social porte un coefficient de *centralité de prestige* élevé. Un sommet connecté à un grand nombre de sommets voisins ou proches possède une *centralité de proximité* élevée. Par induction, centralité de prestige et de proximité importantes pour un même sommet peuvent révéler un coefficient de *confiance* significatif.

Une mesure d'intermédiarité définit l'importance d'un individu pour l'interconnection de ses proches. Elle est formalisée par Freeman (1977) comme suit :

$$Iu = \sum_{ij} \frac{\sigma(i,u,j)}{\sigma(i,j)} \tag{1}$$

où *σ(i, u, j)* est le nombre de plus courts chemins entre les sommets *i* et *j* passant par *u* et *σ(i, j)* le total des plus courts chemins entre *i* et *j*, en somme pour tous les couples (*i, j*) du graphe [5].

2.1.1 Analyse statique structurelle

Classification (*graph-clustering*) et caractérisation de graphes fondent les principes de l'ARS statique. Les propriétés structurelles sont définies pour les principaux types de graphes sociaux et fournissent des éléments d'ARS statique. Dans les graphes aléatoires [Erdõs et Rényi (1959)], le degré des n sommets du graphe est déterminé par une probabilité $p(n)$ avec $p \mapsto [0;1]$. Avec les graphes pseudo-aléatoires, le degré des n sommets est distribué suivant une loi uniforme [6] où $G(V,p)$ possède une densité d'arêtes $p = \mid E \mid \div(\frac{|V|}{2})$. Pour les graphes libres d'échelle [Barabasi et Albert (1999)], les nœuds les plus connectés accroissent leur degré

5. En limitant la somme aux couples (*i, j*) pour lesquels *σ(i, u, j)* est supérieur à 0, il est possible de définir une mesure approximative adaptée aux grands graphes sociaux (*large social networks*).

6. Loi de distribution raffinable - *e.g.* loi de Gauss

de connexion suivant une loi de puissance (*"richers get richer"*). En déterminant des comportements caractéristiques à chaque type de réseaux et sous-graphes, ces propriétés structurelles statiques apportent aussi des éléments d'analyse dynamique des graphes sociaux.

2.1.2 Analyse statique de flux

Les travaux sur les arêtes en théorie des graphes (*e.g.* recherche de flux maximum), sont applicables à l'analyse statique de flux dans les réseaux sociaux, pour certains avec des résultats intéressants. C'est le cas de l'étude du "petit monde" (*small world*) dans laquelle V. LATORA et M. MARCHIORI ont introduit la notion d'*efficacité* (*efficiency*), définie comme une mesure de communication pondérée inversement proportionnelle au plus court chemin entre deux sommets *i* et *j* [Latora et Marchiori (2001)]. Notons ici les travaux de J. LESKOVEC et E. HORVITZ sur les grands graphes sociaux (MSN - 179 millions de sommets) ayant réactualisé la théorie des six degrés de séparation caractéristique du *small world* [7].

Certains modèles physiques sont traités à l'aide de graphes pour la compréhension et la découverte de principes théoriques. Dans le domaine de l'électricité, les *loi des nœuds* et *loi des mailles - Lois de Kirchhoff* - en sont l'illustration la plus connue. Les travaux de Thomassen (1990) sur la résistance et les courants des réseaux infinis, en démontrant l'unité et la continuité des flux dans les grands graphes, apportent une hypothèse à valider en ARS.

Pour résumer, l'ARS statique des réseaux sociaux offre donc un large ensemble de modèles mathématiques, sociologiques et même physiques, basés sur la théorie des graphes et utilisables pour la découverte de connaissances explicites ou implicites dans les graphes sociaux - structures et flux. Certains de ces modèles s'étendent également à l'analyse dynamique des réseaux sociaux - e.g. Zekri et Clerc (2002), que nous n'étudierons pas dans cet article.

2.2 Analyse sémantique

L'analyse sémantique des réseaux sociaux étudie les aspects conceptuels de graphes sociaux. Elle est fondée sur les graphes conceptuels et les ontologies conjuguées aux principes d'ARS [Gruber (2008)]. À l'heure actuelle, peu de travaux significatifs ont été publiés dans ce domaine. L'attrait pour le sujet est bien visible, néanmoins.

J. JUNG AND J. EUZENAT [Jung et Euzenat (2007)] commentent la description d'une vue tridimensionnelle sur l'analyse sémantique de réseaux sociaux, rapprochant graphes sociaux, annotations (*tags*) et ontologies *ERgraphs* (graphes entités-relations). La proposition superpose et fait coïncider les trois dimensions pour construire des ontologies [8] "consensuelles" dont les annotations sont associées au graphe social. Dans Aleman-Meza et al. (2006), ALEMAN-MEZA AND AL. décrivent une application sémantique de détection de conflits d'intérêts dans les réseaux sociaux de publications scientifiques. Basée sur la recherche de patrons syntaxico-sémantiques, l'application mesure la similarité sémantique entre corpus d'auteurs afin de détecter, dans les sujets répartis ou partagés entre plusieurs équipes, d'éventuels redondances et

7. Dans Leskovec et Horvitz (2008), la capacité d'atteindre en 6 sauts 80% des sommets d'un graphe est revue à la baisse avec seulement 48% des sommets atteints. Suivant une courbe *long-tail*, la distribution atteint 78% des sommets en 7 sauts et pour 90% des sommets, la moyenne mesurée est de 7,8 sauts (mesures effectuées sur un tirage aléatoire de 1000 sommets).

8. Une ontologie est une spécification explicite d'une conceptualisation. Elle représente les concepts, objets et autres entités supposés exister sur une aire d'intérêts avec leurs relations [Gruber (1995)].

concurrences. Les premiers travaux de Erétéo et al. (2009) sur l'ARS sémantique ouvrent la voie de l'analyse sémantique statistique et visent à rendre opérationnelles les grandes lignes de l'ARS en les intégrant aux ontologies et langages du Web sémantique (*i.e.* OWL, RIF, FOAF, SIOC, MOAT, POWDER).

Les systèmes de règles et d'inférences en corrélation aux sciences cognitives peuvent tracer un axe de développement à fort potentiel dans le domaine de l'ARS sémantique. Ce développement semble soumis à l'annotation des sommets et arêtes, par des moyens automatiques tels l'apprentissage statistique et le traitement automatique du langage naturel (TALN), ou des traitements humains comme "l'étiquetage social" ou *social tagging*. L'*évaluation réciproque* entre membres d'un réseau social est un exemple d'annotation où l'interaction humaine avec le système produit une valuation des sommets du graphe social sur laquelle un *degré de confiance* relativement fiable et précis peut être calculé [9]. Dans les ontologies appliquées en bactériologie, A-P. MANINE induit des règles sémantiques multiples et récursives extraites de l'information syntaxique par automatismes, puis infère ces règles sur l'ontologie pour l'enrichir automatiquement [10] [Manine (2009)]. L'hypothèse de méthodes dérivées applicables aux graphes sociaux est envisageable. Enfin, l'intégration de sciences cognitives, comme la linguistique, la psychologie ou les neurosciences, produit d'intéressants résultats en permettant, par exemple, la *pragmatisation* d'ontologies [Aimé et al. (2009)]. L'hypothèse de méthodes dérivées spécifiquement adaptées à l'analyse sémantique de réseaux sociaux peut être considérée.

T. GRUBER, l'un des précurseurs du Web Sémantique encourage les initiatives tendant à intégrer les principes et langages du web sémantique aux réseaux sociaux, pour le développement de systèmes d'intelligence et de connaissances collectives [11] [Gruber (2008)]. Des grandes communautés du Web aux réseaux sociaux d'entreprises, l'ARS sémantique peut apporter de réels progrès dans différents domaines tels le marketing global lié à la mondialisation, la gestion du capital humain et social ou l'optimisation des groupes et méthodes de travail au sein d'organisations professionnelles (sociétés, institutions).

3 Synergies multidimensionnelles en ARSEI

Nos travaux s'appliquent à la découverte de synergies multidimensionnelles entre les aspects statiques et sémantiques de l'Analyse de Réseaux Sociaux d'Entreprises et d'Institutions - ARSEI. Les spécificités de l'ARSEI sont : (1) graphe social de 100 000 nœuds maximum, (2) données endogènes restreintes à un ou quelques domaines de connaissances connexes et (3) adoption du principe d'échanges d'informations centrées métiers.

La méthodologie adoptée respecte la segmentation de la problématique :

– L'analyse statique de réseaux sociaux est intégrée comme telle, notre effort de recherche portant sur le rapprochement pertinent de méthodes connues et de modèles identifiés, issus de la physique ou de sciences cognitives. Les résultats proposés touchent de nouvelles mesures de flux sur les graphes sociaux. Ils sont dédiés à l'ARSEI et à la prévention du risque social. À l'heure actuelle, ces résultats consistent en la définition de 2 mesures. La première mesure est dédiée à l'évaluation d'une nouvelle notion baptisée *tension d'un réseau social* (cf. section 3.1.1). La seconde mesure reprend et étend

9. On parle de *réseau de faveurs* quand la structure de graphe dépend des évaluations entre pairs.
10. Utilisation de la Programmation Logique Inductive (PLI).
11. *Collective Intelligence, Collective Knowledge Systems*

l'*intermédiarité* de L.C. FREEMAN (cf. section 3.1.2) qui devient ainsi sémantique (*intermédiarité sémantique*).

– L'analyse sémantique est développée par l'étude de rapprochements entre les graphes sociaux, les graphes conceptuels, les ontologies et règles d'inférences et les sciences cognitives. Les résultats proposés s'appliquent à l'ARSEI et sont dédiés à l'organisation du travail et à la gestion de capital humain et social. À l'heure actuelle, ces résultats consistent en la définition d'une nouvelle mesure de *réactance* destinée à l'évaluation du stress individuel (cf. section 3.2).

Les résultats obtenus sont mis en commun pour converger en un modèle multidimensionnel, propice au développement et à la popularisation d'outils décisionnels pour les Réseaux Sociaux d'Entreprises et d'Institutions - RSEI.

3.1 ARSEI statique, modèles physiques et cognition

Notre modèle adopte les mesures de centralité et d'intermédiarité de FREEMAN sur les graphes non-orientés. Pour les graphes orientés, *Page-Rank* fournit un score assimilable à une mesure de *prestige*. Nous retenons, pour nos travaux, la version de Page et al. (1999) pour un graphe orienté G=(V,ε) de *N* sommets, un sommet (page) *q* référençant un sommet (page) *p* et $0 < \alpha < 1$. Le rang ou degré de centralité d'un sommet est fonction du nombre des sommets pointant dessus et de la somme de leurs rangs respectifs, dans un sous graphe de diamètre [12] fini et supérieur à 2 dont *p* est le centre. Ce score est ponderé par un coefficient dépendant de *N* et d'une constante modératrice α.

$$R(p) = \alpha \cdot \sum_{q:(q,p)\in\varepsilon} r(q)/\omega(q) + (1-\alpha)\cdot 1/N \qquad (2)$$

Une extrapolation intégrant un coefficient d'autorité (réputation de l'auteur), *Trust-Rank - TR*, donne un *score de confiance*, éventuellement adaptable aux graphes non orientés en complément d'autres mesures - *cf.* Gyongyi et al. (2004) .

3.1.1 ARSE statique, flux et modèles physiques

Afin d'introduire de nouvelles mesures de flux, nous éprouvons l'assimilation des arêtes du graphe social à des conducteurs transportant des flux électriques. Notre méthode consiste à quantifier et qualifier les flux par des ratios sémantiques afférents au réseau social d'entreprise, tels que pourcentages de documents communs consultés, rédigés ou échangés (bureautique, courriels, messages instantanés, etc.), d'échanges de paquets de données (ToIp, VoIp) ou d'autres types de communication pouvant caractériser les liens conceptuels entre individus. Des principes électriques sont adaptés à l'analyse statique de flux autour d'un sommet, dont les *lois des nœuds et des mailles* de KIRCHHOFF. La *loi des noeuds* est illustrée en Fig. 1, avec *I* intensité de charges électriques pour un débit de quantité *Q* par unité de temps *t*.

L'originalité de nos travaux consiste à introduire la notion de *tension* du réseau social en relation aux notions d'*intensité* de flux traversant et de *résistance* des sommets du graphe. Un sommet *s* directement connecté à deux autres sommets *r* et *t* est assimilé à un dipôle, de résistance *R*.

12. Le diamètre d'un graphe est le nombre minimum d'arcs reliant ses sommets les plus distants.

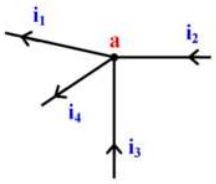

FIG. 1 – *Loi des nœuds :* $\sum Ientrant = \sum Isortant$ *, i2 + i3 = i1 + i4*

Nous utilisons les lois d'Ohm :
$Urt = Rs.Irt$ et $Ps = Rs.Irt^2 = Urt^2/Rs = Urt.Irt$,
où Urt représente la tension électrique dépendante de Rs et de Irt, et Ps représente la puissance délivrée par un sommet, de puissance maximum admissible notée $Pmax$, avec $Umax = \sqrt{R.Pmax}$ et $Imax = \sqrt{Pmax/R}$

En appliquant les lois d'Ohm sur un graphe social, il est possible de calculer un rapport de *charge-capacité* des composants du réseau social d'entreprise, par analogie aux rapports $Ps/Pmax$. L'objectif est de proposer une mesure de stress par individus et communautés. Cette mesure recourt à l'effet Joule pour estimer l'*échauffement* des composants du réseau social d'entreprise et prévenir les risques de dégradation des performances, d'instabilité ou de panne (trouble socio-psychologique). L'échauffement *T* dépend de l'énergie dissipée et de la résistivité ρ du matériau. La valeur de ρ variant selon la diversité des structures moléculaires, son calcul sort du cadre ce cet article. Il faut donc considérer à priori le *matériau social* abstrait comme une constante en initialisant les algorithmes avec $\rho = 1$, soit $T.\rho = W = R.I^2.\Delta t$. Ensuite, ρ pourra être raffiné par $\rho \mapsto [0; 1]$, selon un déterminant défini pour induire l'interaction récursive entre *T* et *R* rencontrée en physique, où ρ varie en fonction de *T*.

3.1.2 ARSEI statique et cognition

L'étiquetage manuel de ressources fait appel aux processus cognitifs. Cette méthode peut, notamment en ARSEI, provoquer un phénomène de rejet psychologique causé par des aspects politiques et éthiques, négativement perçus. L'étiquetage manuel doit être limité aux ressources non-humaines (documents, corpus textuel, base de données) pour être acceptable et la caractérisation des individus et groupes d'individus doit observer des critères respectant la personne et la vie privée.

La sémantisation des annotations, en associant les termes annotant les ressources métiers aux concepts d'une ontologie, permet la découverte de communautés d'usages par le biais des relations implicites entre ressources annotées. Dans cet optique, nous proposons d'utiliser les ontologies métiers pour qualifier l'analyse numérique de graphes sociaux en corrélant les résultats statistiques obtenus sur les flux et structures, aux concepts et graphes conceptuels ontologiques.

À partir de l'équation (1) , nous proposons une nouvelle mesure d'*intermédiarité sémantique* pondérée par des ressources endogènes (*i.e.* principalement des documents annotés à l'aide de termes) où chaque annotation est associée à au moins un individu du réseau social considéré et où la somme d'occurrences d'une annotation calibre la mesure favorablement pour

les individus partageant les ressources associées aux annotations majoritaires. De surcroit, en excluant l'étiquetage entre individus, la proposition respecte l'éthique professionnelle et réduit, faiblement, le risque d'atteinte à la vie privée.

Le cadre formel de cette nouvelle mesure est donné comme suit. Les relations explicites entre l'ensemble des ressources humaines *Rh*, celui des ressources du système d'information *Rsi* et l'étiquetage de contenu *Esi* servent à enrichir l'ARSEI et découvrir des relations implicites *R'*. Nous introduisons donc les ensembles Rh, Rsi, Esi et les relations R, R'. Nous évitons le calcul de relations réflexives inutilement couteuses - *e.g.* relations dans $RsiXRsi, EsiXEsi$ et pour respecter la confidentialité des annotateurs, il n'existe dans notre proposition aucune relation R entre D et D' avec $D = Rh$ et $D' = Esi$ ou $D = Esi$ et $D' = Rh$.

Nous définissons un domaine D, un co-domaine ou image D' et la relation $R(D, D')$:
$D = Rh$ ou $D = Rsi$, $D' = Rh$ ou $D' = Rsi$ ou $D' = Esi$ (note : Esi n'est jamais domaine de Rh).
Nous posons deux variables α, β et leurs contraintes :
- Contrainte 1 : $(\alpha \in Rh \vee \alpha \in Rsi) \wedge (\beta \in Rh \vee \beta \in Rsi \vee \beta \in Esi)$
- Contrainte 2 : si $\alpha \in Rsi \wedge \beta \in Rsi$, il n'existe pas de relation $R(\alpha, \beta)$
- Contrainte 3 : si $\alpha \in Rh \wedge \beta \in Esi$, il n'existe pas de relation $R(\alpha, \beta)$
La réciproque de ces contraintes s'exprime par :
$\forall \alpha \in Rsi \wedge (\beta \in Rh \vee \beta \in Esi) \vee \alpha \in Rh \wedge (\beta \in Rh \vee \beta \in Rsi)$, alors $\exists R(\alpha, \beta)$.

Nous introduisons un nouvel ensemble de mesures par adjonction d'un coefficient de pondération basé sur la cardinalité C, de R. R étant paramétrée par (pD, pD') désignant (D, D') et, facultativement, par eD, eD' restreignant (D, D'), C permet de déclarer toutes les formes acceptées de R entre Rh, Rsi, Esi comme facteurs dans les mesures de graphes sociaux. De plus, R pourra être composée en hiérarchie de relations sur le patron $R'(pD, pD') \rightarrow R(pD, pD')$ avec pD ou pD' communs à R et R' [13]. Nous exprimons ce facteur par :

$$C_p = \frac{1}{C_{R(pD,pD',eD,eD')}} \tag{3}$$

Par exemple, quand D ou D' = Rh, C est utilisable pour pondérer les mesures et équations décrites en sections 2.1,3.1 et 3.1.1. Cette méthode permet d'intégrer à l'analyse du graphe social, des statistiques extraites de données endogènes et sémantiquement connexes.
Avec $eD \in pD \wedge eD' \in pD'$, l'équation (1) sera modifiée comme suit :

$$Iu(C) = \sum_{ij} \frac{\sigma(i,u,j)}{\sigma(i,j)} \div \frac{1}{C_{R(pD,pD',eD,eD')}} \tag{4}$$

L'équation (4) livre une nouvelle mesure d'*intermédiarité sémantique*, basée sur celle de Freeman (1977) et apportant une dimension qualitative en intégrant les informations endogènes contenues dans Rsi et Esi au degré de centralité de l'individu u. C_p est quantifié et qualifié par eD, eD' (facultatifs) via les ontologies permettant l'association sémantique d'éléments de Rsi et Esi. Les connaissances découvertes dans ses associations conceptuelles sont le point fort de cette nouvelle mesure, qu'elles rendent "intelligente".

13. Ceci pourra être rapproché de l'étude des réseaux bayésiens.

3.2 Analyse sémantique de RSEI

Dans les sections 3.1.1 et 3.1.2, nous avons introduit une analogie entre l'analyse de flux et structures des réseaux sociaux, et quelques principes radioélectriques pouvant se révéler pertinents. Ont été avancées les notions de *résistance, charge, capacité, échauffement* et *puissances*. Ces concepts caractérisent des relations sémantiques $R_s(i,j)$, explicites ou implicites, entre les sommets d'un graphe social. Notre intention est de doter ces relations de propriétés sémantiques (propriétés objet ou de données), en les représentant par une ou plusieurs ontologies de domaine qui conceptualisent les interactions avec le graphe social.

La notion de réactance est présente en électrodynamique ou en psychologie sociale. En électrodynamique, la réactance (en Ohms) telle qu'utilisée dans Wang (2009) décrit traditionnellement l'énergie s'opposant à un courant alternatif [14], selon qu'il traverse un élément capacitif ou inductif [15]. En psychologie sociale, la réactance caractérise *un état de motivation négatif faisant suite à une menace (supposée réelle) d'une restriction de la liberté individuelle qui se traduit par une résistance à l'influence* [Brehm (1966)].

Dans le cadre de nos travaux, nous proposons d'utiliser la *réactance* Ψ comme mesure de stress individuel. À partir de la notion de *tension* définie en section 3.1, nous pouvons établir les assertions suivantes :

Soit un graphe $G(V,E)$ où des sommets V sont connectés par des arêtes E, muni des propriétés suivantes :

- Tout élément v de V porte intrinsèquement les coefficients issus de mesures classiques des réseaux sociaux (*cf.* Freeman, etc.) ou leurs raffinements éventuels.
- $\forall (u,v) \in V$ connecté par $e \in E$, u, v portent intrinsèquement des valeurs analogiques calculables dans $\mathbb{Q}$ de *résistance, charge, capacité, échauffement, puissances* dépendant de V, E.
- $\forall e \in E$ assimilé à un flux ϖ quelconque $\vec{\varpi}$ ou de valeur quantifiable $\varphi_\varpi \neq 0$, e porte intrinsèquement des valeurs analogiques calculables dans $\mathbb{Q}$ de *résistance, charge, capacité, échauffement, puissances*. Pour e, $\vec{\varpi}$ ou φ_ϖ sont mesurés en pseudo-tension T_e et pseudo-intensité I_e.

À partir de ces assertions et de nos premières expériences en ARSEI liées au projet SOCIOPRISE (*i.e.* un projet consacré à la gestion du capital humain et social au sein d'organisations métiers), nous proposons un premier ensemble de connaissances dédiées à l'identification du stress individuel. Ces connaissances sont exprimées à l'aide des règles [16] et axiomes [17] suivants :

* règle 1 : Si $CC_u = \frac{charge_u}{capacite_u}$ augmente et $CC_u < 80\%$, alors Ψ_u augmente [18].
* règle 2 : si $P_u = \frac{resistance_u . intensite_{(e1,u,e2)^2}}{Pmax_u}$ augmente et $P_u \leqslant 1$,
  alors Ψ_u et $echauffement_u$ augmentent (P_u représente une puissance utile).

14. Q. Wang utilise la réactance comme paramètre du réseau de neurones, pour contrôler les défauts du réseau électrique.

15. Dans un élément purement résistif, la réactance est nommée impédance et reste égale à la résistance mesurée pour un courant continu.

16. Une règle vérifie en toute circonstance la conclusion de prémisses données.

17. Un axiome affirme une vérité indéniable et indémontrable - *cf.* abduction, induction.

18. Par analogie aux réseaux électroniques de puissance, on intègre la notion de seuil de charge minimal sous lequel le rendement s'effondre.

* règle 2 *bis* (apprentissage par inférence sur règle 2) : si $echauffement_u$ augmente, alors Ψ_u augmente.
* règle 3 : si P_u augmente et $P_u > 1$, alors Ψ_u diminue, $Pmax_u$ diminue et $echauffement_u$ augmente rapidement (P_u a dépassé $Pmax_u$).
* règle 3 *bis* (apprentissage par inférence sur règle 3 et supervision de l'expertise) : si Ψ_u diminue et $echauffement_u$ augmente, alors diminution rapide de $Pmax_u$ et risque de destruction.
* axiome 1 (apprentissage par inférence supervisée sur règle 1) : si $CC_u \leqslant 0.8$, alors risque de pertes de performances socioprofessionnelles.
* axiome 2 (apprentissage par inférences sur règle 3 et 3 *bis*) : si $P_u > 1$, alors risque de troubles socioprofessionnels.
* axiome 3 (apprentissage supervisé sur axiomes 1 + 2 et leurs prémisses) : optimisation de la performance équivaut à $CC_u > 0.8$ et $P_u \leqslant 1$.
* axiome 4 (apprentissage par symétrie sur axiome 3 et ses prémisses) : risque de troubles socioprofessionnels équivaut à risque de pertes de performance économique.

In fine, nous prévoyons de formaliser une mesure scalaire de réactance Ψ_u par le système d'équations sous-jacent à ces règles et axiomes, mesure actuellement inexistante. L'intérêt de l'approche sémantique couplée aux modèles statistiques tient dans la découverte de connaissances conceptuelles implicites pour l'analyse statique des réseaux sociaux.

4 Conclusion

Nos travaux visent à définir un modèle à la fois statique et sémantique d'Analyse de Réseaux Sociaux d'Entreprises et d'Institutions (ARSEI). Leur originalité réside, d'une part, dans l'intégration au sein d'un même modèle des aspects statiques et sémantiques de l'ARSEI et, d'autre part, dans la définition de 3 mesures fondées sur des apports pluridisciplinaires. Ces nouvelles mesures sont respectivement dédiées à l'évaluation des notions de *tension*, d'*intermédiarité sémantique* et de *réactance* en ARSEI.

La *sémantisation* des mesures de FREEMAN, sur le modèle de l'intermédiarité sémantique, permet de qualifier les échanges collaboratifs quantifiés et d'établir au sein des RSEI de nouveaux degrés de centralité sur les individus, corrélant les dimensions statistiques et conceptuelles via les ressources endogènes et l'interdisciplinarité scientifique [19].

Ces travaux sont à la base du développement de nouveaux outils d'aide à la décision, pour la gestion du capital humain et social dans les entreprises et institutions. Plus particulièrement, ils permettent de répondre à des problématiques de prévention du risque de troubles socioprofessionnels, risque de perte de performance économique et risque social. D'un point de vue applicatif, ils sont en cours d'expérimentation dans le cadre du projet SOCIOPRISE [20]. D'un point de vue théorique, ils se poursuivent sur l'intégration d'un aspect dynamique de l'ARSEI.

19. Les raisonnements et procédures de construction de nos hypothèses appelant des processus cognitifs et notions de psychologie complexes, nous sortirions du cadre donné à cet article, comme des contraintes d'édition, en les détaillant.

20. Pour des raisons de maturité et de propriété intellectuelle du projet SOCIOPRISE, nous ne pouvons fournir, à ce stade, de retour d'expérience ou de complément d'information sur l'évaluation des résultats.

La prise en compte des travaux de KIRCHHOFF, d'AMPÈRE et de MAXWELL en électrodynamique ou de MARKOV en statistique est envisagée pour l'étude prédictive de l'évolution structurelle des réseaux sociaux. La perspective applicative de cette démarche est d'assister l'optimisation des groupes de travail et de la performance. La perspective théorique est de formaliser *un modèle complexe et multidimensionnel d'analyse statique, dynamique et sémantique d'analyse de réseaux sociaux d'entreprises et d'institutions*.

Références

Aimé, X., F. Furst, P. Kuntz, et F. Trichet (2009). Gradients de prototypicalité conceptuelle et lexicale : une contribution à la pragmatisation des ontologies de domaine. *Revue des Nouvelles Technologies de l'Information (RNTI) - Extraction et Gestion des Connaissances (EGC'08) 11-1*, 127–132.

Aleman-Meza, B., M. Nagarajan, C. Ramakrishnan, L. Ding, P. Kolari, A. Sheth, I. B. Arpinar, A. Joshi, et T. Finin (2006). Semantic analytics on social networks : experiences in addressing the problem of conflict of interest detection. In *WWW '06 : Proceedings of the 15th international conference on World Wide Web*, New York, NY, USA, pp. 407–416. ACM.

Barabasi, A.-L. et R. Albert (1999). Emergence of scaling in random networks. *Science Magazine Vol. 286*(no. 5439), pp. 509 – 512.

Brehm, J. (1966). *A Theory of Psychological Reactance*. Academic Press.

Burt, R. (1995). Le capital social, les trous structuraux et l'entrepreneur. *Revue Française de Sociologie 36*(4), 599–628.

Erdõs, P. et A. Rényi (1959). On random graphs. *Publicationes Mathematicae 6*, 290–297.

Erétéo, G., F. Gandon, J. De Santo, M. Buffa, et O. Corby (2009). Semantic social network analysis. In *Proceedings of the WebSci'09 : Society On-Line, 18-20 March 2009, Athens, Greece*.

Freeman, L. (1977). A set of measures of centrality based on betweenness. *Sociometry 40*, 35–41.

Freeman, L., W. Bloomberg, S. Koff, M. Sunshine, et T. Fararo (1960). *Local Community Leadership*. Syracuse.

Gruber, R. T. (2008). Collective knowledge systems : Where the social web meets the semantic web. *Web Semantics : Science, Services and Agents on the World Wide Web 6*(1), 4–13.

Gruber, T. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human Computer Studies 43*(5/6), 907–928.

Gyongyi, Z., H. Garcia-Molina, et J. Pedersen (2004). Combating web spam with trustrank. In *30th International Conference on Very Large Data Bases (VLDB 2004)*.

Jean, G. et D. Rallet (2001). Capital humain et capital social. *Nouveaux Regards (revue)* (14). `http://institut.fsu.fr/nvxregards/14/14_rallet_jean.htm`.

Jung, J. et J. Euzenat (2007). Towards semantic social networks. In *ESWC '07 : Proceedings of the 4th European conference on The Semantic Web*, Berlin, Heidelberg, pp. 267–280. Springer-Verlag.

Krivelevich, M. et B. Sudakov (2002). Sparse pseudo-random graphs are hamiltonian.

Latora, V. et M. Marchiori (2001). Efficient behavior of small-world networks. *Physical Review Letters 87*(19).

Lazega, E. (2001). *The Collegial Phenomenon : The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership*. Oxford.

Leskovec, J. et E. Horvitz (2008). Planetary-scale views on a large instant-messaging network. In *WWW 2008, April 21-25, 2008, Beijing, China*. ACM 978-1-60558-085.

Manine, A.-P. (2009). Acquisition de la théorie ontologique d'un système d'extraction d'information. In *Revue des Nouvelles Technologies de l'Information (RNTI E-15). Extraction et Gestion de Connaissances 2009 (EGC'09)*, Volume E-15, pp. 421–426. Editions Cépaduès. ISBN 978.2.85428.878.0.

Page, L., S. Brin, R. Motwani, et T. Winograd (1999). The pagerank citation ranking : Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

Sowa, J. (2000). *Knowledge Representation : Logical, Philosophical, and Computational Foundations*. Brooks Cole Publishing Co.

Thomassen, C. (1990). Resistances and currents in infinite electrical networks. *J. Comb. Theory Ser. B 49*(1), 87–102.

Wang, Q. (2009). Artificial neural network and hidden space svm for fault detection in power system. In *ISNN 2009 : Proceedings of the 6th International Symposium on Neural Networks*. Springer Verlag.

Zekri, N. et J.-P. Clerc (2002). Étude statistique et dynamique de la propagation d'épidémies dans un réseau de petit monde. *Comptes Rendus Physique 3*(6), 741–747.

Summary

Social networks of the Web 2.0 have become global, as FaceBook and MSN show, each one federating 3.6% of the world population. In 1989, L. C. FREEMAN published the first metrics for Social Networks Anaysis (SNA), mainly based on graph-mining models. Our works aim to make converge these models for static analysis, extended by multiple contributions, with the conceptual aspects of social graphs of enterprises and institutions. These conceptual aspects constitute ontologies found in the endogenous information, connate to the studied social networks and trades oriented. This original and multidisciplinary approach aims to discover new multidimensional measures in SNA for new decision-making functions in human resources management. Our approach, involved in a partnership with a leader company in human capital management software, is in line with the SOCIOPRISE project retained by the french State Secretariat at the prospective and development of the digital economy.

Vers une constitution automatique des réseaux d'entreprises collaboratifs à partir du web

Kafil Hajlaoui*, Xavier Boucher*
Michel Beigbeder*

*158 cours Fauriel, centre G2I
42023 Saint Etienne
{hajlaoui, boucher, mbeig}@emse.fr

Résumé. Dans cet article, nous proposons une approche de recherche et de traitement de l'information pour la constitution des réseaux coopératifs d'entreprises à partir de leur site web. Cette approche est basée sur une procédure automatique d'extraction d'information pour la génération des nouvelles connaissances. Ces connaissances sont les raisons qui peuvent et/ou qui doivent amener les entreprises à envisager des coopérations entre elles. Dans cet objectif, connaitre en temps réel et avec précision le secteur d'activité de l'entreprise est indispensable. L'approche abordée dans cet article repose sur des outils et des méthodes de recherche d'information, à savoir, l'indexation contrôlée et le matching. La bonne connaissance de cette activité permet de faire émerger des réseaux coopératifs d'entreprises de divers types.

1 Introduction

L'évolution de l'économie, la concurrence, la pression des donneurs d'ordre et l'impact des nouvelles technologies de l'information et de la communication (TIC) sont quelques unes des raisons qui amènent les entreprises à envisager des collaborations techniques et économiques. La collaboration inter-entreprises intervient lorsque plusieurs entreprises décident de mettre en commun des informations, des ressources ou des compétences dans la poursuite d'objectifs conjoints, qui pourront déboucher sur des activités coordonnées voire intégrées. Par exemple, deux entreprises peuvent collaborer parce que chacune possède une partie de l'information, de l'expertise et des ressources nécessaires à la mise au point d'un produit. Cet aspect collaboratif dans les réseaux des entreprises nécessite de mettre en place différentes architectures pour la gestion des processus de collaboration et différentes méthodes et outils d'aide à la décision stratégique pour l'entreprise. Le développement d'approches de type décisionnel requiert de déployer des solutions pertinentes de traitement de l'information, qui pourront devenir le support de processus de pilotage des activités et processus ou encore de processus de pilotage des systèmes de compétences.

Dans l'environnement économique moderne, caractérisé par des mutations incessantes, les entreprises sont appelées à être adaptatives, flexibles et proactives. Pour cela, elles construisent des espaces coopératifs dans lesquels elles travaillent et réagissent ensemble. Ces espaces coopératifs, appelés le plus souvent "nouvelles formes organisationnelles", ont émergé dans les années 80 sous diverses formes (réseau d'entreprises, entreprise virtuelle,

cluster, groupement de PME…). Une entreprise possède souvent des liens et des relations de différents types avec divers partenaires en fonction de ses objectifs, besoins et caractéristiques. Cette multiplicité et diversité des liens a amené les dirigeants, mais aussi les chercheurs, à prendre en compte l'entreprise avec ses ramifications.

Des travaux antérieurs au sein de notre équipe ont porté sur l'entreprise virtuelle (Burlat et Benali, 2007). Ces travaux ont proposé des méthodes et des outils d'aide à la décision pour la construction des réseaux d'entreprises basés sur deux critères économiques clés de rapprochement d'entreprises : la complémentarité des activité et la similarité des compétences. Ces outils sont basés sur la collecte et le traitement des données concernant les entreprises. Ces données sont collectées manuellement à partir d'un questionnaire rempli par les dirigeants d'entreprises. Il s'avère que les dirigeants ne sont pas toujours collaboratifs et actifs pour fournir l'information pertinente. Ce qui est une limite majeure pour ces outils. Notre contribution vise à résoudre cette limite en proposant des méthodes automatiques de collecte et de traitement des données pour la détection des activités d'entreprises complémentaires. Ces méthodes sont développées dans un environnement ouvert basé sur l'information publique, sans frontière restreinte sur la recherche des partenaires. Elles reposent sur la recherche et l'extraction d'information à partir des sites web des entreprises.

2 Organisations virtuelles et recherche d'informations

2.1 Organisations virtuelles

Le concept de l'organisation virtuelle (OV) représente un des exemples les plus discutés des réseaux de collaboration, qui a soulevé des espérances considérables dans beaucoup de domaines d'application (réseaux d'entreprises, les hôpitaux, les universités, les organisations gouvernementales etc). La possibilité de former rapidement une OV, déclenchée par une opportunité commerciale et spécifiquement conçue en fonction des conditions de cette occasion, est fréquemment mentionnée comme expression d'un mécanisme d'agilité et de survie face à la turbulence du marché. La même idée est également très attrayante dans d'autres contextes orientés affaires. Dans la suite nous allons expliciter la problématique des organisations virtuelles dans le cadre de notre travail, montrant en particulier comment est justifié le besoin de la recherche et l'extraction d'information pour la construction des réseaux d'entreprises.

Si nous nous intéressons à cet objectif de recherche et d'extraction de l'information pertinente qui permettent la construction des réseaux, il y a en littérature beaucoup de recherche traitant les données caractéristiques sur les partenaires potentiels pour des organismes gérés en réseau (Camarinha-Matos et Afsarmanesh 2003, Plisson.J et al. 2007, Ermilova et al. 2005). Cependant, ces approches sont développées dans un environnement virtuel fermé (Virtuel Breeding Environment). Ce VBE fournit déjà une présélection des partenaires potentiels, dans lesquels tous les organisations donnent volontairement les données caractéristiques exigées. Au contraire, l'approche que nous nous présentons dans cet article est basée sur l'hypothèse d'un environnement ouvert des partenaires potentiels, de ce fait ayant une plus large application.

Dans le cycle de vie des organisations virtuelles, on considère que la création d'entreprise virtuelle à court terme requiert la mise en place préalable de réseau à long terme nommé VBE (FIG 1).

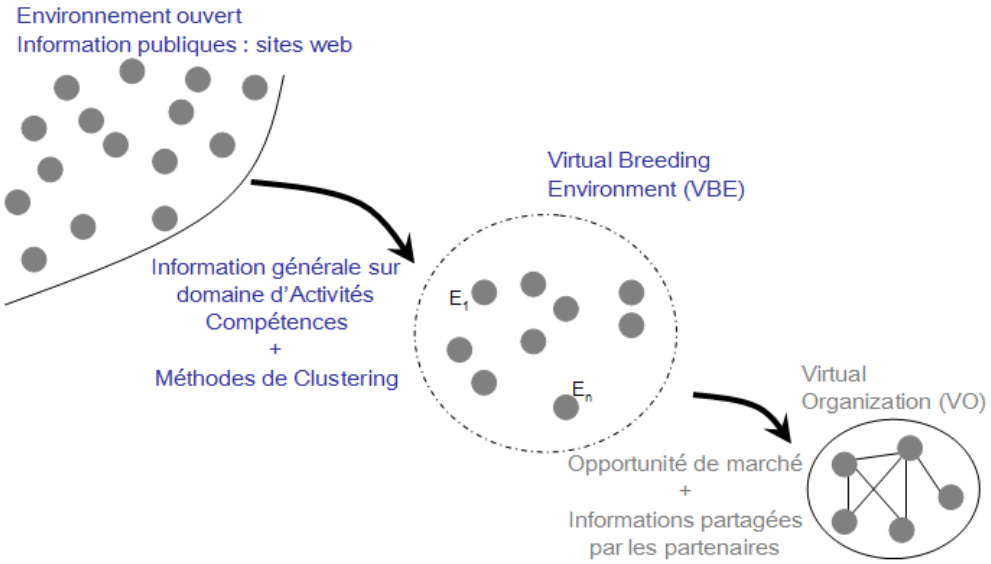

FIG. 1 – *Cycle de vie des Organisation Virtuelles.*

Un VBE est une association d'organisation adhérant à un accord d'organisation à long terme où ses membres sont recrutés dans un univert Ouvert selon des critères définis par les administrateurs. Une VO est une organisation provisoire déclenchée pour une occasion spécifique de collaboration. L'objectif principal du VBE est d'améliorer l'état de préparation de ses membres pour créer efficacement des VO.

2.2 Besoin de recherche d'informations

Pour faciliter la coopération, ces organisations ont besoin d'une infrastructure leur permettant de partager des documents, de travailler et de communiquer ensemble malgré les contraintes géographiques. C'est pourquoi les organisations virtuelles s'appuient fortement sur les technologies de traitement de l'information.

Pour construire un système d'aide à la décision pour la gestion de la collaboration inter-organisations, les approches de recherche et d'extraction d'informations sont sollicitées pour mettre en exergue l'information caractérisant le réseau (Yeong 2009, Camarinha-Matos et Afsarmanesh 2003, Plisson.J et al. 2007, Ermilova et al. 2005). Ces approches de recherche et d'extraction d'information gèrent la création dynamique des organisations virtuelles. Il existe deux types de recherche pour la gestion dynamique de ces organisations virtuelles:

- Une recherche dans un environnement fermé où les organisations se mettent d'accord d'avance pour travailler ensemble à court terme (pour une durée précise). Pour ce faire, elles partagent leurs connaissances et leurs informations (savoir faire, compétences …) sous un format donné et une structure homogène. Cette alliance est en général définie sur un court terme, une fois le bien ou le service livré, le regroupement est dissocié. Ce type de réseau est caractérisé par des frontières très nettes et les nouveaux venus ne sont autorisés qu'en cas d'incident (Exemple : un partenaire quitte le réseau).

- Une deuxième recherche qui se fait dans un environnement ouvert où les organisations ne se connaissent pas et ont une information hétérogène publique et non restreinte. Ce type d'information rend la recherche plus difficile car on est face à des documents mal structurés. Ce type de réseau est réalisé pour un nombre non prédéfini de processus, ce sont des alliances à caractère stratégique. Toutes les organisations intéressées et correspondantes aux objectifs du réseau peuvent y adhérer.

Notre travail se situe dans le deuxième type de recherche. Des travaux antérieurs au sein de notre équipe ont proposé une typologie des modes de coordination entre les différentes entreprises du réseau (Peillon 2001, Burlat et al. 2001). Cette typologie est basée sur deux paramètres : la complémentarité des activités et la similarité des compétences. Ces deux paramètres ont été identifiés comme étant discriminants pour justifier le choix d'un type de coopération industrielle. C'est pourquoi notre besoin d'information s'articule autour de deux systèmes d'extraction d'information (complémentarité des activités et similarité des compétences).

Nous nous limitons dans ce papier à la recherche et l'extraction d'information sur les secteurs d'activités d'entreprises. Nous proposons une approche basée sur des méthodes et outils de la recherche d'information. Les informations extraites sur les activités et les savoir faire, nous les utiliserons dans une deuxième étape pour montrer comment elles génèrent des nouvelles connaissances et permettent de faire émerger des propositions opérationnelles de mise en réseaux des entreprises.

3 Approche proposée

L'objectif principal de notre approche est de détecter automatiquement le secteur d'activité de l'entreprise. Connaitre l'activité principale d'une entreprise donnée est une question importante pour la gestion d'un réseau de collaboration. C'est aussi une question pertinente pour l'entreprise elle-même pour savoir quels sont ses concurrents ou simplement pour s'assurer qu'elle met suffisamment d'information publique à propos de son activité, par exemple sur son site web.

3.1 Démarche de recherche

Notre démarche de recherche est basée sur le modèle de transformation données, informations, connaissances : Les données sont les sites web des entreprises représentés par les pages html contenant des textes mal structurées. Les informations sont les secteurs d'activités que nous souhaitons les détecter automatiquement par l'application des modèles et des méthodes de la recherche d'informations. Les connaissances sont les regroupements d'entreprises qui sont émergés par l'application des méthodes formelles d'aide à la décision.

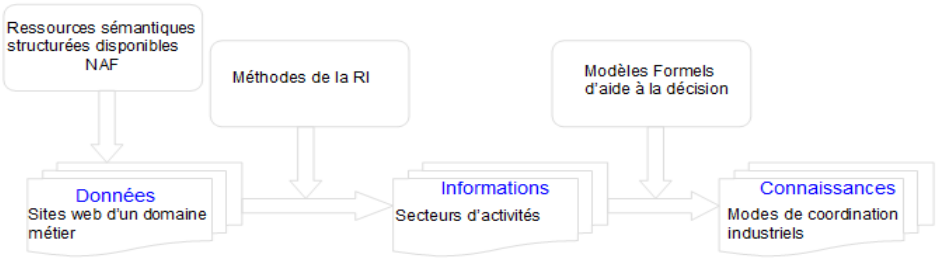

FIG. 2 – *Démarche de recherche.*

Avec cette démarche de recherche, nous cherchons aussi à étudier la question des performances et de l'adéquation éventuelle des techniques de la recherche d'information dans une application spécifique à un domaine d'information métier ciblé (secteurs d'activités des entreprises). Le domaine métier est en premier lieu caractérisé par une complexité importante liée au fait que l'information s'y exprime de manière peu contrôlée et peu structurée : les textes composant le corpus ne suivent aucune structure standard ; la sémantique du vocabulaire utilisé est très lié au domaine métier (vocabulaire contextualisé) ; la structure linguistique des textes est parfois absente ; l'ensemble de ces facteurs induisent de forts risques d'ambiguïté. Mais le domaine métier est également caractérisé par un ensemble de spécificités dont on peut tirer parti de manière formelle, permettant de réduire cette complexité intrinsèque. Dans notre démarche de recherche, nous n'avons pas de réponse a priori sur l'efficacité des techniques de RI lorsqu'elles sont confrontées à la réalité de l'information métier : l'évaluation de leurs performances font partie de l'étude.

Ayant délimité ce "domaine informationnel métier", nous avons cherché à tirer parti de ses spécificités en cherchant des caractérisations de ce domaine, afin d'accroitre l'efficience des dispositifs de RI : par quelle unité informationnel est exprimé ce domaine (mot, expression ou phrase)? Quelle granularité peut-on avoir sur les secteurs d'activités des entreprises? Quelle ambigüité informationnelle et sémantique peut-on croiser dans ce domaine et par quelle ressource sémantique (taxinomie, thesaurus) peut-on guider la recherche?

3.2 Approche de détection des secteurs d'activités

Notre approche se déroule en trois phases décrites par la figure 4. Tout d'abord, dans la première phase nous utilisons un thésaurus qui reflète une représentation sémantique et conceptuelle de tous les domaines d'activités. Dans notre cas le thesaurus est le code NAF[1]. C'est l'un des codes d'INSEE, il permet la codification de l'activité principale exercée dans l'entreprise ou l'association. Le code NAF nous fournit une représentation conceptuelle hiérarchisée de tous les secteurs d'activités d'un domaine industriel : c'est une structure hiérarchique de classes et sous-classes de secteurs d'activités. Ce code NAF va être utilisé comme ressource sémantique externe, afin d'améliorer l'expressivité du besoin d'information avant de le soumettre au système de recherche d'information. L'intérêt du code NAF est qu'il délimite le domaine de recherche en explicitant ses caractéristiques et ses

[1]. http://www.insee.fr/fr/nom_def_met/nomenclatures/naf/pages/naf.pdf

spécificités. Le système de détection des secteurs d'activités que nous réalisons traite des entreprises françaises, mais il est facilement exploitable à l'international pour tout pays francophone : la détection automatique du NAF permet de traiter toutes les entreprises, indépendamment du fait que leur NAF soit ou non répertorié dans les bases de données institutionnelles. Dans notre recherche, le NAF est utilisé pour améliorer l'efficacité du processus d'indexation des sites web des entreprises. Il va servir à contrôler l'information qui circule dans le texte pour ne laisser passer que celle pertinente à notre domaine informationnel. Cette indexation conceptuelle tend à ne sélectionner que les plus importants concepts figurant dans le NAF, au contraire d'une indexation classique qui a pour but de couvrir tout le document. Parallèlement nous utilisons cet apport sémantique de manière plus large grâce aux techniques d'apprentissage par réseau de neurones en créant des liens sémantiques (synonymie, généralisation...) entre les termes du domaine.

FIG. 3 – *Extrait des classes et sous-classes NAF utilisées*

Notre thesaurus est utilisé en amont du moteur de recherche. Il sert de ressource sémantique externe pour améliorer l'expressivité du besoin d'information (quelle est mon code NAF à partir de mon site web ?) avant de le soumettre au système de recherche d'information. Cette technique peut s'avérer efficace, notamment lorsqu'il s'agit d'information traitant d'un domaine spécifique (activités des entreprises par exemple), dans la mesure où elle permet à l'utilisateur d'exprimer son besoin d'information dans un langage contrôlé. Nous effectuons la lemmatisation (avec l'outil TreeTagger) des termes du thesaurus ainsi qu'une élimination des mots vides. Le résultat est le Vocabulaire Contrôlé Hiérarchique (VCH) qui est un ensemble de termes (mots simples et mots composés), par exemple : usinage, emboutissage, machines-outils.

Dans une première phase une pondération manuelle est faite sur ce vocabulaire contrôlé ; elle permet d'attribuer, par expertise, un poids (1, 2 ou 3) pour chaque terme. Le poids d'un terme dans un document traduit l'importance de ce terme dans le document. En réorganisant l'ensemble des termes du VCH selon la structure initiale du NAF, nous obtenons un vecteur pour chaque classe NAF (vecteur classe). Dans une deuxième phase, nous utilisons le VCH pour réaliser une pondération automatique du site web de l'entreprise. Cette pondération est basée sur le calcul de la fréquence du terme dans le texte du site de l'entreprise après avoir effectué un filtrage pour ne garder que les termes qui sont présents dans le VCH.

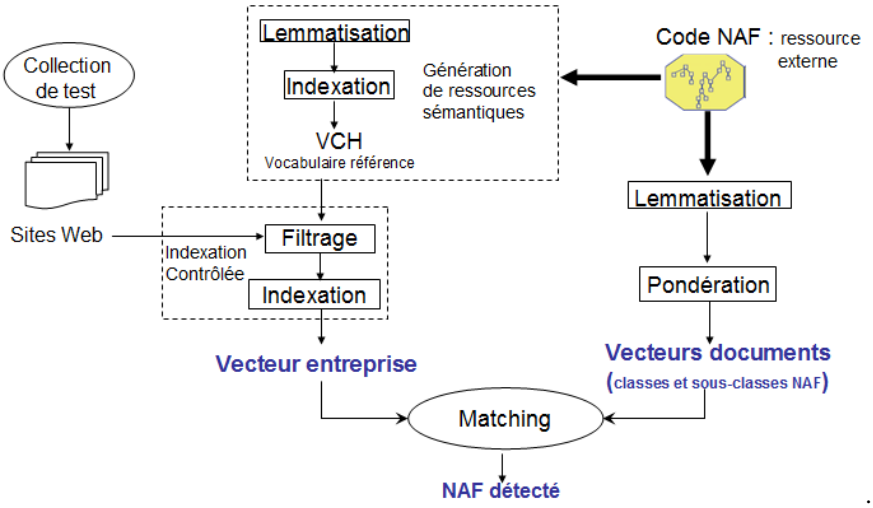

FIG. 4 – *Approche visée pour l'extraction des activités d'entreprises*

Notre approche (FIG 4) repose sur l'idée qu'il existe un rapport entre le contenu véhiculé par un texte et les mots utilisés dans le texte, que ce rapport est en fonction de la fréquence d'usage des mots, et qu'il existe une relation entre la capacité d'un mot à être choisi comme terme d'indexation et sa fréquence d'emploi. Avec ces deux phases, nous voulons construire des vecteurs pour toutes les classes et les sous-classes du NAF, i.e C28, C28.1, C28.2, etc, et construire un vecteur pour chaque site web d'entreprise. Chaque vecteur est l'ensemble des descripteurs d'un document (classes ou sous-classes NAF) ou d'une requête (site web d'une entreprise) avec leurs pondérations (poids informationnels). Pour cela, on utilise les techniques traditionnelles de la RI et une représentation vectorielle des termes des libellés des classes et sous-classes NAF. Dans une troisième phase, on effectue un matching entre le vecteur classe et le vecteur entreprise pour mesurer le degré de rapprochement.

3.3 Matching

L'étape de matching (appariement), quant à elle, constitue une phase de calcul de similarité entre « vecteur requête » et « vecteur document ». Elle se déroule en deux étapes : dans un premier temps on cherche à se positionner sur la classe du NAF la plus pertinente pour l'entreprise ; dans un deuxième temps on explore les sous-classes de cette classe pour se positionner de nouveau sur une sous-classe. En vue d'optimiser la performance finale, nos travaux nous ont conduits à développer d'une part des mesures de similarité utilisant trois fonctions traditionnelles de la recherche d'information (Produit scalaire, Cosinus et Jaccard) et d'autre part une mesure de similarité basée sur un modèle connexionniste (mise en place d'un réseau de neurones pour le calcul de l'appariement (Hajlaoui et al. 2009)). Dés le départ une fonction de similarité de type td-idf nous a semblée mal adaptée à notre cas, car les documents de la collection ont des petites tailles. C'est pourquoi notre choix de fonctions de similarité s'est porté sur ces trois fonctions principales. Nous nous limitons ci-dessous à synthétiser les performances obtenues.

4 Evaluation

L'évaluation de la performance du système est basée sur le calcul des deux indicateurs de performance « précision » et « rappel ». La précision est la capacité du système à rejeter les documents non pertinents, le rappel est la capacité du système à retrouver les documents pertinents. Notre objectif est d'augmenter la précision du système ainsi que son rappel, mais aussi éviter le plus possible d'avoir des valeurs nulles qui signifient que le système ne retrouve pas de documents pertinents. La performance du système dépend non seulement de la mesure de la similarité mais de la façon de définir l'ensemble final de documents appropriés.

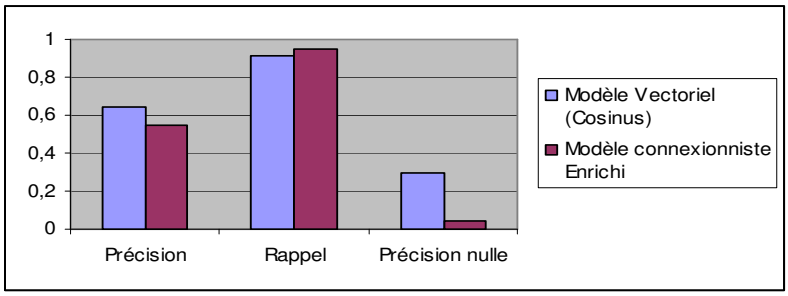

FIG. 5 – *Evaluation de l'identification des secteurs d'activités*

La figure 5 ci-dessus compare les performances entre la mesure de similarité basée sur la fonction cosinus (plus performante que les fonctions « Produit Scalaire » ou « Jaccard » suite à nos expérimentations), et un matching mise en œuvre par un réseau de neurones (tel que spécifié dans (Hajlaoui et al. 2009)). Ces résultats mettent en évidence que la précision est légèrement meilleure pour la fonction cosinus, mais que le modèle connexionniste a permis d'améliorer le rappel et l'indicateur de précision nulle. Concernant la comparaison entre ces 2 appariements, d'autres expérimentations seraient nécessaires dans le futur pour obtenir des conclusions plus définitives. En revanche, nous pouvons d'ore et déjà confirmer que la capacité à identifier correctement un code NAF est d'ores et déjà élevée (TAB 1). Ainsi, ces techniques de recherche d'information s'avèrent efficaces, lorsqu'elles sont enrichies par l'usage d'une ressource sémantique externe spécifique au métier, du type du code NAF.

	modèle Vectoriel	Modèle connexionniste 2
Classes NAF identifiées	92%	88%
Sous-classes identifiées	76%	88%

TAB. 1 – *Performance obtenues pour les deux modèles vectoriel et connexionniste*

5 Application à la constitution des réseaux d'entreprises

Cette application constitue un test de faisabilité sur l'usage de l'information extraite des sites web sur les codes NAF d'entreprise. L'usage que nous testons concerne la recherche de réseaux collaboratifs inter-entreprises. Comme nous l'avons introduit plus haut, il s'agit ici

de tester la possibilité d'appliquer une méthode déjà existante et développé par (Benali, 2005). En ce qui concerne l'analyse des activités d'entreprises, cette méthode préconise d'identifier un Graphe de Complémentarité des Activités (GCA), afin d'appliquer à ce graphe un algorithme de clustering.

Afin de déterminer un graphe de complémentarité entre les secteurs d'activités d'entreprises considérées, nous avons utilisé les caractéristiques du code NAF. Le code NAF présente l'intérêt d'être générique, et reconnu par les acteurs du monde industriel (notamment en mécanique). Nous avons ainsi recueilli des informations complémentaires auprès d'experts du domaine de la mécanique pour évaluer un degré de complémentarité générique entre les classes ou sous-classes du code NAF. (Hajlaoui et al., 2008). L'information recueillie auprès de ces experts a été formalisée sous forme d'une matrice de degrés de complémentarité entre les différents secteurs d'activités associés aux codes NAF (domaine mécanique). Cette Matrice peut être également représentée sous forme d'un graphe (figure 4).

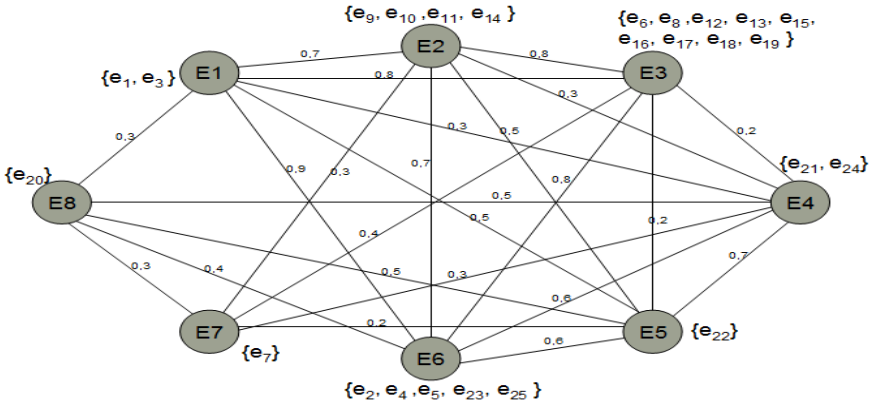

Figure 4. *Résultat du positionnement automatique des 25 entreprises sur le GCA*

Les 25 entreprises étudiées lors de l'étape de détection des secteurs d'activités sont distribuées sur 8 secteurs d'activités. Une entreprise représentative est choisie pour chaque secteur. Nous réutilisons l'algorithme proposé par (Burlat et Benali, 2007). L'objectif de cet algorithme est d'isoler des sous-graphes fortement interconnectés en minimisant la perte d'information (perte d'arcs, perte de complémentarité potentielle). Ces sous-graphes représenteront les entreprises très complémentaires qui permettront de justifier d'une relation de type « réseau proactif » ou de type « firme ». L'algorithme est basé sur du partitionnement et il prend en compte plusieurs aspects spécifiques du graphe de complémentarité des activités établi par l'expert. Il prend en compte non seulement la quantité d'information perdue, mais aussi la qualité d'information : la quantité d'information c'est le nombre d'arcs éliminés et la qualité d'information est donnée par le degré de complémentarité. L'algorithme regroupe les entreprises en petits réseaux (dénommés clusters) en éliminant le moins des arcs possibles et les moins significatifs (de poids faible). Un indicateur caractéristique des regroupements obtenus est l'indicateur I qui traduit l'intensité de coopération au sein d'un cluster.

Etapes	Arc(k)	Arcs éliminés	I	Sous-groupe	Qualité
1	0,1	Ø	0	Ø	Faible
2	0,2	{E7, E5}{E5, E3}{E3, E4}	0,05	Ø	Faible
3	0,3	{E1, E4}{E1, E8}{E2, E4}{E2, E7}{E4, E7}{E8, E7}	0,2	Ø	Faible
4	0,4	{E8, E6}{E7, E3}	0,27	{E7}{E1, E2, E3, E4, E5, E6, E8}	Moyenne
5	0,5	{E1, E5}{E5, E2}{E8, E4}{E8, E5}	0,44	{E7}{E8}{E1, E2, E3, E4, E5, E6}	Bonne
6	0,6	{E4, E6}{E5, E6}	0,54	{E7}{E8}{E5, E4}{E1, E2, E3, E6}	Bonne
7	0,7	{E1, E2}{E2, E6}{E5, E4}	0,72	{E7}{E8}{E5}{E4}{E1, E2, E3, E6}	Bonne
8	0,8	{E1, E3}{E2, E3}{E3, E6}	0,92	{E7}{E8}{E5}{E4}{E1, E6}{E3}{E2}	Bonne
9	0,9	{E1, E6}	1	{E7}{E8}{E5}{E4}{E1}{E6}{E3}{E6}{E2}	Bonne

TAB. 2 – *Construction des groupes d'entreprises en coopération*

L'algorithme de division procède par itérations successives pour déterminer les sous-groupes d'entreprises qui ont des activités complémentaires. Le nombre pertinent d'itérations peut être choisis lors de l'expérimentation en référence à deux indicateurs de qualité de la solution fournie (TAB 2) Par exemple après six itérations les sous-groupes suivants sont obtenus :

$$G_1 = \{E7\}; G2 = \{E8\}; G3 = \{E5, E4\}; G4 = \{E1, E2, E3, E6\}$$

L'entreprise E7 se retrouve isolée dès le début des itérations, c'est-à-dire avec une très faible intensité de coopération. Nous avons pu vérifier a posteriori la cohérence de ce résultat, compte tenu du code NAF de E7. Deux clusters G3 et G4 apparaissent. Ces 6 itérations correspondent à une intensité de coopération moyenne (I=0,54), au sein des clusters, avec une logique de coopération intra-cluster de type Réseau Proactif (Burlat et Benali, 2007).

Cette première application, met en évidence la faisabilité l'ensemble de la démarche proposée, consistant à extraire des sites web d'entreprises une information synthétique sur les secteurs d'activités d'entreprises, afin d'appliquer dans une seconde étape des outils d'aide à la décision facilitant l'émergence de réseaux collaboratifs inter-entreprises. Cependant les premiers résultats montrent les limites de n'utiliser qu'une information concernant les « secteurs d'activités » : en effet nous avons souligné que les 25 entreprises de l'expérimentation se répartissaient au final sur 8 secteurs d'activités. Cette donnée seule ne suffit pas à traiter la problématique de création des réseaux collaboratifs.

6 Conclusion

Nous avons présenté une contribution de détection automatique des activités des entreprises. Cette contribution présente un système automatique d'extraction d'information sur les activités des entreprises à partir de leurs sites web. Elle est basée sur des méthodes et des

outils de recherche d'information. Les mesures de similarité utilisées s'appuient sur les indicateurs standards de la RI (Précision et Rappel) et montrent une performance autour de 80% de bonnes réponses. Cependant la complémentarité des activités des entreprises est insuffisante pour regrouper correctement les entreprises d'un même réseau de coopération. C'est pourquoi nous avons besoin du second système d'extraction concernant cette fois les compétences des entreprises. La question d'extraction des compétences est une question plus complexe à résoudre, en absence des ressources sémantiques structurés propres au domaine métier et décrivant la notion de compétences d'entreprises. Nos travaux de recherche actuels sollicitent cette problématique par le recours à des techniques d'extraction plus avancés : Analyse de texte, traitement de la langue naturelle et à la construction et utilisation des ontologies du domaine concernée.

Références

Burlat P., Benali M. *A methodology to characterise co-operation links for networks of firms. Production Planning & Control* Vol. 18, No. 2 March 2007, 156-168

Burlat P., Vila D., Besonbes B., et Deslandres V., "Un cadre de modélisation des trajectoires d'évolution des groupements d'entreprises", Congrès International de Génie Industriel, Marseille, France 2001.

Camarinha-Matos, LM., Afsarmanesh H. *Elements of a base VE infrastructure*,. *Computers in Industry,* 51, pp. 139-163 2003.

Ermilova, E., Galeano N., Afsarmanesh H. *ECOLEAD deliverable D21.2a. Specification of the VBE competency/profile management,* 2005.

Hajlaoui K. and Boucher X. Neural network based text mining to discover enterprise networks. In 13th IFAC Symposium on Information Control Problems in Manufacturing (INCOM'2009). Moscow, Russia, 2009.

Hajlaoui K., Boucher X., Mathieu M. (2008). Data Mining To Discover Enterprise Networks, 9 th IFIP Working Conference on Virtual Entreprises (PRO-VE'08) Poznan, POLAND, 8-10 September 2008.

Plisson, J. Ljubic P, Mozetic I, Lavrac N. (2007). *An ontology for Virtual Organisation Breeding Environments*. *To appear in IEEE Trans. On Systems, Man, and Cybernetics, 2007*

Peillon S., *"Le pilotage des coopérations inter-entreprises : le cas des groupements de PME", Doctorat d'Economie de l'Université Jean Monnet 2001*.

Yeong Su Lee and Michaela Geierhos. *Business specific online information extraction from german websites*. In CICLing '09 : Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing, pages 369–381, Berlin, Heidelberg, 2009. Springer-Verlag. 63

Vers une constitution automatique des réseaux d'entreprises collaboratifs

Summary

In this paper, we propose a search system and automatic extraction of information starting for the web for decision aid approach to support the constitution of collaborative corporate networks. This approach based on an automated procedure of information extraction aiming at identifying key features of potential partner. This knowledge is the reason that may or must lead firms to consider cooperation between them. In this goal, to precisely know the activity of a company in real-time and its savoir faire is indispensable. In this article we present an approach based on the tools and the methods of information retrieval: controlled indexing and measuring similarity. This approach defines an IRS that allows automatic detection of the activity and the savoir faire from the company web site. The detection makes possible to emerge cooperative networks companies with various types

Robustesse de communautés émergées des graphes sociaux issus des réseaux de communication

Slimane Lemmouchi,
Mohammed Haddad, Hamamache Kheddouci

Laboratoire LIESP Université Claude Bernard
43, Bd du 11 novembre 1918 - 69622 Villeurbanne
{slemmouc, mhaddad, hkheddou}@bat710.univ-lyon1.fr

Résumé. Les réseaux sociaux et l'étude des communautés d'intérêt sont devenus un véritable challenge et ce, dans différents domaines de recherche. Les interactions entre les nœuds dans les réseaux de communications (exemple: les réseaux pair à pair) induisent des structures de graphes ayant des caractéristiques semblables à celles trouvées dans les réseaux sociaux (effet du petit monde, petit diamètre, faible densité globale et forte densité locale). Ces structures ou topologies virtuelles, appelées aussi graphes de communication, représentent les échanges entre les nœuds dans un réseau. Plusieurs aspects tels que l'étude de la nature de ces topologies (arbre, clique, étoile, …), leurs construction et évolution dans le temps, leur robustesse face aux perturbations (disparition d'un lien, panne sur nœud de fort degré,…) ont suscité l'intérêt de la communauté scientifique. Dans ce papier, nous nous intéressons à l'étude de l'émergence de telles structures ainsi qu'à leur robustesse dans les réseaux de communications à grande échelle avec une application aux réseaux pairs à pairs. Nous nous focaliserons sur la quantification de la robustesse de ces structures de communautés ayant émergées des communications issues de réseaux à nature pair-à-pair.

1 Introduction

Les réseaux « pair à pair » ou « P2P » permettent d'échanger et de partager des ressources. La principale caractéristique de ces réseaux est l'absence d'entités centrales et ainsi, les pairs peuvent communiquer directement entre eux pour trouver et partager des ressources. Contrairement aux systèmes client-serveur où peu de serveurs sont connectés à plusieurs clients, les systèmes « pair à pair » sont définis comme des systèmes distribués où les ressources et services sont partagés directement entre les pairs. Ces systèmes sont robustes, anonymes, flexibles et auto-organisables. Ils peuvent fonctionner avec une grande stabilité et une bonne résistance face, éventuellement, à une haute dynamique.

Les réseaux P2P sont devenus ces derniers temps un centre d'intérêt au regard du contenu échangé entre les utilisateurs sur internet. Des études récentes [1, 9] estiment que les fichiers P2P échangés comptent pour environ 70% du trafic réseau. En effet, plusieurs études ont été menées pour diminuer le trafic dans les réseaux P2P en améliorant la recherche de ressources dans ces réseaux. A ce titre, plusieurs approches proposent le concept de « P2P social network » pour capturer les caractéristiques sociales des pairs et des ressources partagées [7].

Semblables à des réseaux sociaux humains, un « réseau social P2P » est une collection de pairs informatiques connectés (nœuds), dont chacun de ces pairs connaît un sous-ensemble d'autres pairs. Les liens sociaux des pairs indiquent qu'un pair est un fournisseur de ressources ou peut fournir des informations sur d'autres fournisseurs de ressource.

Dans ce type de réseaux, les pairs ne connaissent pas tous les autres pairs dans le réseau. Ils peuvent communiquer directement avec certains pairs. Dans le cas idéal, un pair voudrait communiquer uniquement avec les pairs ayant les ressources qui l'intéressent ou partageant avec lui un centre d'intérêt commun. Le centre d'intérêt d'un nœud est l'ensemble de ressources qu'il veut trouver sur un réseau. Cette communication entre pairs fait ressortir une topologie virtuelle qui représente les relations entre les pairs ayant échangé des ressources. Cette topologie émergée, qu'on appellera *graphe de communication*, possède une structure semblable à celle dans les réseaux sociaux. Dans les réseaux sociaux, les individus tendent à communiquer et à partager plus de choses avec des individus se trouvant dans un même groupe. Ces groupes sont en général appelés *clusters* ou *communautés*.

La découverte des structures de communautés émergées dans les réseaux a fait l'objet de plusieurs travaux. Les techniques pour détecter les communautés d'un réseau sont un moyen couramment utilisé pour simplifier l'étude des réseaux, étant donné que le nombre de clusters est en général bien plus petit que le nombre de nœuds. Cependant, des interrogations ont été suscitées telles que : la robustesse de ces communautés émergées? Leurs structures changent-elles aux perturbations ? C'est ce que nous essayons d'y répondre dans ce travail.

Ce papier est organisé comme suit. La section 2 présente brièvement quelques concepts sur les réseaux P2P, les techniques de détection des structures de communautés ainsi que leur robustesse. La section 3 sera consacrée à la présentation de notre approche. La section 4 est dédiée aux expérimentations. La section 5 à la description des résultats préliminaires et enfin, la section 6 clôturera notre papier.

2 Travaux antérieurs

Dans cette section, nous définissons quelques concepts utilisés dans ce papier.

2.1 Les réseaux sociaux P2P

La notion de réseaux sociaux et les méthodes d'analyse des réseaux sociaux ont considérablement suscité, ces dernières années, l'intérêt de la communauté scientifique spécialisée dans les sciences sociales. Plusieurs recherches se sont intéressées à l'étude des types des relations entre les entités et leur impact sur le réseau. En effet, Stanley Milgram fut l'un des premiers ayant étudié les réseaux sociaux [20] à travers l'expérimentation (effet du petit monde) dans laquelle, il est montré que chaque individu peut être connecté à n'importe quel autre individu par une courte chaine de relations sociales. Ce concept a donné naissance à un autre concept appelé « six degrees of separation ». Ce dernier concept montre que deux ci-

toyens américains, choisis d'une manière aléatoire, sont connectés entre eux, en moyenne, par une chaine de six relations.

Les réseaux sociaux sont des groupes d'individus d'une organisation, connectés entre eux par des relations. Les individus développent des relations (liens) avec d'autres individus dans différents contextes et utilisent ces relations pour trouver des informations ou des services, dépendant du contexte.

Comme déjà mentionné plus haut, plusieurs études récentes proposent des réseaux P2P sociaux qui capturent des associations sociales des pairs à partir des ressources partagées. Ces associations sociales sont caractérisées notamment par la *distribution de degrés*, le *coefficient de clustering*, le *plus court chemin*, la *corrélation entre la betweenness et le degré*. Cette analyse donne une meilleure compréhension des associations des pairs dans le partage de ressources et apporte un plus dans la conception des réseaux P2P.

Dans [10], les auteurs proposent une approche basée sur l'aiguillage de requêtes vers des pairs fortement susceptibles de satisfaire ces requêtes avec une certaine similarité. Cette approche, qui utilise la simplicité de Gnutella, stipule que si un pair a un morceau particulier de contenu qui interesse un autre pair, alors il est très probable qu'il ait d'autres ressources qui l'intéressent aussi. Une approche aussi semblable à cette dernière et qui se fonde sur l'information fournie par la corrélation des utilisateurs, est également présentée dans [5].

D'autres approches utilisant les concepts « schema based peer-to-peer networks » en combinant la notion de la technologie web sémantique et les bases de données ont été également présentées dans [2, 11]. Dans ces cas, la description du nœud est prise en compte pour grouper avec lui les nœuds ayant une similarité de contenu.

D'autres travaux dans la thématique sont également élaborés et on retrouve également les approches utilisant les schémas ontologiques. L'ontologie rapporte différents concepts employés pour décrire le contenu informationnel des pairs. De cette façon, la recherche est exécutée non seulement sur des pairs de contenu semblable, mais également sur les pairs dont le contenu est connexe par l'ontologie avec le contenu de la requête [21]. Cette recherche n'est pas uniquement effectuée sur les pairs dont le contenu a une similarité avec la requête, mais également sur les pairs dont leurs contenus sont connexes à travers l'ontologie. Dans [8], les auteurs ont défini une méthode, dénommée *Semantic Partition Tree* « SPT » qui utilise l'ontologie et qui consiste à partitionner les réseaux en clusters et les requêtes sont routées d'une manière sémantique et chaque nœud indexe les adresses des autres nœuds qui possèdent le contenu exprimable par le concept qu'il maintient.

Dans [16], les auteurs ont utilisé le concept de proximité sémantique qui exploite les intérêts exhibés parmi des pairs afin de décomposer le réseau en clusters sémantiques. Dans ce travail, les auteurs ont introduit la notion de « user fileset » qui permet à chaque pair d'exprimer son intérêt à travers un ensemble de noms de fichiers qui correspond à un nombre représentatif de fichiers qu'il gère.

2.2 Détection des communautés

La détection de structures de communautés dans un réseau à grande échelle a fait l'objet de plusieurs travaux. Détecter des communautés dans un réseau revient à le découper en sous-ensembles tel que chaque sommet d'un groupe possède plus de liens à l'intérieur du groupe qu'à l'extérieur.

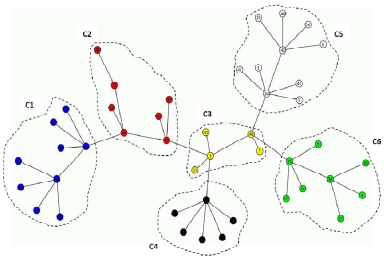

FIG. 1 – *Détection de communautés dans un réseau.*

Un grand nombre de méthodes pour trouver des communautés ont été proposées ces dernières années. Il existe essentiellement deux classes d'approches : les approches séparatives et les approches agglomératives.

2.2.1 Les approches séparatives :

L'idée commune à toutes ces méthodes est d'essayer de scinder le graphe en plusieurs communautés en retirant progressivement les arêtes reliant des communautés distinctes. Les arêtes sont retirées une à une, et à chaque étape les composantes connexes du graphe obtenu sont identifiées à des communautés. Le processus est répété jusqu'au retrait de toutes les arêtes. On obtient alors une structure hiérarchique de communautés (dendrogramme). Les méthodes existantes diffèrent par la façon de choisir les arêtes à retirer.

Dans [12], Newman *et al.* ont défini un algorithme pour trouver les arêtes avec une haute betweenness ou centralité d'intermédiarité. La centralité d'intermédiarité est définie pour une arête comme étant le nombre de plus courts chemins passant par cette arête. Cette méthode retire les arêtes de plus forte centralité d'intermédiarité. Il existe en effet peu d'arêtes reliant les différentes communautés et les plus courts chemins entre deux sommets de deux communautés différentes ont de grandes chances de passer par ces arêtes.

Les algorithmes de Radicchi *et al.* [6] et d'Auber *et al.* [4] basés sur le clustering d'arêtes ainsi que l'algorithme de Fortunato *et al.* [18] basé sur la centralité d'information font aussi partie de la famille des approches séparatives.

2.2.2 Les approches agglomératives :

Dans cette famille d'approches, chaque nœud est considéré comme étant une communauté, *c.à.d.* : il y a autant de communautés que de nœuds dans le graphe initial. Les communau-

tés sont regroupées deux à deux et la structure produite par ces algorithmes est un dendrogramme (un arbre montrant l'ordre de jointure des nœuds). La métrique utilisée dans [13] pour assembler deux communautés est la modularité : $\Delta Q = 2(e_{ij} - a_i a_j)$, où : e_{ij} est la fraction d'arêtes dans le graphe qui connectent les nœuds de la communauté i à la communauté j et $a_i = \sum_j e_{ij}$.

A chaque étape, les paires des communautés qui produisent la meilleure valeur de la modularité sont jointes.

Toutes ces approches, séparatives et agglomératives, peuvent donner différentes partitions (communautés détectées) et, pour mesurer la qualité de la partition obtenue, des fonctions de qualité ont été introduites. La fonction définie dans [14] en est une.

2.3 Robustesse des communautés

Les techniques pour détecter les communautés d'un réseau sont un moyen couramment utilisé pour simplifier les réseaux, étant donné que le nombre de clusters est en général bien plus petit que le nombre de nœuds. Cependant, plusieurs interrogations concernant notamment les structures de ces communautés ont été posées. Parmi ces interrogations, nous pouvons citer : la nature des structures détectées (arbre, chaine, étoile, ….), leurs construction dans le temps, leur robustesse face à des perturbations ou bruits introduits sur ces structures (par exemple : disparition d'un lien entre 2 nœuds, crash d'un nœud de fort degré,…)…etc. Dans notre étude, nous nous intéressons à la quantification de la robustesse des structures de communautés ayant émergées des communications entre les nœuds dans le réseau P2P. Mais avant d'introduire notre méthode, un bref état de l'art de quelques principales méthodes utilisées pour perturber et quantifier des communautés est présenté :

Karrer *et al.* [3] ont proposé une méthode pour perturber des réseaux et mesurer les changements résultants dans les communautés. Ces changements sont utilisés pour comprendre la signification des communautés dans une variété de réseaux (réels et générés). Dans leur papier, il a été montré que la signification de la structure de communauté peut-être effectivement quantifiée en mesurant sa robustesse à des petites perturbations dans la structure du réseau.

Dans [19], les auteurs ont exploré la robustesse par la mesure de la centralité aux différentes perturbations. Dans ce travail, il a été introduit quatre types de perturbations : suppression d'arête, suppression de nœud, ajout d'arête et ajout de nœud. Les réseaux utilisés sont des réseaux générés aléatoirement, de différentes tailles et de différentes densités.

Un autre travail intéressant a été détaillé dans [17] où les auteurs ont étudié la réaction des réseaux complexes sujets à des attaques sur des nœuds ou des arêtes. Dans ce travail, les auteurs ont défini quatre stratégies d'attaque pour étudier la corrélation entre la centralité d'intermédiarité et le degré sur six graphes différents.

3 Méthodologie

Comme déjà dit plus haut, notre travail étudie les structures de communautés émergées des échanges entre les pairs. Notre modèle (cf. Figure 2) simule le fonctionnement d'un système P2P. Ce modèle se décompose en trois parties :

FIG. 2 – *Framework de détection de communautés sociales.*

3.1 Modèle du réseau P2P :

Soit G = (V, E) tq : V :ensemble des sommets et E : ensemble des arêtes.

Pour simuler plusieurs réseaux, nous avons généré plusieurs graphes avec différentes tailles et différentes densités. La taille du graphe varie selon la variation du nombre de nœuds et la densité selon le nombre d'arêtes dans le graphe.

3.1.1 Génération du graphe aléatoire :

Les graphes utilisés dans nos simulations sont générés d'une manière totalement aléatoire selon le modèle d'Erdös-Rényi. Du moment que ce modèle n'assure pas systématiquement la connexité du graphe, une propriété Hamiltonienne est ajoutée.

Pour générer ces graphes, nous utilisons une probabilité p ϵ [0,1] d'existence des arêtes. D'une manière générale, plus p est grand, plus la densité du graphe est forte et vice versa. (p=1 : graphe complet)

3.1.2 Modélisation des ressources et des centres d'intérêt :

Pour chaque nœud désigné *fournisseur* ou *serveur*, nous attribuons, aléatoirement, un certain nombre de ressources qu'il devra fournir et, pour l'ensemble des nœuds du graphe, qu'ils soient demandeur ou fournisseur, nous attribuons un certain nombre ressources qui représentent ce qu'on appelle « Centre d'intérêt du nœud « CIN »». CIN est l'ensemble des ressources dont un nœud (client ou serveur) peut chercher sur le réseau. Là aussi, les CIN

sont aussi définis d'une manière complètement aléatoire et la taille du CIN diffère d'un nœud à un autre.

3.2 Communication dans le graphe :

Après avoir défini le graphe physique, le nombre de liens entre les nœuds, les ressources par serveurs ainsi que les centres d'intérêt des nœuds des nœuds (serveurs et clients), nous commençons l'exécution des requêtes de recherche de ressources dans le réseau. Les requêtes, dans notre modèle, est modelisé par l'envoi de message d'un nœud à un autre, contenant la ressource à trouver. Mais, avant d'envoyer chaque requête, le nœud consulte sa liste de nœuds voisins pour déterminer s'il existe parmi eux un nœud avec lequel il a déjà communiqué, auquel cas la requête va lui être transmise. Dans le cas contraire, celle-ci va être transmise à un nœud choisi aléatoirement parmi ses voisins. Dans ce cas, 2 cas se présentent :

- Le nœud recevant la requête peut la satisfaire et la recherche est stoppée ;
- Le nœud recevant la requête ne peut la satisfaire et, dans ce cas, la requête sera transmise à l'un de ses voisins.

3.3 Structure émergées :

A cette étape, le graphe de communication, qui représente l'ensemble des échanges (requêtes) entre les nœuds dans le réseau, est obtenu. Ce graphe possède une structure similaire aux structures de graphe dans les réseaux sociaux [21].

Pour détecter les structures de communautés émergées, nous avons utilisé l'algorithme de Newman décrit dans [13]. Cet algorithme, qui fait partie de la famille des approches agglomératives, permet d'étudier des grands réseaux et utilise une fonction appelée « Modularité » qui permet dévaluer la qualité de la partition obtenue.

3.4 Quantification de la robustesse des communautés :

3.4.1 Perturbation du réseau :

Notre approche pour évaluer la robustesse des communautés émergées se déroule en deux étapes :

1- perturber le graphe de communication obtenu des échanges entre les pairs et,
2- quantifier les changements opérés dans les structures de communautés.

La méthode que nous avons adoptée pour perturber le graphe de communication est similaire à celle utilisée par Karrer *et al.* [3]. Cette méthode consiste en la génération d'un $2^{ème}$ graphe identique au graphe de communication obtenu (même nombre de nœuds et même nombre de liens). L'unique différence entre ces deux graphes est la distribution des arêtes.

En effet, la position des arêtes dans le 2ème graphe est changée selon une probabilité p. Si p=0, aucune arête n'est déplacée et dans ce cas, les deux graphes sont isomorphes. Si p=1, tous les liens sont déplacés. L'arête déplacée est placée entre deux sommets non liés initialement. Dans notre étude, nous introduisons une faible perturbation (déplacement d'un faible nombre d'arêtes). Un déplacement d'un grand nombre d'arêtes génère un graphe aléatoire non corrélé avec le graphe original.

3.4.2 Quantification des différences dans les structures de communautés

Pour évaluer la robustesse des structures de communautés émergées dans les réseaux de communication, nous considérons deux aspects : le nombre de liens déplacés et le nombre de nœuds instables.

Pour le 1er aspect « nombre de liens déplacés », il existe plusieurs méthodes pour quantifier les différences entre les 2 graphes (original et perturbé). En effet, des méthodes pour mesurer les similarités et les différences entre les partitions d'un réseau ont été proposées dans la littérature. Nous pouvons classer ces méthodes en 3 catégories : les méthodes basées sur le comptage de pairs, les méthodes basées sur le « matching » de clusters et enfin, les méthodes théoriques [15].

Le 2ème aspect «nombre de nœuds instables» détermine les nœuds qualifiés d'instables du fait qu'ils n'appartiennent pas à une seule partition. En effet, les algorithmes de détection de communautés affectent les nœuds à des groupes de telle sorte que chaque nœud d'un groupe ait autant de liens dans son groupe que vers d'autres nœuds appartenant à d'autres groupes. Mais, le problème peut se poser pour les nœuds ayant presque autant de liens vers l'intérieur du groupe que vers un autre groupe et dans ce cas précis, la majorité de ces algorithmes affecte ces nœuds dits *instables* à l'une ou l'autre de ces deux partitions. L'étude de ces nœuds particuliers dans un réseau de communication est très intéressante dans le cas où le taux de présence de ce type de nœud est élevé. En effet, leur forte présence dans un réseau de communication peut affecter considérablement son bon fonctionnement.

4 Expérimentation

Dans cette section, nous discutons certains résultats préliminaires à travers des simulations que nous avons effectuées sous NS2 (Network Simulator 2).

4.1 Construction du réseau physique

Pour la construction du réseau P2P, nous utilisons le modèle d'Erdös Rényi. Nous générons plusieurs réseaux de 50, 100, 150, 200, 500 et 1000 nœuds. Pour chaque taille, nous faisons varier la densité (Nb arêtes existantes/Nb arêtes possible) du réseau généré selon les valeurs suivantes : 0.25, 0.5, 0.75 et 1 (Si densité = 1, le graphe est complet). Le nombre de nœuds faisant office de fournisseur de ressources dans le réseau varie selon des proportions

(25%, 50%, 75% et 100%) par rapport au nombre total des nœuds dans le réseau (par exemple, dans un réseau de 100 nœuds, le taux de 50% représente 50 serveurs). Pour ce qui de la distribution de ressources sur les serveurs, elle varie également selon les proportions suivantes : 5%, 15%, 25%, 50%, 75% et 100% du nombre total de ressources dans le réseau. Enfin, le Centre d'Intérêt des Nœuds (clients et fournisseur) varie lui aussi selon les proportions suivantes : 5%, 15%, 25%, 50%, 75% et 100% du nombre de ressources dans le réseau.

Les proportions attribuées à la densité, nous les avons choisies de manière à obtenir des petits graphes à faible densité, des petits graphes à forte densité, des grands graphes à faible densité et des grands graphes à forte densité (du graphe sparse à des graphes plus ou moins complets). Les proportions attribuées à la distribution des ressources et l'attribution des centres d'intérêts des nœuds dans le réseau obéissent à une catégorisation prédéfinie. En effet, nous avons défini 3 catégories de proportions : faible (5% et 15%), moyenne (25% et 50%) et forte (75% et 100%). Le but principal de ce choix est de concevoir des modèles de réseaux P2P proches des systèmes P2P réels.

4.2 Modélisation de la perturbation

La perturbation introduite se résume en le déplacement d'un certain nombre d'arêtes sur le graphe de communication selon les proportions suivantes : 1%, 2%, 3%, 4% et 5%.

Pour chaque réseau généré, comme décrit dans la sous-section précédente, nous faisons varier la quantité de perturbation à introduire.

Pour chacune des configurations, nous calculons la moyenne des propriétés en utilisant les méthodes de quantification des différences dans les structures de communautés, citées plus haut.

4.3 Exemple d'exécution

Dans cette patrie, nous allons donner un exemple du scénario d'une simulation. Nous générons un réseau avec les caractéristiques suivantes : taille = 20 nœuds, densité=0.5, Nb de serveurs = 25%, Nb de ressources/serveur = 50%, Taille CIN = 75% ; perturbation = 3%.

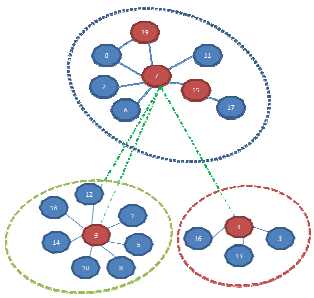

FIG. 3 – *Communautés émergées dans le réseau initial.*

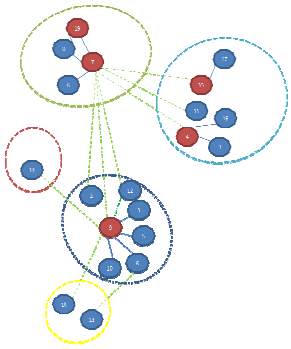

FIG. 4 – *Communautés émergées dans le réseau perturbé.*

Dans cet exemple, notre réseau physique est constitué de 20 nœuds dont 5 serveurs (les nœuds en rouge dans les figures 3 et 4) et 23 liens.

La première phase de notre framework est de construire le réseau P2P physique selon les paramètres en entrée (taille, densité, nombre de serveurs, nombre de ressources/serveurs, taille du CIN des nœuds). Une fois le réseau physique construit, chaque nœud (serveur ou client) recherche les ressources qui constituent son centre d'intérêt. Les échanges entre tous les nœuds du réseau représentent le graphe de communication sur lequel, nous appliquons l'algorithme de Newman décrit dans [13].

La figure 3 illustre les 3 communautés obtenues dans le graphe de communication initial de tailles respectives, 8, 8 et 4 nœuds tandis que la figure 4 représente les communautés obtenues (5 communautés de tailles respectives 7, 6, 4, 2 et 1 nœud) après introduction d'une perturbation (déplacement de 3% des liens) sur le graphe de communication initiale.

5 Résultats préliminaires

Dans toutes les expériences que nous avons effectuées, nous avons remarqué que plus le graphe de communication est dense (par exemple une clique), plus sa structure est résistante aux perturbations. Un graphe de communication plus dense signifie que le pourcentage des serveurs est élevé et que ces derniers ne possèdent toutes les ressources, c'est-à-dire, que tout le monde communique avec tout le monde. A l'inverse, un graphe est moins dense (arbre ou étoile) est sensible aux faibles perturbations.

6 Conclusion

Dans ce travail, nous nous sommes intéressés aux structures de communautés dans les réseaux de communication, avec une application aux réseaux pairs à pairs. Dans un premier temps, nous avons défini un modèle, constitué en trois parties, pour simuler le fonctionnement d'un réseau de communication et déterminer les communautés construites suite aux

échanges sociaux entre les entités du réseau. La construction de ces communautés dépend des paramètres physiques en entrée (taille du graphe, densité, nombre de serveurs, la distribution de ressources et le centre d'intérêt des nœuds). Si cette première partie de notre étude s'intéresse à l'étude globale des structures émergées (type de la structure : clique, arbre, étoile, ….), la seconde partie de notre travail, quand à elle, s'intéresse à l'étude détaillée de ces mêmes structures et de leur robustesse. Le but recherché dans cette seconde partie du travail est de déterminer l'influence des paramètres physiques sur la construction des structures sociales des communautés.

Notre modèle a été implémenté et testé avec toutes les configurations possibles. Nous envisageons d'élargir notre expérimentation pour des réseaux de plus grandes tailles et d'appliquer notre approche sur d'autres réseaux de communications.

Références

[1] A. Madhukar, C. Williamson. A longitudinal study of P2P trafic classification. In : Proceedings of MASCOTS'06, Monterey, USA, August 2006.

[2] A.Y.Halevy, Z.G. Ives, P.Mork, I. Tatarinov, Piazza:data management infrastructure for semantic web applications, in: Proceedings of the 12th International World WideWeb Conference (WWW2003), Budapest, Hungary 2003.

[3] Brian Karrer, Elizaveta Levina, and M.E.J Newman. Robustness of community structure in networks. Phys. Rev. E 77, 046119, 2008.

[4] David Auber, Yves Chiricota, Fabien Jourdan, and Guy Melançon. Multiscale visualization of small world networks. In Proceedings of the 9th IEEE Symposium on Information Visualization (InfoVis 2003), page 10, Seattle, USA, 2003. IEEE Computer Society.

[5] E. Cohen, A. Fiat, H. Kaplan, Associative search in peer to peer networks: harnessing latent semantics, in: Proceedings of the 22nd IEEE INFOCOM, vol. 2, April. 2003, pp. 1261–1271.

[6] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. PNAS, 101(9) :2658_2663, 2004.

[7] F. Wang, Self-organizing Cognitive Peer-to-Peer Social Networks, unpublished (2005).

[8] Habib Rostami, Jafar Habibi and Emad Livani, 2008, Semantic partitioning of peer-to-peer search space. Computer Communications, Volume 32, Issue 4, pp. 619-633, 2009.

[9] Jeffrey Erman, Anirban Mahanti, Martin Arlitt, Carey Williamson,, Identifying and Discriminating Between Web and Peer-to-Peer Traffic in the Network Core, Canada 2007.

[10] K. Sripanidkulchai, B. Maggs, H. Zhang, Efficient content location using interest-based locality in peer-to-peer systems, in: Proceedings of the 22nd IEEE INFOCOM, vol. 3, April 2003, pp. 2166–2176.

[11] K. Aberer, P. Cudré-Mauroux, M. Hauswirth, The chatty web: emergent semantics through gossiping, in: Proceedings of the 12th International World WideWeb Conference (WWW2003), Budapest, Hungary 2003.

[12] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. Physical Review E, 69:026113, 2004.

[13] M. E. J. Newman. Fast algorithm for detecting community structure in networks. Physical Review E, 69:066133, 2004.

[14] M. E. J. Newman. Modularity and community structure in networks. PNAS, 103(23):8577, 2006.

[15] Marina Meila. Comparing clusterings– an information based distance. J. multivar. Anal., 98(5): 873-895, 2007.

[16] Nikolaos D. Doulamis, et al. Exploiting semantic proximities for content search over P2P networks. Computer Communications Volume 32, Issue 5, pp. 814-827, 2009

[17] Petter Holme, and Beom Jun Kim. Attack vulnerability of complex network. Phys. Rev. E 65, 056109 (2002). DOI: 10.1103/PhysRevE.65.056109

[18] Santo Fortunato, Vito Latora, and Massimo Marchiori. Method to find community structures based on information centrality. Physical Review E, 70(5) :056104, 2004.

[19] Stephen P. Borgatti, Kathleen M. Carley, David Krackhardt. On the robustness of centrality measures under conditions of imperfect data. Doi:10.1016/j.socnet.2005.05.001.

[20] S. Milgram. The small world problem. Psychology Today, 2(1):60{67, 1967.

[21] Vincenza Carchiolo et al. Social behaviours applied to P2P Systems:An efficient algorithm for resources organisation. arXiv:cs/0702085v1, Feb 2007.

Summary

Social networks and the study of communities of interest have become a real challenge in different research areas. The interaction between the nodes in communication networks (eg. peer to peer networks) induce graph structures with characteristics similar to those found in social networks (small world effect, small diameter, low density global and local density). These structures or virtual topologies, also called *communication graph*, represent exchanges between nodes in a network. Several aspects such as the study of the nature of these topologies (tree, clique, star, ...), their construction and development over time, their robustness against disturbances (disappearance of a link failure on node of high degree, ...) have attracted the interest of the scientific community. In this paper, we focus on the study of the emergence of such structures and their robustness in communications networks on a large scale with an application to peer to peer networks. We focus on quantifying the robustness of the structures of communities with emergent communications from issued from peer-to-peer networks.

Stochastic Networks

Cynthia Basileu*, Sofian Benamor**
Marc Bui**, Ahmed Bounekkar*
Nadia Kabachi*, Michel Lamure*
Mondher Toumi*

*Laboratoire ERIC - University of Lyon - University Claude Bernard
Batiment Nautibus, 43 boulevard du 11 novembre 1918
69622 Villeurbanne Cédex France
cbasileu@yahoo.fr,
{bounekkar, kabachi, lamure, mondher.toumi}@univ-lyon1.fr
http://recherche.univ-lyon2.fr/eric
**Laboratoire ERIC - University of Lyon - EPHE
41 rue GayLussac,75005 Paris
Marc.Bui@ephe.sorbonne.fr
s.benamor@gmail.com
http://recherche.univ-lyon2.fr/eric

Abstract. Social networks are quite often used in social sciences to model the links between individuals. Several problems are analyzed in their light: what nodes are of importance regarding connectivity, what diffusion level for given nodes? Graph theory and more specifically random graphs are the basis of the most of works. However, a social network generally consists of various types of links between the individuals, what implies to use a more adequate model than graph theory. So, our proposal is to work with families of graphs which first enable us to study the topological properties of a network by means of a version weakened by the topology, called pretopology and second enable us handle uncontrolled factors which may influence the network structure. Indeed, in real world, these events are not predictable, we thus introduce stochastic aspects in our model. This leads us to the notion of stochastic network. In this paper, we give at first basic definitions of a network, of a stochastic network and some elements on the mathematical theory of the pretopology and random sets. Then, we give first results concerning the modeling of a stochastic network, with some particular points concerning topological analysis. We terminate with perspectives for the future works.

1 Introduction

Social sciences widely use the concept of network for modeling links between individuals. (Social networks). The references of Andersson (1999) and this one of Sattenspiel and Simon (1988) show the interest for the social network in the spread of disease by using percolation

theory to model epidemics. A lot of works have been proposed on different problems related to this concept. For example, what percentage of nodes of the network is required to observe a significant change in the connectivity of the network? Or, what are nodes the most linked to others? The common point of these Works is that they all model a social network by a graph, in the mathematical sense (see Albert and Barabasi (2002), Newman (2003), Barabasi et al. (2004)). However, we think a social network is composed from different types of links between individuals, which implies using more than one graph. So we propose a new formalism for modeling social networks, based on families of graphs. This formalism enables to study topological properties of a network by means of a weakened version of mathematical topology (pretopology, see Belmandt (1994), Auray et al. (2009), Dalud Vincent et al. (2007), Lamure et al. (2009)). As a second point, we assume that links between individuals cannot be considered as certain and that many events can influence existence of these links. In most cases, these events are not predictable, so we introduce stochastic aspects in modeling a network. This leads us to the concept of stochastic network. In this paper, we first give basic definitions of a network, of a stochastic network and some elements about underlying mathematical theories: pretopology and random sets. Then, we give first results about modeling the topology of a stochastic network, with some particular points concerning transmission and connectivity. We conclude with perspectives of further works.

2 Basic definitions

Given a finite population E, with n individuals, given a probability space $(\Omega, \mathcal{A}, p)$, we consider the following operator $\mathcal{R}(.)$ defined as:
$\mathcal{R}(.) : (\Omega, \mathcal{A}, p) \longmapsto \mathcal{R}(E)$
where $\mathcal{R}(E)$ denotes the set of all binary relationships (graphs) on E. By definition, $\mathcal{R}(E)$ is a family of subsets of $E \times E$. So, we assume that $\mathcal{R}(.)$ is a random set, i.e. a measurable correspondence from $(\Omega, \mathcal{A}, p)$ into $E \times E$ (see Debreu (1967), Matheron (1975), Lamure (1978)).

Definition 1. $\mathcal{R}(.)$ is called a stochastic graph operator.

Definition 2. We define a network as a family $\{\mathcal{R}_i\}_{i=1,..,p}$ of binary relationships on E.

Without loss of generality, we can assume that relationships $\mathcal{R}_i$ are reflexive ones.
Example :
Let us consider $E = \{a, b, c, d, e\}$ and three relationships $\mathcal{R}_1$, $\mathcal{R}_2$ and $\mathcal{R}_3$, described by the following diagram (see figure 1). For convenience, in figure 1, loops on any node are omitted.
Now, we can give definition of a random network. For that, we can note that any binary relationship $\mathcal{R}$ is characterized by the following family:

$$\forall x \in E, \Gamma(x) = \{y \in E / x \mathcal{R} y\}$$

which is the set of "children" of x in E.
In case of a random network, this is expressed in a more complicated way, but based on the same principle. We consider a family $\{\mathcal{R}_i(.)\}_{i=1,..,p}$ of stochastic graph operators defined on

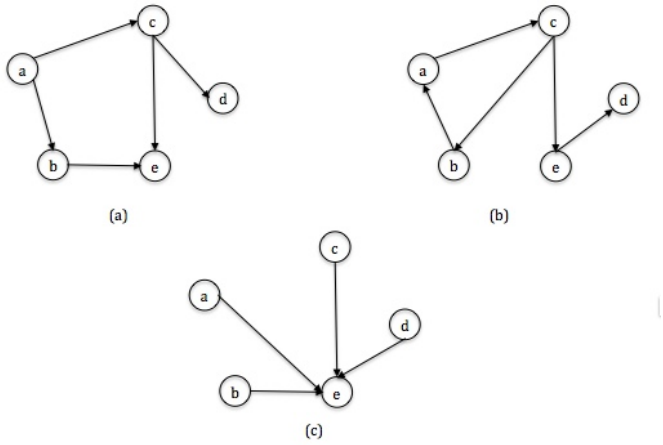

FIG. 1 – *A simple Network*

a probability space $(\Omega, \mathcal{A}, p)$. Let us consider:

$$\forall x \in E, \forall \omega \in \Omega, \forall i = 1, .., p, \Gamma_i(\omega, x) = \{y \in E / x \mathcal{R}_i(\omega) y\}$$

The family $\{\Gamma_i(., x), x \in E\}_{i=1,..,p}$ characterizes $\{\mathcal{R}_i(.)\}_{i=1,..,p}$. If we assume that:

$$\forall x \in E, \Gamma_i(., x) : (\Omega, \mathcal{A}, p) \longmapsto \mathcal{P}(E)$$

is a random correspondence, then $\{\mathcal{R}_i(.)\}_{i=1,..,p}$ is a family of stochastic graph operators and we put:

Definition 3. A stochastic network is a family $\{\mathcal{R}_i(.)\}_{i=1,..,p}$ of stochastic graph operators

Example
Let us consider $\Omega = \{\omega_1, \omega_2, \omega_3\}$, $\mathcal{A} = \mathcal{P}(\Omega)$ and p defined by $p(\omega_i) = \frac{1}{3}, \forall i = 1, 2, 3$
Let us consider: Let us consider $E = \{a, b, c, d\}$ and:
$\Gamma_1(\omega_1, a) = \{a\}, \Gamma_1(\omega_1, b) = \{b, c\}, \Gamma_1(\omega_1, c) = \{a, b, c\}, \Gamma_1(\omega_1, d) = \{a, d\}$
$\Gamma_2(\omega_1, a) = \{a, c\}, \Gamma_2(\omega_1, b) = \{b, d\}, \Gamma_2(\omega_1, c) = \{b, c\}, \Gamma_2(\omega_1, d) = \{a, b, d\}$
$\Gamma_3(\omega_1, a) = \{a, b, d\}, \Gamma_3(\omega_1, b) = \{b, d\}, \Gamma_3(\omega_1, c) = \{a, c, d\}, \Gamma_3(\omega_1, d) = \{a, c, d\}$
$\Gamma_1(\omega_2, a) = \{a\}, \Gamma_1(\omega_2, b) = \{b, c\}, \Gamma_1(\omega_2, c) = \{a, b, c\}, \Gamma_1(\omega_2, d) = \{a, d\}$
$\Gamma_2(\omega_2, a) = \{a\}, \Gamma_2(\omega_2, b) = \{b, d\}, \Gamma_2(\omega_2, c) = \{c\}, \Gamma_2(\omega_2, d) = \{d\}$
$\Gamma_3(\omega_2, a) = \{a, b\}, \Gamma_3(\omega_2, b) = \{b, d\}, \Gamma_3(\omega_2, c) = \{a, c\}, \Gamma_3(\omega_2, d) = \{c, d\}$
$\Gamma_1(\omega_3, a) = \{a\}, \Gamma_1(\omega_3, b) = \{b\}, \Gamma_1(\omega_3, c) = \{c, d\}, \Gamma_1(\omega_3, d) = \{b, d\}$
$\Gamma_2(\omega_3, a) = \{a, c\}, \Gamma_2(\omega_3, b) = \{b, d\}, \Gamma_2(\omega_3, c) = \{c\}, \Gamma_2(\omega_3, d) = \{c, d\}$
$\Gamma_3(\omega_3, a) = \{a, b\}, \Gamma_3(\omega_3, b) = \{b, d\}, \Gamma_3(\omega_3, c) = \{c\}, \Gamma_3(\omega_3, d) = \{a, c, d\}$
We get a stochastic network which is described by figure 2, figure 3 and figure 4 (loops on nodes are omitted) Thus, a stochastic network is fully defined by a family of random correspondences (or random sets) which each give the set of "children" of any element of the reference set E, under random events driven by a probability law.

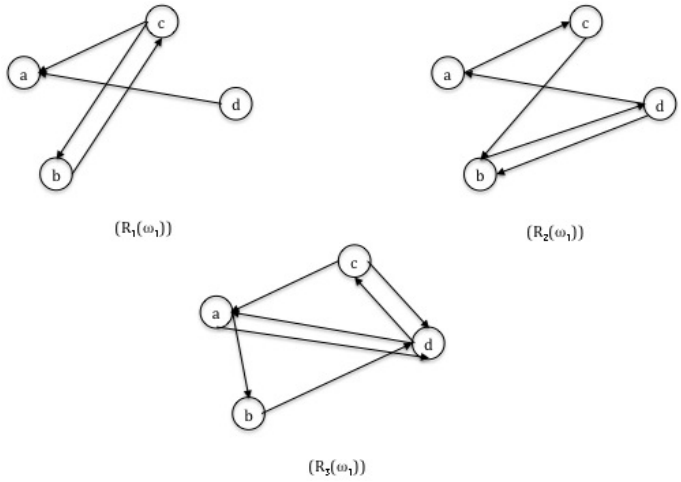

FIG. 2 – *A stochastic network, scenario* ω_1

At this point, a question is to know how it is possible to model the structure of the network and to know what are its properties regrading the stochastic aspect; That means we have to:

- define a pretopological structure onto the network,
- analyze the properties of this structure regarding measurability and integrability concepts.

3 Pretopology on a stochastic network

Basic concepts of pretopology and of random correspondences are respectively recalled in annex 1 and annex 2. In this section, we will define a pretopological structure on a stochastic network and list its basic properties.
Let E be a finite set. Let $\mathcal{R}_i(.)$ a stochastic network defined on a probability space $(\Omega, \mathcal{A}, p)$. Let us consider, for any subset A of E, the function $a(.,.)$ defined by:
$a(.,.) : (\Omega, \mathcal{A}, p).\mathcal{P}(E) \longrightarrow E$ such as:
$a(\omega, A) = \{x \in E / \forall i, \Gamma_i(\omega, x) \cap A \neq \emptyset\}$
where $\Gamma_i(\omega, x) = \{y \in E / x\mathcal{R}_i(\omega)y\}$ Then:

Theorem 4. *$(E, a(.,.))$ with $a(.,.)$ as previously defined is a stochastic pretopological space.*

Proof. Let $x \in E$, we put $\phi_i(\omega, x) = 1$ if $\Gamma_i(\omega, x) \cap A \neq \emptyset$ and $\phi_i(\omega, x) = 0$ otherwise. So, $\{x \in E / \Gamma_i(\omega, x) \cap A \neq \emptyset\} = \{x \in E / \phi_i(\omega, x) = 1\}$.As for any i and any $x \in E$, $\Gamma_i(., x)$ is a random correspondence, the function $\phi_i(., x)$ is a random variable for any $x \in E$ and then $\omega \longrightarrow a_i(\omega, A) = \{x \in E / \Gamma_i(\omega, x) \cap A \neq \emptyset\}$ is a random correspondence for any $A \subset E$.

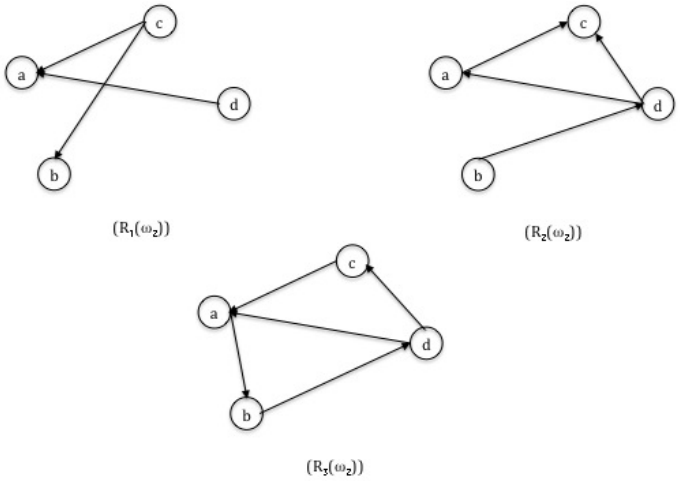

FIG. 3 – *A stochastic network, scenario* ω_2

As for any $A \subset E$, $a(\omega, A)$ is a finite union of $a_i(\omega, A)$, the correspondence $\omega \longrightarrow a(\omega, A)$ is a random correspondence. So, $(E, a(., .))$ is a stochastic pretopology. □

According to definition of $a(., .)$ we have $x \in a(\omega, A) \Leftrightarrow \forall i, \Gamma_i(\omega, x) \cap A \neq \emptyset$.
In other words, $x \in a(\omega, A) \Leftrightarrow \forall i, \exists y \in A, x\mathcal{R}_i(\omega)y$. So $x \in a(\omega, A)$ means that, for any kind of relationship, ther exists a link between x and at least one element of A. $\parallel a(\omega, A) \parallel$ then is a good indicator of influence of A in the network in the sense the greater it is, the greater is the number of elements outside A linked to, at least, one element of A, whatever the nature of the link.
By definition (see annex 1) $\forall A \subset E, \forall \omega \in \Omega, i(\omega, A) = (a(w, A^c))^c$. Then, $i(\omega, A)$ is the subset of elements of A for which it is possible to fond out at least one relationship such as all children of x are in A, so $\parallel i(\omega, A) \parallel$ also is a good indicator. The greater it is, the greater is the number of elements of A for which we can find out at least one relationship for which their children are in A. This leads us to the following:

Definition 5. We call pseudoclosure ratio, the quantity

$$pcr(\omega, A) = \frac{\parallel a(\omega, A) \parallel}{\parallel A \parallel}$$

We call interior ratio the quantity

$$ir(\omega, A) = \frac{\parallel i(\omega, A) \parallel}{\parallel A \parallel}$$

Then:

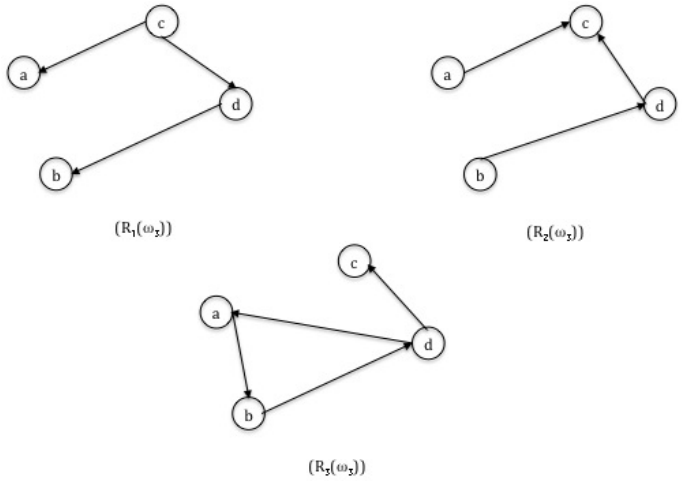

FIG. 4 – *A stochastic network, scenario* ω_3

Theorem 6. $\forall A \subset E, \omega \to pcr(\omega, A)$ *is a random variable*
$\forall A \subset E, \omega \to ir(\omega, A)$ *is a random variable*

Proof. It is sufficient to note that $\parallel a(\omega, A) \parallel = \sum_{x \in E} 1_{a(\omega,A)}(x)$ and that, as $a(., A)$ is a random correspondence, $1_{a(\omega,A)}(x)$ is a random variable. □

Obviously, there is a strong link between stochastic networks and random graphs. In fact, in cases where there is only one relationship in the network, we are faced to a random graph; The difference between our approach and usual approaches is that pretopology and stochastic pretopology enables us to provide a topological analysis of the network based on concepts fully adapted to discrete spaces. Another advantage is to be able to compute statistics, to perform statistical analysis on indicators such as $pcr(.,.)$ and $ir(.,.)$ and to use new concepts of connectivity defined in the framework of pretopology.

4 Conclusion

In this paper, we give first concepts on stochastic networks, as an extension of random graphs, with a new approach mixing pretopology and theory of random sets. New indicators of structure of the network can be defined, as $pcr(.,.)$ and $ir(.,.)$. With pretopology, we can use different types of connectivity, fully adapted to graphs: hyperconnectivity, ultraconnectivity, apoconnectivity,... Each of them gives a specific information about the structure. Another aspect of future works is related to probability computation. In brief, this approach leads to a lot of works with applications in various fields: social sciences, computer sciences, epidemiology,...

5 Annex 1: Basics on pretopology

Let's consider a non empty set E. We define a function $a(.)$ from $\mathcal{P}(E)$ into itself such as:
$(P_1)\ a(\emptyset) = \emptyset$
$(P_2)\ \forall A, A \subset E, A \subset a(A)$
Then, the couple $(E, a(.))$ is called "pretopological space".
As in topology, we can define the interior function $i(.)$ by putting:
$\forall A \subset E, i(A) = (a(A^c))^c$ where A^c denotes the complementary of A in E. Thus, related to usual concepts of topology, we only keep two first properties of the topological closure mapping. Such a function $a(.)$ is called pseudoclosure function in pretopology.

5.1 Different pretopological spaces

A basic pretopological space $(E, a(.))$ is such as:
$(P_1)\ a(\emptyset) = \emptyset$
$(P_2)\ \forall A, A \subset E, A \subset a(A)$

5.1.1 $\mathcal{V}$ type space

Let us consider the following axiom:
$(P_3)\ \forall A, A \subset E, \forall B, A \subset E, A \subset B \Rightarrow a(A) \subset a(B)$

Definition 7. if $a(.)$ fulfills P_1, P_2 and P_3, we say that $(E, a(.))$ is a $\mathcal{V}$ type space.

In this case, the concept of neighborhood becomes a quit interesting one. In pretopology, this concept is defined in the same way as in topology.

Definition 8. Let $(E, a(.))$ be a $\mathcal{V}$ type space. Any subset V of E is said a neighborhood of x, $x \in E$ if and only if $x \in i(V)$.

However, in pretopology, the family $\mathcal{V}(x)$ of neighborhoods of any x does not fulfills the same properties. In fact, generally speaking, the only thing we can say is that $\mathcal{V}(x)$ is a prefilter of subsets of E, i.e.:
$\forall x \in E, \emptyset \notin \mathcal{V}(x)$
$\forall x \in E, \forall V \in \mathcal{V}(x), \forall W \subset E, V \subset W \Longrightarrow W \in \mathcal{V}(x)$

5.1.2 $\mathcal{V}_D$ type space

Let us consider the following axiom:
$(P_4)\ \forall A, A \subset E, \forall B, A \subset E, a(A \cup B) = a(A) \cup a(B)$

Definition 9. if $a(.)$ fulfills P_1, P_2 and P_4, we say that $(E, a(.))$ is a $\mathcal{V}_D$ type space.

Obviously, if $(E, a(.))$ is a $\mathcal{V}_D$ type space, it also is a $\mathcal{V}$ type space.
And the family of neighborhoods of any x in E is a filter, i.e. $\mathcal{V}(x)$ is a prefilter and satisfies the following property:
$\forall V \in \mathcal{V}(x), \forall W \in \mathcal{V}(x), V \cap W \in \mathcal{V}(x)$.

5.1.3 $\mathcal{V}_s$ type space

Let us consider the following axiom:
(P_5)

$$\forall A, A \subset E, a(A) = \bigcup_{x \in A} a(\{x\})$$

Definition 10. if $a(.)$ fulfills P_1, P_2 and P_5, we say that $(E, a(.))$ is a $\mathcal{V}_D$ type space.

Clearly, if $(E, a(.))$ is a $\mathcal{V}_s$ type space, it also is a $\mathcal{V}_D$ type space and then a $\mathcal{V}$ type space. Moreover, the family of neighborhoods of x satisfies the following property

$$\bigcap_{V \in \mathcal{V}(x)} \in \mathcal{V}(x)$$

This last property is interesting from a computational point of view as it implies it is sufficient to compute pseudoclosure of singletons of E to get pseudoclosure of any subset of E.

6 Annex 2: Basics on random correspondences

Three concepts of measurabilty have been defined Lamure (1978).

6.0.4 Definition I

Let us consider a measurable space $(\Omega, \mathcal{A})$ and a correspondence Γ into $\mathbf{R}^n$. Γ is assumed a non empty compact valued correspondence. We also suppose that Ω is locally compact and $\mathcal{A}$ is defined as follows:
Starting from $\mathcal{B}$ the σ-algebra of borelians of Ω, we complete it to obtain $\mathcal{B}_p$ (p being the probability on $\mathcal{B}$) and we consider:

$$\mathcal{A} = \{A, A \subset \Omega / A \cap K \in \mathcal{B}_p, \forall K \in \mathcal{K}(\mathbf{R}^n)\}$$

where $\mathcal{K}(\mathbf{R}^n)$ is the family of compacts of Ω.

Definition 11. Let us consider $(\Omega, \mathcal{A})$ a measurable space, $\Gamma \longmapsto \mathbf{R}^n$. We say that Γ is measurable in the sense I if and only if for all F, closed subset of $\mathbf{R}^n$,

$$A = \{\omega \in \Omega : \Gamma(\omega) \cap F \neq \emptyset\} \in \mathcal{A}$$

We can note this definition can be rewritten as follows: for any O open subset of IR^n,

$$B = \{\omega \in \Omega : \Gamma(\omega) \subset O\} \in \mathcal{A}$$

6.0.5 Definition II

In this subsection, $G(\Gamma)$ denotes the graph of the correspondence Γ, i.e. $G(\Gamma) = \{(\omega, x) \in \Omega \times \mathbf{R}^n / x \in \Gamma(\omega)\}$ and $\mathcal{B}_n$ denotes the σ-algebra of borelians of $\mathbf{R}^n$.

Definition 12. Let us consider $(\Omega, \mathcal{A})$ a measurable space, $\Gamma \longmapsto \mathbf{R}^n$. We say that Γ is measurable in the sense II if and only if $G(\Gamma) \in \mathcal{A} \otimes \mathcal{B}_n$

6.0.6 Definition III

As correspondences are compact valued, a third proposition can be proposed. Let us consider the following families:

$$\mathcal{U}^w = \{K, K \in \mathcal{K}(\mathbf{R}^n) / K \cap U \neq \emptyset, \forall U \in \mathcal{O}(\mathbf{R}^n)\}$$

$$\mathcal{U}^s = \{K, K \in \mathcal{K}(\mathbf{R}^n) / K \subset U, \forall U \in \mathcal{O}(\mathbf{R}^n)\}$$

where $\mathcal{O}(\mathbf{R}^n)$ denotes the family of open subsets of $\mathbf{R}^n$.
These two families define a topology $\mathcal{T}$ on $\mathcal{K}(\mathbf{R}^n)$ which is equivalent to the topology generated by the Hausdorff metric. Thus $\mathcal{K}(\mathbf{R}^n)$ also is a separable metric space. Let us consider Σ_n the σ-algebra of borelians of $\mathcal{K}(\mathbf{R}^n)$. Γ can be considered not as a correspondence from Ω into $\mathbf{R}^n$ but as a function form Ω into $\mathcal{K}(\mathbf{R}^n)$. It is possible to consider for Γ, the usual definition of measurability for functions.

Definition 13. Let us consider $(\Omega, \mathcal{A})$ a measurable space, $\Gamma \longmapsto \mathbf{R}^n$. We say that Γ is measurable in the sense III if and only if

$$\forall A \in \Sigma_n, \Gamma^{-1}(A) \in \mathcal{A}$$

where

$$\Gamma^{-1}(A) = \{\omega \in \Omega / \Gamma(\omega) \in A\}$$

We get the following result:

Theorem 14. *Let us consider $(\Omega, \mathcal{A})$ a complete measurable space, Ω being locally compact, $\Gamma \longmapsto \mathbf{R}^n$. The three definitions are equivalent ones.*

Proof. First, lest us prove equivalence of definitions of I and II. For that, we use the following result.
Let $(\Omega, \mathcal{A}, p)$ a complete measurable space, Let E a complete metric separable space and Γ a correspondence defined upon Ω, valued in the family of closes subsets of E, then

$$G(\Gamma) \in \mathcal{A} \otimes \sigma(E) \Leftrightarrow \{\omega \in \Omega / \Gamma(\omega) \cap F \neq \emptyset\} \in \mathcal{A}$$

where $\sigma(E)$ denotes the σ-algebra of borelians of E. As $\mathbf{R}^n$ and Γ verify properties of this result, definitions I and II are equivalent.
Now, let us suppose Γ measurable according to definition III and let us consider, for any closed subset F of $\mathbf{R}^n$, the set $A = \{\omega \in \Omega / \Gamma(\omega) \cap F \neq \emptyset\}$.
$A = \{\omega \in \Omega / \Gamma(\omega) \subset F^c\}^c$, where F^c denotes the complementary of F in $\mathbf{R}^n$. F^c is an open subset of $\mathbf{R}^n$ and $A = \Gamma^{-1}((F^c)^s)$. As Γ is measurable according to definition III, $\Gamma^{-1}((F^c)^s) \in \mathcal{A}$ and Γ is measurable according to definition I.
To prove that definition I implies definition III, it is sufficient using the following result:
$(\Omega, \mathcal{A})$ is a measurable space, Ω is locally compact, if f is a function from $(\Omega, \mathcal{A})$ in E, E is a separable metric space endowed with its borelians, then f measurable is equivalent to f p-measurable.
This result is applied to Γ as a function from $(\Omega, \mathcal{A})$ into $\mathcal{K}(\mathbf{R}^n)$. This leads to the result. □

References

Albert, R. and A. L. Barabasi (January 2002). Statistical mechanics of complex networks. *Review of modern physics 74*.

Andersson, H. (1999). Epidemic models and social networks. *The Mathematical Scientist 24*, 128–147.

Auray, J. P., S. Bonnevay, M. Bui, G. Duru, and M. Lamure (2009). Prétopologie et applications : un état de l'art. *Studia Informatica Universalis 7.1*, 25–44.

Barabasi, A. L., A. Vasquez, R. Dobrin, D. Sergi, J. P. Eckmann, and N. Oltval (December 2004). The topological relationship between the large-scale attributes and local interaction patterns of complex networks. *PNAS 101*, 52.

Belmandt, Z. (1994). *Manuel de prétopologie et applications*. Editions Hermés.

Dalud Vincent, M., M. Brissaud, and M. Lamure (2007). Closed sets and closures in pretopology. *International Journal of Applied Mathematics*.

Debreu, G. (1967). *Integration of correspondences*. Fifth symposium of barkeley.

Lamure, M. (1978). *Contribution á la théorie de la multiestimation*. Vandenhœck & Ruprecht Edition.

Lamure, M., S. Bonnevay, M. Bui, and S. Benamor (2009). Modélisation et simulation de la pollution urbaine - un modèle aléatoire, prétopologique pour la ville de ouagadougou. *Santé et systémique 10 numéro 3-4/2007*, 75–87.

Matheron, G. (1975). *Random sets and integral geometry*. New york: J. Wiley editions.

Newman, M. E. J. (2003). The structure and function of complex networks.

Sattenspiel, L. and C. P. Simon (1988). The spread and persistence of infectious diseases in structured populations. *Math Biosci 90*, 341–366.

Résumé

Les réseaux sociaux sont souvent utilisés pour modéliser les relations entre individus. plusieurs problèmes sont examinés à leur lumière : quels sont les noeuds importants du point de vue de la connexité ? quel est le niveau de diffusion d'un noeud ? La théorie des graphes, plus particulièrement les graphes aléatoires sont à la base des travaux du domaine. Cependant, un réseau social est généralement constitué d'une famille de relations entre individus, ce qui implique l'emploi d'un modèle plus approprié que celui de la théorie des graphes. Notre proposition est de travailler avec des familles de graphes de manière à en étudier les propriétés topologiques au moyen de la prétopologie d'une part et également de prendre en compte des facteurs incontrôlables. Dans les faits, ces derniers influencent les relations entre individus et ne sont pas prévisibles. Nous introduisons donc le concept de réseau stochastique. Dans ce travail, nous posons les définitions de base des réseaux stochastiques, ceux de prétopologie et ceux relatifs aux ensembles aléatoires. Nous donnons de premiers résultats, notamment du point de vue topologique et terminons par les perspectives de travaux futurs.

Une architecture multi-agents pour la découverte et la construction de profils utilisateurs distribués

Anis Chouchane, Amel Bouzeghoub

Institut TELECOM SudParis, Département Informatique
9, rue Charles Fourier, 91011 Evry, France
{Anis.Chouchane, Amel.Bouzeghoub}@it-sudparis.eu

Résumé. Nous décrivons dans ce papier les techniques de découverte et de construction de profils utilisateurs distribués afin de leur proposer des services adaptés à leurs besoins dans le cadre d'une application d'apprentissage pervasif. Vu l'influence des données émergentes sur le web, notre système traite du problème de distribution des informations du profil utilisateur auquel font face actuellement les systèmes de profils inférés. L'objectif de ce travail est double. Il s'agit d'une part de proposer un modèle pour la gestion du profil utilisateur et d'autre part de proposer une architecture pour la découverte et la construction du profil dans un contexte mobile. Un prototype a été implémenté permettant à un service donné de proposer des recommandations à un apprenant adaptées à son profil. Ce dernier étant construit dynamiquement et à la demande.

Mots-clés : apprentissage pervasif, profil utilisateur distribué, systèmes multi-agents, web sémantique.

1 Introduction

Un nouveau concept a émergé ces dernières années pour traduire le potentiel de l'informatique ubiquitaire dans le domaine de l'apprentissage. Cette nouvelle façon d'utiliser des technologies pour soutenir les processus d'apprentissage est appelée "Apprentissage pervasif". Cette évolution s'intensifie ces dernières années avec l'émergence des terminaux mobiles et ultra-mobiles (ex : ordinateurs portables, téléphones mobiles, Pocket PC, PDA) et des réseaux mobiles (GSM, 3G+, réseaux sans fil, Bluetooth, etc.). L'apprentissage pervasif utilise ces nouvelles technologies comme support pour améliorer l'apprentissage traditionnel et élargir les perspectives du processus d'apprentissage lui-même. L'objectif principal dans un environnement d'apprentissage pervasif est de fournir aux apprenants la bonne ressource au bon moment et de la meilleure façon. En outre, il offre un champ d'application en expansion qui peut accroitre la diffusion des Nouvelles Technologies de l'Information et de la Communication. Toutefois, cela ne se réalisera que si l'on peut proposer facilement des services adaptés et simples d'utilisation. Pour cela, il est nécessaire de pouvoir adapter les services, les documents multimédia au contexte d'utilisation ainsi qu'à l'utilisateur. Dans ce contexte, le profil apprenant est un critère fondamental. Il existe un certain nombre de modèles de profil et d'architectures de gestion du profil. Mais les travaux dans ce domaine

n'ont abouti que dans les systèmes de profil centralisés. La modélisation des profils distribués est encore un point à étudier, notamment dans un environnement mobile. Ainsi deux questions se posent. Comment stocker les données du profil utilisateur de la meilleure façon ? Et quelle architecture adopter pour intégrer plusieurs profils utilisateurs dans un environnement mobile ?

Dans ce travail, nous prenons le terme d'apprentissage dans son sens le plus large. Par exemple, un utilisateur, au cours d'une activité d'apprentissage, peut effectuer une recherche sur le web, où différents services stockent une partie de son profil. Nous partons de l'hypothèse qu'un profil utilisateur est stocké en différents points du réseau, par des services tels que Facebook, Flickr, Amazon, etc. Le profil utilisateur est donc fragmenté. Ceci implique la nécessité du partage et de l'échange des fragments du profil entre les différents services. De plus, les données du profil, tout comme l'identification des fragments du profil sont utiles pour les services d'adaptation et de recommandation. Plus particulièrement, le scénario que nous avons adopté est le suivant : un service de recommandation souhaite récupérer le profil de l'utilisateur afin de lui proposer un service. L'utilisateur lui donne son identité dans le système de gestion de profil. Le service de recommandation se connecte au système de gestion de profil et demande le profil correspondant à cette identité. Le système de gestion de profil se base sur les paramètres configurés par l'utilisateur pour générer le profil et le renvoie au service de recommandation.
Le plan de cet article est décrit comme suit : un état de l'art est présenté dans la section suivante. Il s'agit d'étudier les différents standards existants pour la modélisation du profil utilisateur, ainsi que les travaux de recherche portant sur la gestion du profil dans les systèmes d'adaptation et de personnalisation. En se basant sur l'analyse de cet état de l'art, nous proposons dans la section 3 un modèle de profil dans un contexte mobile. Ce dernier sera mis en œuvre dans la section 4 qui décrit l'architecture multi-agents permettant de gérer efficacement les profils utilisateurs distribués. Nous citons dans la section 5 les caractéristiques de notre approche. Enfin, l'article se termine par une conclusion et des perspectives dans la section 6.

2 Etat de l'art

Proposer un modèle pour la gestion des profils utilisateurs dans un contexte mobile nécessite d'analyser les différents standards existants sur la structuration des données du profil. De la même manière, définir une architecture pour l'intégration du profil apprenant requiert une étude des différents travaux de recherche portant sur l'intégration du profil dans les systèmes d'adaptation et de personnalisation. Ainsi, cet état de l'art comprend deux parties correspondant à ces deux objectifs.

Un profil est un modèle utilisateur ou source de connaissance qui contient des acquisitions sur tous les aspects de l'utilisateur pouvant être utiles pour le comportement du système. Outre les informations d'identification de base, le profil utilisateur peut regrouper des informations très diverses selon les besoins. Parmi celles-ci, Jameson (1999) propose:

- Des caractéristiques personnelles pouvant influencer fortement l'interaction (âge, sexe, etc.).

- Les intérêts et les préférences générales relatives à la tâche à accomplir, qui permettent une adaptation aux attentes de l'utilisateur.
- Les compétences ou le niveau d'expertise relatifs à la tâche (pour déterminer par exemple un degré d'autonomie et déceler un besoin d'aide ou de formation).
- Le but courant de l'utilisateur.

Sur les sites web, le profil utilisateur est souvent assimilé à un curriculum vitae court, avec (ou non) une photo et quelques informations statistiques. Mais dans les services de réseaux sociaux en ligne tels que Facebook, Google profile, Linkedin… un profil peut être plus compliqué puisque l'utilisateur a la possibilité de décrire son identité, ses intérêts, ses préférences, ses compétences...

Les standards les plus importants sont PAPI (Public And Private Information) et IMS LIPS (Learner Information Package Specification). Ces standards, développés par le W3C, permettent le partage du modèle apprenant dans un environnement d'apprentissage.

PAPI Learner (2000) est un standard proposé par le groupe Learner Model Working Group de l'IEEE, qui décrit les informations sur l'apprenant utiles pour la communication entre les systèmes coopératifs. Il se focalise sur les performances et les interactions entre apprenants.
PAPI décompose le profil en six catégories : *Informations personnelles, Relations, Sécurité, Préférences, Performances, Portfolio.*

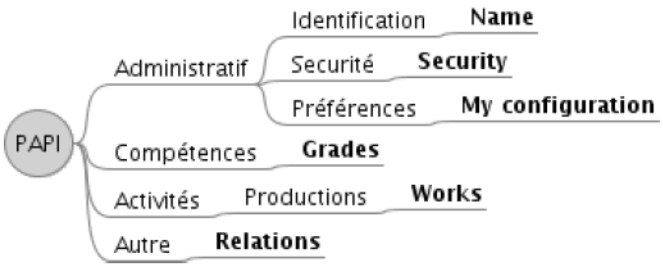

Figure 1 : Eléments de PAPI

IMS LIP (2001) est une spécification décrivant une approche classique de CV structuré. Elle se focalise sur l'historique de l'apprenant et de son expérience d'apprentissage. Le but de ce standard est de faciliter l'échange des informations sur les apprenants entre systèmes éducatifs, systèmes de gestion d'apprentissage, etc. IMS LIP est structurée en onze catégories de base : *Identification, But, Qualifications, Certifications & licences (QCL), Activité, intérêts, Relations, Compétences, Accessibilité, Transcription, Affiliation, Sécurité.*

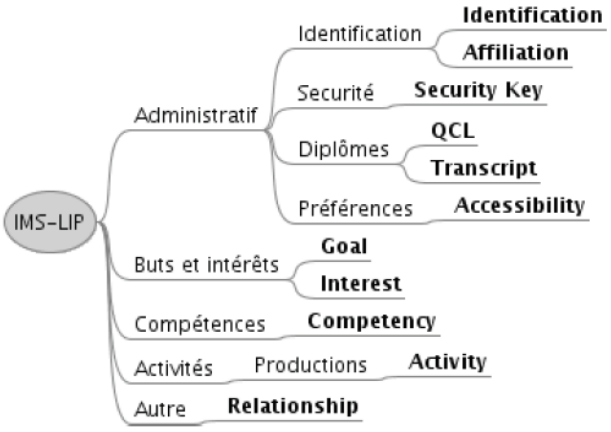

Figure 2 : Eléments d'IMS LIP

FOAF (*Friend Of A Friend*) (2007) est un vocabulaire basé sur RDF, défini dans le cadre d'un projet open source, permettant de décrire des personnes et les relations qu'elles entretiennent entre elles. Il a été développé pour la construction de groupes sociaux. FOAF distingue 5 catégories pour décrire un profil : FOAF *Basics* comprend la description de base comme le nom, l'adresse e-mail, les images. *Personal Information* décrit plus d'informations personnelles telles que le blog, les intérêts, les publications et les relations aux autres profils qui connaissent cette personne. *Online Accounts* décrit les informations sur les comptes qu'une personne possède. *Projets and Groups* définit les informations sur les projets, les groupes ou les organisations dans lesquelles la personne est membre. *Documents and Images* décrit les documents et les images relatifs à l'apprenant, par exemple: document de profil, logo…

La figure 3 ci-dessous représente un tableau mettant en évidence la comparaison entre les modèles apprenant décrits précédemment. Dans ce tableau, la représentation de la taxonomie a été simplifiée, en gardant les grandes catégories et les sous catégories.

Annotation utilisée dans le tableau :

+ : support total p : support partiel x : capacité à être étendu

Catégories	Sous-catégories	IMS LIP	PAPI	FOAF
Données personnelles		+	+	+
Relations	A d'autres profils		+	+
	Référence aux autres			+
	Groupe / Org description	p		+
But		+		
Réalisations et Historique	Activité	+	p	
	Compétences	+	+	
	Certification	+	+	
	Transcript	+	x	
Intérêts		+		+
Accessibilité	Langue	+		+
	Style d'apprentissage	+		
	Eligibilité	+		
	Incapacité	+		+
Sécurité		+	+	x
Systèmes éducatifs		+	+	

Figure 3 : Tableau comparatif des standards du modèle utilisateur

D'après ce tableau comparatif, nous pouvons déduire que FOAF possède l'avantage de prendre en compte tous les types de relations entre les profils. Néanmoins, il se trouve que tous les modèles étudiés, même s'ils supportent tous la description des données personnelles de l'apprenant, ils ne décrivent pas le contexte courant de l'apprenant, alors que ce type d'informations est essentiel dans un environnement ubiquitaire, surtout dans l'apprentissage mobile. Par conséquent, le modèle apprenant attendu qui peut être adapté à l'apprentissage mobile doit avoir tous les avantages des normes plus des informations sur le contexte de l'utilisateur.

La deuxième partie de l'état de l'art concerne l'étude des travaux de recherche portant sur l'intégration du profil apprenant afin de proposer ensuite une architecture convenable pour mettre en œuvre le profil apprenant. Cette étude nous a permis de distinguer deux types de modélisation : centralisée et distribuée. Différents scénarios ne sont pas adaptés à l'architecture centralisée : notamment dans l'informatique ubiquitaire, où l'utilisateur dispose d'informations en différents points sur le réseau. Bien que ce type de modélisation permette d'assurer la cohérence des données de l'utilisateur, il nécessite une représentation standard des données de l'utilisateur : toutes les applications doivent partager le même schéma de métadonnées. De plus, les applications utilisent seulement un fragment des données du profil stocké sur le serveur centralisé. Enfin, les données du profil se trouvent hors du contexte dans lequel il a été récupéré, les données peuvent être interprétées différemment dans un autre contexte.

Certains travaux de modélisation centralisée se basent sur les Web Services. Dans D.L. Musa et J.P.M de Oliveira (2005), les auteurs proposent une architecture pour l'intégration du profil apprenant en utilisant les Web Services. L'objectif est d'assurer la coopération entre différents systèmes d'apprentissage, comme la plateforme de formation à distance *Claroline*, afin d'obtenir un modèle apprenant plus riche. Les services de l'application proposée permettent de gérer la confidentialité des données du profil en utilisant la norme P3P.

K. Kabassi et Maria Virvou, (2003) décrivent un système d'apprentissage personnalisé Web *F-SMILE* (Web File-Store Manipulation Intelligent Learning Environment). Ce système déploie deux modèles pour l'apprenant : un modèle stocké localement sur son PC accessible via une application locale, et un deuxième modèle stocké sur le serveur et accessible via une application Web. L'inconvénient dans cette approche est que l'apprenant peut avoir plusieurs dispositifs, ce qui nécessite de déployer l'application sur autant de dispositifs qu'il possède. Par ailleurs, vu que F-SMILE stocke deux profils apprenants, le système peut rencontrer des problèmes de connexions entre le PC de l'apprenant et le serveur.

Dans Andreas von Hessling et al. (2005) une architecture dans un environnement P2P a été proposée. Parmi les avantages de cette approche, il n'y pas besoin de serveur central vu qu'il s'agit d'une architecture décentralisée. Il y a juste besoin d'une connexion wifi qui permet de se connecter à des services proches. Toutes les données de l'utilisateur sont stockées sur son propre dispositif mobile. Ceci permet une meilleure gestion de la confidentialité des données de l'utilisateur, qui a le contrôle sur ses informations. Toutefois, cette application est définie dans un périmètre limité à la zone wifi dans laquelle se trouvent les services auxquels l'utilisateur peut accéder (exemple : salles de cinéma), ce qui réduit le contexte de mobilité de ce dernier. La modélisation du profil établie se résume à un ensemble d'intérêts et de désintérêts, et donc ne prend pas en compte d'autres informations qui permettent de décrire le profil de manière exhaustive.

Différents travaux ont été effectués dans le cadre des systèmes adaptatifs éducatifs. Dans ce cadre, Mohammad Alrifai et al. (2006) proposent des solutions au problème d'interopérabilité des contenus éducationnels sur le web. Dans cette approche, la modélisation du profil est décrite par une combinaison des standards LIP et PAPI étendus à de nouvelles propriétés.

En se basant sur ces travaux, nous pouvons conclure que chaque approche possède des avantages et des inconvénients selon le contexte d'utilisation. L'architecture centralisée permet un contrôle central qui réalise des opérations comme : le stockage et la récupération des données, l'analyse des demandes, etc. L'architecture distribuée est adaptée lorsque les données sont distribuées sur plusieurs « nœuds ».

3 Proposition d'un modèle de profil dans un contexte mobile

Notre objectif est de proposer un modèle apprenant et une architecture permettant de gérer le profil. A partir d'une étude comparative entre les différents standards de modélisation du profil, nous avons construit un modèle apprenant en combinant les avantages des standards IMS LIP et FOAF, et en ajoutant les composants manquants, en l'occurrence le contexte de l'utilisateur (le dispositif) et l'agenda. Ce dernier constitue une source d'informations très utile pour des systèmes de recommandation. Le standard IMS LIP fournit un vocabulaire qui a été approuvé dans un contexte large. De plus, il définit une structure de données plus riche que PAPI, en introduisant des éléments tels que les objectifs, les intérêts et les préférences de l'apprenant, éléments indispensables pour les applications d'adaptation. La combinaison avec la norme FOAF est expliquée par la capacité de cette dernière à gérer les relations entre les apprenants.

Le modèle peut donc avoir les catégories suivantes :

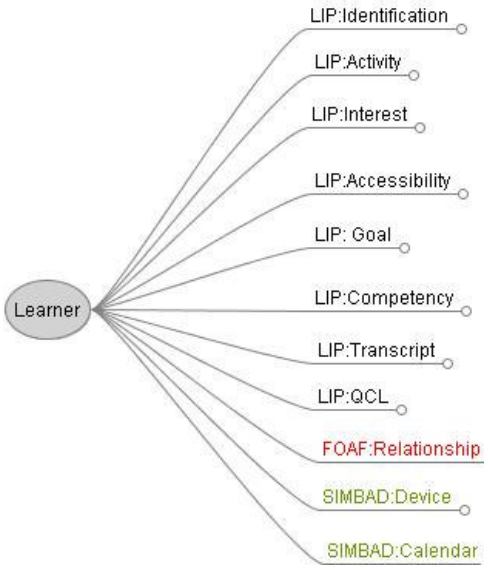

Figure 4 : Un extrait du modèle apprenant proposé

Un profil apprenant contient plusieurs informations. Chaque type d'information a des caractéristiques différentes. Pour simplifier la gestion, une classification des informations est

nécessaire. Il existe plusieurs critères pour les classifier, mais nous nous intéressons ici à deux critères: la stabilité et la taille. Avec le premier critère, nous avons deux types d'information :

- Les informations stables : ce sont les informations comme les données personnelles de base (ex : nom, âge, adresse, etc.) et les données éducatives (ex : but, certification, etc.). Elles ne changent pas souvent.
- Les informations non stables : ce sont les informations comme les tâches, l'agenda, les intérêts etc. Elles changent selon le contexte de l'utilisateur.

Le critère de la taille comprend également deux types d'information : petite taille correspondant aux informations sous forme de texte, et grande taille pour les informations comme les images, les documents, etc.

Dans le cadre de ce travail, les données du profil d'un apprenant sont fragmentées et maintenues par des fournisseurs de profil qui se situent au niveau de serveurs différents sur le réseau Internet. Notre objectif est de concevoir un système ayant un rôle d'intermédiaire, c'est à dire qu'il récupère les fragments du profil, construit un profil complet, et le partage à d'autres services. Construire un profil une seule fois au début, puis le stocker dans le système pour le réutiliser plusieurs fois, signifie qu'une partie des données devient périmée quand l'utilisateur effectue la mise à jour de ses informations dans les fournisseurs originaux, notamment les informations non stables. Par contre, si le profil est construit au moment du besoin, cela peut s'avérer coûteux en termes de performances.

Notre solution est donc la fusion des deux approches. C'est-à-dire que l'on va stocker les données stables et récupérer les données non stables au moment du besoin. Les informations stables ne changent pas souvent mais ce n'est pas toujours le cas. Alors, un mécanisme de mise à jour pour eux va résoudre ce problème. Il en va de même pour le cas de la taille des données. Ainsi, notre modèle de gestion du profil est une fusion du modèle centralisé et décentralisé.

4 Architecture multi-agents

En se basant sur le modèle de gestion du profil abordé ci-dessus, notre système est à la fois centralisé et décentralisé. Il est centralisé parce qu'il a besoin d'un contrôle central qui réalise des opérations comme : le stockage des données, l'analyse des demandes, la récupération des données, la construction profil, etc. En outre, il est décentralisé parce que ses données sont distribuées sur plusieurs « nœuds ». Chaque nœud stocke localement un fragment particulier du profil dans son propre langage de représentation. Le système permet de récupérer et gérer efficacement des fragments de profil de l'utilisateur, afin de pouvoir partager et/ou réutiliser le profil facilement. Dans le contexte du projet SIMBAD[1], le système dispose aussi d'un module de recommandation qui permet de recommander les ressources convenables à l'apprenant en se basant sur son profil. Ainsi, le système de profil mentionné ci-dessus lui sera une source utile.

[1] SIMBAD (Semantic Interoperability for Mobile collaborative and ADaptive application) est un projet de l'INT qui s'intéresse à la description et à la composition de ressources pédagogiques et de workflows.

L'architecture basée sur les agents a plusieurs avantages par rapport à celle des services web dans le contexte ubiquitaire, surtout pour les applications distribuées. Ainsi, nous avons implémenté un prototype basé sur l'architecture multi-agents. Un système multi-agents est composé d'un groupe d'agents autonomes ou semi-autonomes qui interagissent entre eux, afin de réaliser des tâches ou atteindre quelques buts.

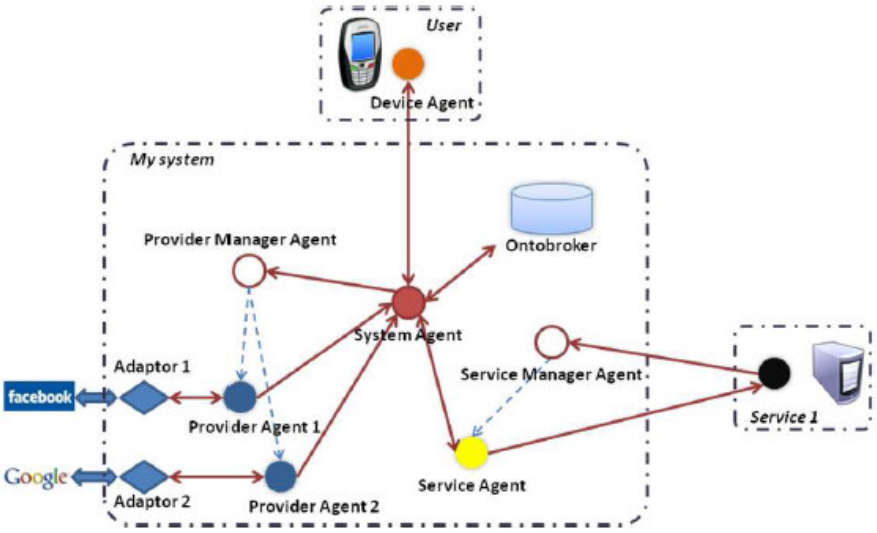

Figure 5: Communication entre les agents

Le système se compose d'une partie Médiateur (système central), la partie des services, la partie des utilisateurs et la partie des fournisseurs. La communication entre ces parties est basée sur la communication entre agents, sauf la partie des fournisseurs. L'architecture du système est composée de 6 agents : l'Agent Système, l'Agent Gestionnaire de Fournisseurs, l'Agent Gestionnaire de Service, l'Agent Fournisseur, l'Agent Service, et l'agent Dispositif. L'architecture est illustrée dans la figure 5. Les agents collaborent ensemble afin de suivre l'apprenant, lui fournir des conseils personnalisés et des recommandations lorsque cela est nécessaire. Les trois premiers sont des agents stables, ils fonctionnent continuellement et forment la base du système. Les agents fournisseur et les agents service sont créés automatiquement en fonction de l'état du système.

L'**Agent Système**: c'est l'agent le plus important dans notre système. Il réalise la plupart des tâches importantes du système: réception des demandes des utilisateurs et des agents service, analyse des demandes, envoi de requêtes à l'agent gestionnaire de fournisseur, synthèse des résultats envoyés par des agents fournisseurs, la récupération des données sur Ontobroker[2]. L'**Agent Gestionnaire de Fournisseur**, comme son nom l'indique, a pour fonction de gérer des agents fournisseurs. Il reçoit des requêtes adressées par l'agent système et crée alors des

[2] Ontobroker : comprend des langages et des outils qui permettent d'améliorer l'accès par requêtes et des services d'inférence dans le WWW

agents fournisseur correspondants. Chaque **Agent Fournisseur** récupère le profil utilisateur à partir de différents fournisseurs (Facebook, Google, LinkedLn, etc) et fournit à l'utilisateur un login et un mot de passe afin d'accéder au fournisseur. Notre système peut fonctionner avec de nombreux fournisseurs différents, la communication entre chaque fournisseur est mise en œuvre par un adaptateur spécifique. Le rôle d'un adaptateur et de vérifier quelle information est nécessaire et comment la récupérer à partir des fournisseurs. L'agent fournisseur reçoit ces résultats et les envoie à l'agent système. Comme l'agent gestionnaire de fournisseur, mais du côté service, l'**Agent Gestionnaire de Service** s'occupe de gérer des agents service. Lorsqu'un service externe veut proposer une recommandation adaptée, l'agent gestionnaire de service crée l'agent service correspondant. L'**Agent Service** transfère les requêtes à l'Agent Système et envoie les résultats au service.
Côté utilisateur, chaque dispositif de l'utilisateur implémente un **Agent Dispositif**. Il permet à notre système de communiquer avec le dispositif de l'utilisateur. Il possède une interface permettant à l'utilisateur d'envoyer des demandes ou de recevoir des notifications du système.
Le framework adopté pour la mise en place du prototype est le langage Java. Ontobroker a été utilisé comme moteur d'inférence et base de données. Pour implémenter l'architecture multi-agents, nous avons choisi la plateforme JADE, qui est un framework logiciel implémenté en Java. LEAP-JADE, une extension de JADE, a été utilisée pour lui permettre de fonctionner sur les appareils mobiles et le PDA à ressources limitées.

5 Caractéristiques de notre approche

Dans ce qui suit, nous citons quelques caractéristiques de notre approche :
- **Mobilité** : notre prototype est développé dans un contexte où l'utilisateur peut se connecter à partir de différents dispositifs mobiles (smartphone, laptop, ordinateur, etc.), et peut accéder aux informations sur son profil n'importe où, n'importe quand et en temps réel, sur différents points sur le réseau.
- **Intégration selon le type de données** : l'intégration du profil se fait selon le type des données du profil. Pour les données assez stables l'intégration se fait une seule fois à partir du serveur de profil. Pour les données susceptibles de changer l'intégration se fait seulement au moment du besoin et pour un objectif spécifique.
- **Evolutivité** : comme l'intégration des fragments du profil est effectuée de manière décentralisée, il n'y a pas besoin de base de données afin de gérer le profil central. Ainsi, de nouveaux profils peuvent être facilement ajoutés au système.
- **Collaboration** : il est possible de partager les profils ente les utilisateurs si des accords existent entre eux.
- **Confidentialité** : La communication entre profils est effectuée à travers des accords entre les deux partis en utilisant P3P (Platform for Privacy Preferences) (2002). Il s'agit d'une norme du W3C décrivant une architecture qui permet de partager et de stocker les données de l'utilisateur de manière sécurisée, en établissant une politique de confidentialité consistant en des accords entre les sites web gérant les profils d'un côté, et les utilisateurs de l'autre.
- **Autonomie** : les agents réduisent le trafic au niveau du réseau. Ils sont exécutés d'une manière asynchrone et autonome. Le choix de l'architecture multi-agents permet de

déployer et d'élargir le système facilement dans un environnement ubiquitaire, surtout sur des dispositifs ayant une configuration faible (ex : Smartphones). Elle permet de réduire la charge du réseau et de la communication entre le système central et le dispositif mobile. Les traitements et la construction du profil au niveau du système central permettent de réduire la charge du dispositif mobile dont les capacités sont assez limitées. Mais ceci est aussi un inconvénient du système, puisqu'il dépend d'une unité centrale. Le serveur central (ou médiateur) peut tomber en panne en cas de surcharge de requêtes. La figure 5 présente des copies d'écran du prototype développé.

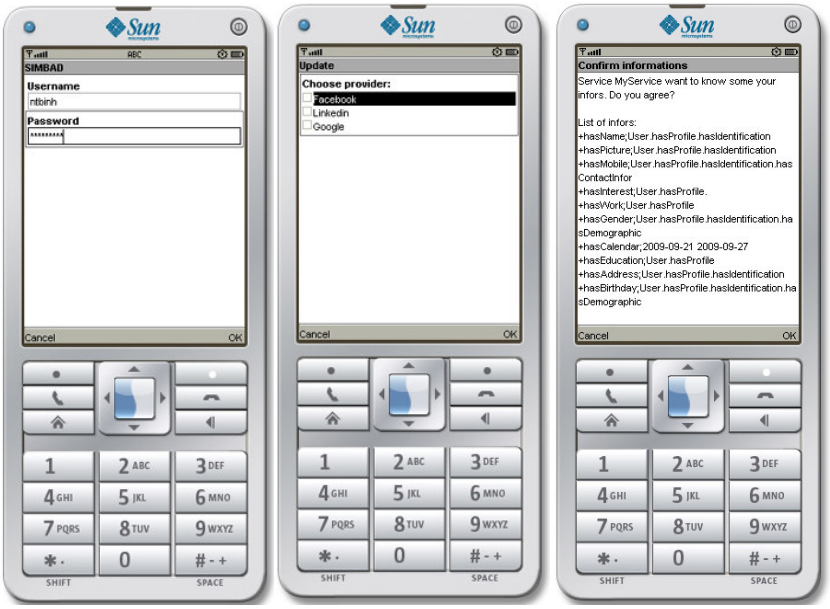

Figure 5: Les trois principales étapes vues par l'utilisateur : identification, choix du fournisseur de profil et liste des recommandations proposées par un service adaptée à son profil

6 Conclusion

Dans ce papier nous avons présenté des techniques pour la découverte et la construction de profils utilisateurs, en utilisant un agent de base de l'architecture distribuée pour construire le profil utilisateur le plus approprié d'un service. Nous avons présenté une solution pour améliorer l'information de profil avec des données contextuelles, et de résoudre le problème de distribution du profil utilisateur. Cependant, le processus d'intégration des fragments du profil pose d'autres problèmes à résoudre. Les données du profil étant considérées comme des informations confidentielles, il convient alors d'étudier les problèmes de confidentialité et de sécurité dans les travaux futurs. En outre, comme l'information sur le contexte du système est maintenant limitée au dispositif utilisé par l'utilisateur, une étude plus riche sur la gestion du contexte sera très utile pour les services de recommandation.

Références

L Aroyo, P Dolog, GJ Houben, M Kravcik, A Naeve, M Nilsson, F Wild Interoperability in Personalized Adaptive Learning Educational Technology & Society (Projet Prolearn), 2006.

Andreas von Hessling, Thomas Kleemann, and Alex Sinner, Semantic User Profiles and their Applications in a mobile Environment, 2005.

D.L. Musa, J.P.M de Oliveira, Integration of Distributed Learner Information through Web Services, 2005.

FOAF (Friend Of A Friend, projet Open Source), 2007. http://www.foaf-project.org/ ; http://fr.wikipedia.org/wiki/FOAF

Houben, J., Geert-Jan Houben, Ad Aerts, Lora Aroyo, Kees van der Sluijs, Bas Rutten, Paul De Bra., State of the art: semantic interoperability for distributed user profiles, Telematica Institut Report, 2005.

IMS LIP (Information Model Specification, Learner Information Packaging), 2001.

http://www.imsglobal.org/profiles/

Jameson A., User Adaptive Systems An integrated Overview. Tutorial presented at the 7th International Conference on User Modeling, June 20-24, 1999.

K. Kabassi et Maria Virvou, Using Web Services for Personalised Web-based Learning, Educational Technology & Society, 6(3), 61-71, 2003.

Katerina Kabassi and Maria Virvou, 2003. Using Web Services for Personalised Web-based Learning

Mohammad Alrifai, Peter Dolog, Wolfgang Nejdl Learner Profile Management for Collaborating Adaptive eLearning Applications., 2006.

P3P (Platform for Privacy Preferences, W3C) http://www.w3.org/P3P/, 2002.

PAPI Learner (Public and Private Information, IEEE), 2000. http://edutool.com/papi/

Summary

In this paper we describe techniques for the discovery and construction of user profiles. Leveraging from the emergent data web, our system addresses the problem of sparseness of user profile information currently faced by both asserted and inferred profile systems. The objective of this work is twofold. First, we study the various existing standards for modeling the user profile and the research work on the profile management systems in the adaptation and customization, and then we propose a model to manage learner profiles. On the other hand, we propose an architecture for discovery and construction of the user profile in a mobile context An agent-based profile system that dynamically builds the most suitable user profile for a particular service or interaction in real-time, is employed in our prototype implementation.

Une méthode mixte d'analyse d'un réseau social: classification prétopologique et centralité d'intermédiarité

Vincent Levorato *

*LIFO (Laboratoire d'Informatique Fondamentale d'Orléans)
Batiment IIIA, Rue Léonard de Vinci, B.P. 6759 F-45067 ORLEANS Cedex 2
vincent.levorato@univ-orleans.fr

Résumé. Dans cet article, nous proposons de modéliser les réseaux sociaux par la théorie de la prétopologie comme une généralisation de la théorie des graphes. Après quelques définitions, nous expliquons comment nous pouvons généraliser par la prétopologie des méthodes d'analyse connues (k-moyennes, centralité d'intermédiarité) dans l'optique d'obtenir des résultats nouveaux. Pour appuyer notre modélisation, nous donnerons un exemple d'application sur un réseau social réel du Web.

1 Introduction

Actuellement, la modélisation des réseaux complexes est utilisée dans de nombreux domaines scientifiques, et se base principalement sur la théorie des graphes. Les graphes sont utilisés, par exemple en sciences sociales, afin de modéliser les interactions entre entités. La plupart de ces études considèrent les individus comme des entités uniques, un groupe étant formé par plusieurs individus, les uns interagissant avec les autres. En effet, la plupart des travaux portant sur l'analyse des réseaux sociaux modélisent un groupe comme une combinaison d'individus, non comme une entité propre. Les réseaux sociaux étant des réseaux complexes (Newman et al. (2006)), un phénomène d'émergence peut apparaître, et le comportement d'un groupe de personnes peut être différent de la "somme" des comportements de chaque individu. De notre point de vue, la théorie des graphes paraît insuffisante pour modéliser toutes les interactions complexes qui ont lieu dans un réseau social : nous proposons l'utilisation d'une théorie plus générale, la théorie de la *prétopologie* (Belmandt (1993)).
Cet article est structuré en trois parties :

- dans la première partie, nous donnons les définitions de la prétopologie et la définition d'un réseau (social ou non) qui en découle.
- dans la deuxième partie, nous explicitons notre apport : une nouvelle méthode d'analyse d'un réseau social en se basant sur l'algorithme des k-moyennes et un indice de centralité connu (ici la centralité d'intermédiarité) adapté au cas prétopologique général.
- enfin, pour illustrer notre discours, nous montrerons les résultats que l'on obtient sur un réseau social réel du Web.

2 Modélisation prétopologique d'un réseau social

Avant d'entrer dans la définition des concepts prétopologiques, nous allons étayer notre propos par un exemple simple d'interactions dans un petit réseau social composé de quatre individus : John, Tim, Ben et Ed (voir Fig. 1). Si on considère qu'une arête représente une relation d'amitié, il est aisé de savoir qui est l'ami de qui. On peut facilement trouver les amis de John par exemple qui sont Tim, Ben et Ed. En revanche, peut-on trouver facilement les amis du groupe $\{John, Ed\}$? Cela pose un problème car Tim est ami avec John mais pas avec Ed. Pour résoudre ce problème, nous devons nous référer à une certaine proximité entre les éléments, ce qui revient à définir la notion de *voisinage* du modèle. On peut définir par exemple que les amis d'un groupe de personnes sont ceux qui ont au moins une relation d'amitié avec un des individus du groupe. Mais on peut également définir qu'un ami d'un groupe de personnes doit être en relation avec tous les membres du groupe. On se rend bien compte de la complexité des interactions qui peuvent intervenir dans un tel modèle, d'où l'intérêt d'utiliser une théorie permettant de modéliser ces phénomènes : la théorie de la *prétopologie*.
La prétopologie est un outil mathématique définissant la proximité entre les éléments d'un espace discret. Cette théorie généralise la topologie, permettant d'analyser un système complexe *pas à pas*, grâce à des processus d'adhérence, d'intérieur, que nous définissons ci-après.

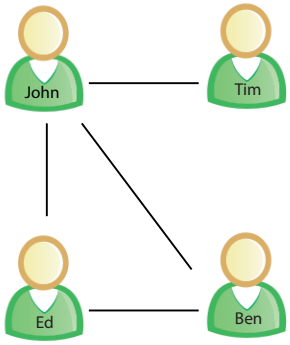

FIG. 1 – *Exemple d'un petit réseau social avec une relation d'amitié*

2.1 Concepts prétopologiques

Soit E un ensemble non vide, et soit $\mathcal{P}(E)$ l'ensemble des parties de E.

Soit une application $a : \mathcal{P}(E) \rightarrow \mathcal{P}(E)$ appelée *adhérence* et définie comme suit :
$\forall A$, $A \subseteq E$ l'adhérence de A, $a(A) \subseteq E$ est telle que :

- $a(\emptyset) = \emptyset$ (P_1)
- $A \subseteq a(A)$ (P_2)

L'adhérence est associée au processus de *dilatation*. De plus, $a(.)$ peut être appliquée à A selon une séquence : $A \subseteq a(A) \subseteq a^2(A) \subseteq ...$. Cela signifie que l'on peut suivre le processus

pas à pas, ce qui n'est pas possible avec la topologie, qui conserve la propriété d'idempotence ($a(A) = a^2(A)$) (Bourbaki (1971)). Grâce à l'adhérence, on peut directement modéliser la notion de *proximité*.

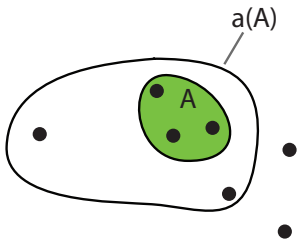

FIG. 2 – *Adhérence de A*

Soit une application $i : \mathcal{P}(E) \rightarrow \mathcal{P}(E)$ appelée *intérieur* et définie comme suit : $\forall A$, $A \subseteq E$ l'intérieur de A, $i(A) \subseteq E$ est telle que :

- $i(A) = [a(A^c)]^c$ (P_1)
- $i(A) \subseteq A$ (P_2)

avec A^c le complémentaire de A soit $E - A$.

L'intérieur est quant à lui associé au processus *d'érosion*. Notons que la propriété 1 de l'intérieur amenant la dualité n'est pas toujours vraie. Il est possible de définir une application intérieur indépendamment de l'adhérence.

On appelle *espace prétopologique* le triplet (E, i, a) dont les applications i et a sont définies précédemment.

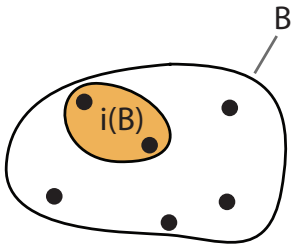

FIG. 3 – *Intérieur de B*

L'intérêt des précédentes définitions pour la modélisation des réseaux sociaux peut être expliqué ainsi : on peut dire que les éléments de $a(A)$ sont proches de A (voisins "directs"), et pour chaque adhérence, on absorbe de nouveaux éléments. On est capable de modéliser des dynamiques complexes comme la diffusion d'une information dans un réseau par exemple.

Nous avons également l'application intérieur : celle-ci permet d'exclure des éléments en périphérie d'un groupe social. Pour reprendre l'exemple précédent, on pourrait retrouver grâce à une série d'intérieurs, l'origine de la diffusion de l'information.

Le processus de dilatation généré par l'adhérence s'arrête à un instant donné et n'évolue plus. Dans ce cas, on a $a^{k+1}(A) = a^k(A)$. On nomme A comme étant un sous ensemble *fermé*. De la même manière, l'évolution de l'intérieur va cesser, ce qui nous donne $i^{k+1}(A) = i^k(A)$. Cette fois, on nomme A comme étant un sous ensemble *ouvert*.
Respectivement, on utilise les notations $F(A)$ pour la fermeture de A et $O(A)$ pour l'ouverture de A.

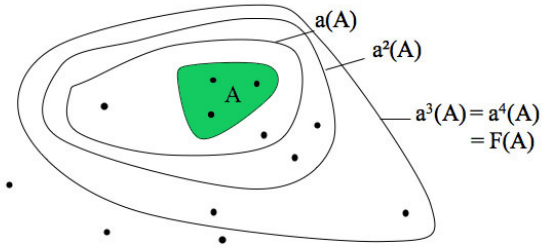

FIG. 4 – *Illustration d'adhérences successives menant au fermé*

On appelera fermé élémentaire et on notera F_x, la fermeture d'un singleton $\{x\}$ de E. On note $\mathcal{F}_e(E, a)$ ou $\mathcal{F}_e$, l'ensemble des fermés élémentaires de E :

$$\mathcal{F}e(E, a) = \{F_x, x \in E\}$$

On appelle fermé minimal de E, tout élément de $F(E, a)$, minimal au sens de l'inclusion. L'ensemble des fermés minimaux est noté : $\mathcal{F}_m(E, a)$ ou $\mathcal{F}_m$.
Un résultat important est que tout fermé minimal est obligatoirement élément de $\mathcal{F}_e$, c'est à dire un fermé élémentaire. Déterminer les fermés minimaux revient donc à explorer les éléments de $\mathcal{F}_e$ et en extraire les éléments minimaux par la relation d'inclusion.

2.1.1 Espace prétopologique de type $\mathcal{V}$

Un espace prétopologique général comme défini ultérieurement ne présente que peu d'intérêt en l'état, car il est difficile d'en faire une analyse. Il faut donc amener une nouvelle propriété pour rendre cet espace prétopologique plus "intéressant", d'où la définition d'un nouvel espace prétopologique : le type $\mathcal{V}$.

Un espace prétopologique de type $\mathcal{V}$ (E, a) est défini comme suit :

$$\forall A, B, A \subseteq E, B \subseteq E \; et \; A \subset B \; avec \; a(A) \subseteq a(B)$$

2.1.2 Espace prétopologique de type $\mathcal{V}_d$

Un espace prétopologique de type $\mathcal{V}_d$ (E, a) est défini comme suit :

$$\forall A, B, A \subseteq E, B \subseteq E \; et \; A \subset B \; avec \; a(A \cup B) = a(A) \cup a(B)$$

Tout espace de type $\mathcal{V}_d$ est de type $\mathcal{V}$.

2.1.3 Espace prétopologique de type $\mathcal{V}_s$

Un espace prétopologique de type $\mathcal{V}_s$ (E, a) est défini comme suit :

$$\forall A, A \subseteq E, avec \; a(A) = \bigcup_{\forall \; x \in A} a(\{x\})$$

Un espace de type $\mathcal{V}_s$ est clairement de type $\mathcal{V}_d$. Les applications a et i ne sont pas forcément idempotentes. On ne doit pas confondre une prétopologie de type $\mathcal{V}_s$ et une topologie. Les types d'espaces les plus utilisés dans nos études sont les types $\mathcal{V}$ et $\mathcal{V}_s$.

2.2 Définition d'un réseau en prétopologie

Maintenant les concepts prétopologiques présentés, nous sommes en mesure de définir un réseau (social) de manière prétopologique. Un réseau social peut être défini comme une famille de relations binaires ou valuées définies sur une population donnée (Degenne (2004)). La dynamique d'un réseau est basée sur des opérations telle que l'arrivée de nouveaux éléments, l'éviction d'éléments existants, la formation de groupe ou la séparation en sous-groupes. Ces phénomènes sont souvent observables dans les réseaux sociaux sous forme de communautés (Backstrom et al. (2006)) mais également dans le cas des réseaux de manière plus générale.

Dans le cadre de la prétopologie, un réseau est une famille de prétopologies sur un ensemble donné (Fig. 5), d'où la définition suivante (Dalud-Vincent (1994)) :

Soit X un ensemble :
soit I une famille dénombrable d'indices ;
soit $\{a_i, i \in I\}$ une famille de prétopologies sur X_i ;
la famille d'espaces prétopologiques $\{(X, a_i), i \in I\}$ constitue un réseau sur X.

On peut représenter ainsi des relations de natures différentes : par exemple, on pourra modéliser un réseau social où les individus sont reliés entre eux par une relation d'amitié (relation binaire) et où leur emplacement géographique est nécessaire (métrique). Le voisinage d'un individu pourra être défini selon les besoins de la problématique : sont voisins ceux qui sont amis et qui habitent dans un rayon de x km. La définition de l'adhérence et/ou de l'intérieur dépend donc de la nature de la problématique : un certain nombre de travaux dans ce domaine ont déjà montré d'intéressants résultats (Bonnevay et al. (1999); Largeron et Bonnevay (1997); Levorato et Bui (2007); Levorato et al. (2009)).

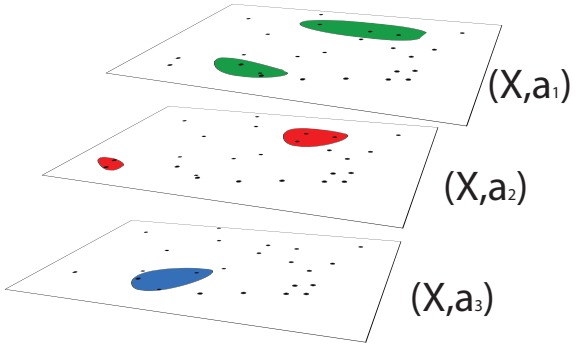

FIG. 5 – *Exemple de trois prétopologies différentes sur X*

3 Analyse d'un réseau social par une méthode mixte : k-moyennes et centralité d'intermédiarité

L'idée est la suivante :
- on partitionne notre réseau social grâce à une méthode des k-moyennes,
- on classe les partitions obtenues selon leur centralité d'intermédiarité.

Dans cette section, nous présentons un algorithme de partionnement des k-moyennes basé sur la théorie de la prétopologie qui a déjà fait l'objet de travaux et qui a été introduit par Le et al. (2008). Nous présenterons ensuite nos contributions : la centralité d'intermédiarité prétopologique, puis la méthode finale permettant une analyse de réseau social.

3.1 MCPR : Méthode de Classification Prétopologique avec Réallocation

La base de cette méthode reprend l'algorithme original de MacQueen (1967). Cet algorithme assigne chaque objet au sous-ensemble dont le centre est le plus proche de l'objet en question :
- Choisir le nombre de sous-ensembles k à obtenir.
- Choisir k groupes de manière aléatoire et en déterminer les centres, ou prendre aléatoirement k objets comme étant les centres initiaux.
- Assigner chaque objet au groupe dont le centre est le plus proche.
- Recalculer les centres de chaque groupe.
- Répéter les deux étapes précédentes tant que la composition des groupes change.

La performance de cette technique est "proportionnelle" à la qualité de la fonction de mesure de distance utilisée. En prétopologie, nous ne sommes pas forcément dans un espace métrique, donc nous ne disposons pas d'une distance à proprement parler. Une pseudo-distance doit être définie : Le et al. (2008) définissent $\delta(A, B)$ comme la pseudo-distance entre deux sous-ensembles A et B d'un ensemble fini E. Celle-ci est calculée comme la distance de Hausdorff.

- $k_0 = min(min\{k|A \subset a^k(B)\}, \infty)$
- $k_1 = min(min\{k|B \subset a^k(A)\}, \infty)$
- $\delta(A, B) = min(k_0, k_1)$

La famille $\mathcal{F}_m$ des fermés minimaux de E représente le nombre k de partitions à obtenir. Concernant la détermination du centre d'un ensemble F, on procède comme suit :
On note x_0 le centre de l'ensemble.
Avec $F = \bigcup_{x \in F}\{x\}$, nous devons décider quel $\{x_i\}$ choisir. Pour cela, nous calculons $Card(a(x_i))$ avec $i \in [1, Card(F)]$. Nous choisissons x_0 tel que $Card(a(x_i))$ soit l'adhérence contenant le plus grand nombre d'éléments.
Au cas où plus d'un x_0 existe, on choisit x_0 de manière à ce que celui-ci minimise la pseudo-distance avec *le plus grand fermé élémentaire qui le contient*.

L'algorithme MCPR se basant sur l'algorithme des *k-moyennes*, on retrouve ainsi son déroulement dans ce qui suit :

1. Choisir k groupes initiaux par les fermés minimaux puis en calculer les centres en formant ainsi k classes.
2. (Ré)affecter chaque objet x à la classe C_i de centre M_i tel que $\delta(x, M_i)$ soit minimale
3. Recalculer le centre M_i de chaque classe.
4. Aller à l'étape 2 jusqu'à ce que les objets ne changent plus de classe.

Nous avons là une méthode prétopologique de partionnement, applicable à des espaces non-métriques ou mixtes. Nous voulons aller au-delà du simple partitionnement d'un réseau social, en classant ces partitions selon leur importance. Ici, nous nous penchons sur le rôle que chaque partition peut jouer dans le réseau en terme de diffusion de l'information.

3.2 Centralité d'intermédiarité prétopologique

La centralité d'intermédiarité a été proposée par Freeman (1977) et défend l'idée qu'un individu peut bien être faiblement connecté aux autres et même relativement éloigné, mais servir d'intermédiaire dans bon nombre des échanges entre les autres membres du groupe. Plus il sert ou peut servir d'intermédiaire pour tous les membres, plus il est en position de contrôler la communication ou d'être indépendant des autres pour communiquer. Un tel individu peut influencer le groupe plus facilement en filtrant ou distordant les informations qui y circulent. Sa position lui permet également d'assurer la coordination du groupe. D'où la définition suivante :

Centralité d'intermédiarité : *Soit n le nombre de sommets d'un graphe, g_{jk} le nombre de chemins géodésiques* [1] *reliant le nœud j au nœud k, et $g_{jk}(i)$ le nombre de ces chemins passant par le nœud i, on définit $C_{AI(i)}$ l'indice de centralité absolu d'intermédiarité du sommet i par :*

$$C_{AI}(i) = \sum_{j}^{n} \sum_{k=1}^{n} \frac{g_{jk}(i)}{g_{jk}}$$

1. plus courts chemins.

avec : $j \neq k \neq i$ et $j < k$

La propriété de Freeman est intéressante, et il nous a paru intéressant d'en adapter une version prétopologique plus générale :

Algorithme 1 ***Algorithme d'intermédiarité prétopologique***
Méthode : ***PretopoBetweeness****(Ensemble A)*
Variables :
A : ensemble de départ tel que $A \subset E$
g_{jk}, $g_{jk}i$, $g_{jk}tmp$: entier
$Bdeg$: réel

Début
$Bdeg \leftarrow 0$
$g_{jk} \leftarrow 0$
$g_{jk}i \leftarrow 0$
Pour *i de 0 à $Card(E)$* ***Faire***
  Pour *j de 0 à $Card(E)$* ***Faire***
    $elt_i \leftarrow$ *singleton de E*
    $elt_j \leftarrow$ *singleton de E*
    $g_{jk}tmp \leftarrow$ *nb_chemins_geo(elt_i,elt_j)*
    Si $g_{jk}tmp > 0$ ***Alors***
      $g_{jk} \leftarrow g_{jk} + g_{jk}tmp$
      $g_{jk}i \leftarrow g_{jk}i$ + *(nb_chemins_geo(elt_i,A) $\times$ nb_chemins_geo(A,elt_j))*
    FinSi
  FinPour
FinPour
Si $g_{jk} > 0$ ***Alors***
  $Bdeg \leftarrow g_{jk}i / g_{jk}$
FinSi
Renvoyer $Bdeg$
Fin

Exemple

Voici un exemple concret sur un passage de la boucle ci-dessus. Soit E un espace prétopologique de type $\mathcal{V}_s$ avec des relations inter-éléments de nature binaire. Pour une plus grande facilité de lecture, nous représentons l'espace E comme un graphe (Fig. 6). Après avoir classé les éléments selon l'adhérence dans laquelle il se trouvent, en supposant que la classe 1 représente les éléments contenus dans l'adhérence de degré 1 (moins l'élément j), nous avons :

1. {A,B,C}
2. {D,E,F}
3. {G,k}

De manière intuitive, nous excluons d'emblée l'élément G. Ainsi, les plus courts chemins entre j et k sont : j-A-D-k, j-A-E-k, j-B-E-k.
Dans ce cas précis, il y a 3 chemins géodésiques entre j et k. Nous voulons calculer la centralité d'intermédiarité d'un élément, par exemple E (nommé i dans l'algorithme). On remarque qu'il y a 2 chemins géodésiques entre j et k passant par E. Donc, la centralité d'intermédiarité de l'élément E, pour un seul passage de la boucle est de $\frac{2}{3}$. Pour avoir le résultat final, il faut bien sûr finir l'algorithme en prenant toutes les paires (j, k) du réseau avec E comme élément i. L'intérêt de cet algorithme est que si pour l'exemple et la compréhension, on ne travaille qu'avec des singletons, dans la pratique, on peut calculer l'intermédiarité d'un ensemble, permettant ainsi de généraliser l'algorithme original.

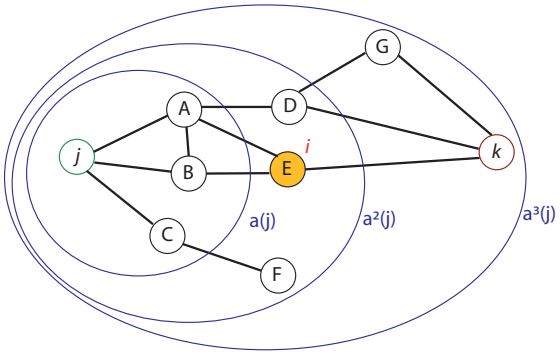

FIG. 6 – *Adhérences successives de j*

3.3 Méthode mixte : MCPR & centralité d'intermédiarité

La méthode d'analyse qui consiste à coupler MCPR et la centralité d'intermédiarité généralisée reprend l'idée énoncée en début de section : on partitionne tout d'abord notre espace en groupes grâce à MCPR puis on classe ces partitions selon leur centralité d'intermédiarité entre elles. C'est à dire qu'on considère uniquement les chemins d'une classe à une autre. Pour une classe étudiée, plus il y a de chemins entre deux autres classes passant par cette classe, plus celle-ci est intermédiaire.

Algorithme 2 ***Algorithme d'analyse mixte***
Méthode : ***PretopoMCPR&Between****(Espace E)*
Variables :
E *: espace prétopologique*
$listeMCPR$ *: liste d'ensembles (partitions)*
$listeResultats$ *: couple ensemble-réél (composition de l'ensemble et score associé)*

Début
$listeMCPR \leftarrow MCPR(E)$
Pour *i de 0 à* $listeMCPR.taille - 1$ ***Faire***
    $listeResultats.index(i) \leftarrow couple(listeMCPR.get(i), PretopoBetweenessClasses(listeMCPR.get(i))$
FinPour
Renvoyer $listeResultats$
Fin

On peut donc associer un poids d'intermédiarité à chaque classe. Nous avons appliqué cette méthode sur un réseau social Web : YouTube. Les données ont été extraites par Cheng et al. (2008) et représentent les liens "Vidéos similaires" qu'il peut y avoir entre les vidéos. Notre méthode a été appliquée sur un réseau non-connexe de 953 vidéos et 3037 liens orientés. Le sens de la centralité d'intermédiarité dans ce cas est que plus une vidéo est intermédiaire, plus son rôle dans le ou les plus courts chemins entre deux vidéos quelconque est important. Si on exclut ce genre d'éléments du réseau, il est probable que la taille du plus court chemin entre couples d'éléments augmente ou même qu'il n'y ait plus du tout de chemin. En terme d'interprétation, cela signifie que certaines vidéos permettent de faire découvrir un maximum d'autres vidéos par leur biais. Avoir cette information peut, par exemple, permettre de promouvoir toute une catégorie de vidéos en mettant en avant seulement quelques vidéos clés. Voici comment est

défini notre espace prétopologique avec R des relations binaires réflexives non-symétriques :

$$R_1(x) = \{y \in E, xRy\}, R_2(x) = \{y \in E, yRx\}$$

$$\forall A \in \mathcal{P}(E), a_1(A) = \{x \in E, R_1(x) \cap A \neq \emptyset\} \text{ et}$$
$$\forall A \in \mathcal{P}(E), a_2(A) = \{x \in E, R_2(x) \cap A \neq \emptyset\}$$

On utilise la première adhérence pour MCPR car ce qui nous intéresse dans un premier temps, ce sont les voisins d'une vidéo qui pointent vers celle-ci. Puis on utilise la deuxième adhérence pour la partie intermédiarité car ce sont les chemins dans le réseau qui nous intéressent par la suite. Nous donnons une illustration de ce que l'on veut obtenir sur une partie de réseau YouTube Fig. 7. Les classes sont en vert, il peut y avoir des éléments non classés, et les éléments colorés appartiennent aux classes les plus intermédiaires. La non-connexité est tout à fait compréhensible pour un réseau tel que YouTube, en tout cas en ce qui concerne la récupération des données. Evidemment, cette exemple, servant uniquement à la visualisation du problème, n'est pas très représentatif puisque l'on a trop peu d'éléments pour pouvoir faire une quelconque interprétation. Néanmoins, cela permet de comprendre le principe : après avoir partitionné le réseau, on recherche les classes les plus intermédiaires, et à fortiori les éléments les composant.

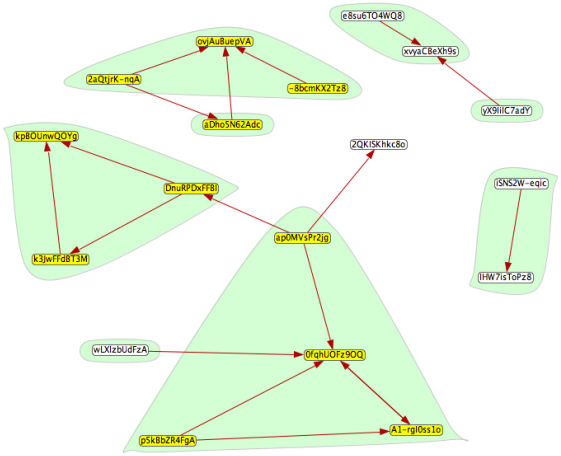

FIG. 7 – *Exemple de résultats obtenus*

Sur le réseau de 953 vidéos et 3037 liens, nous obtenons 171 groupes dans un premier temps, composés au maximum de 20 éléments. Pour ces groupes, seuls 29 d'entre eux possèdent une intermédiarité supérieure à 0. Sur ces 29 groupes, 5 groupes se détachent avec des valeurs supérieures aux 24 restants, représentant 42 éléments du réseau. Les scores se situent entre 0.023 et 0.011, ordre de grandeur que l'on retrouve avec la version standard de la centralité d'intermédiarité. D'ailleurs, en comparant les scores noeud par noeud du réseau avec la centralité d'intermédiarité originale, et le score calculé par notre méthode, si on retrouve

Noeud	Score
$d8nWIqxo0U$	0.028
$L95Sv5aLtZg$	0.015
$Q7Cpi5t-YQI$	0.015
$drW1zIv4wnA$	0.013
$PyC3Byv\_0mM$	0.012
$4LK3KVSKebI$	0.011
$cGY5lhFZFpc$	0.011
$6QVtiaBImlw$	0.011
...	...
Partition	**Score**
[dl4-v0oK7PY, bCZZnJYcPZ4, QBvyHHwQLdw]	0.02353
[mUko-hI7rV8]	0.02353
[Oh40Lm4sRoU, aHer7D0USEc, L95Sv5aLtZg, Q1U6LjALlgo, 4LK3KVSKebI, hECwBuJN2us, eiApTvWkkBI, PyC3Byv_0mM, aWpkJIOwH5c, 2J-Moz5j_do, U1I6H3hTzeI, vLRWAnJQjbU, RYETO3-FTek, Q7Cpi5t-YQI, JWp_gxJORAQ, uRkkmVR6ATU, 9HTzWE8WeCc, S5rU71HA69Y, 9RDCWaSERoY]	0.02353
[_d8nWIqxo0U, 2_Wlu-K7dS8, UPumjbRBqZM, X1lwXjlKIvA, kBYoYeIPBUc, cGY5lhFZFpc, vPXKq2HWHJM, hI4ixSkCgas, P2JYz22fwAI, du1SK1SYgEo, drW1zIv4wnA, b5RVxE4jLD0, PX7Yujz6Kj8, NmzplkH4q88]	0.02353
[SsHjkq-_RCw, 68gwSEpog6g, 81UN9izhrjM, 6QVtiaBImlw, V8pVrKvofpY]	0.02353
...	...

TAB. 1 – *Extrait de résultats de centralités d'intermédiarité : méthode standard et méthode prétopologique mixte*

une certaine cohérence, on a cependant des noeuds qui, seuls, ont une centralité d'intermédiarité proches de zéro, et qui, en groupe, ont un score placé dans les premiers (Tab. 1). Nous observons une propriété émergente que l'on voit apparaître seulement si on regroupe certains éléments qui, ensemble, auront un comportement différent que s'ils agissaient chacun séparément. Notre méthode nous permet dans ce cas de détecter des phénomènes que l'on n'aurait pas pu déceler avec une modélisation et des méthodes d'analyse classiques.

4 Conclusion

Le travail présenté dans ce document participe à la généralisation de la modélisation des réseaux sociaux du Web, de part la modélisation utilisée (théorie de la prétopologie, qui allie qualitatif et quantitatif), et des méthodes algorithmiques proposées. Outre le fait d'avoir donné un algorithme alliant les k-moyennes et la centralité d'intermédiarité généralisés, notre méthode amène une vision différente de celles que l'on connaît habituellement dans ce domaine, en prenant en compte l'émergence de propriétés qui apparaît quand plusieurs éléments forment un groupe, celui-ci ayant un comportement différent des seuls éléments le composant. Il nous est donc susceptible d'obtenir des résultats "plus fins". Bien entendu, le problème de l'interprétation des résultats ne disparaît pas pour autant, nécessitant l'avis d'experts du domaine (sociologues), mais en nous plaçant à un niveau de modélisation plus général, nous sommes désormais capables d'analyser de manière plus précise et plus fine la dynamique et la structure des réseaux complexes, nous permettant de mieux comprendre les phénomènes émergents qui s'y déroulent.

Références

Backstrom, L., D. Huttenlocher, J. Kleinberg, et X. Lan (2006). Group formation in large social networks : Membership, growth, and evolution. *Proc. 12th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*.

Belmandt, Z. (1993). *Manuel de prétopologie et ses applications : Sciences humaines et sociales, réseaux, jeux, reconnaissance des formes, processus et modèles, classification, imagerie, mathématiques*. Hermes Sciences Publications.

Bonnevay, S., M. Lamure, C. Largeron, et N. Nicoloyannis (1999). A pretopological approach for structuring data in non-metric spaces. *Electronic Notes in Discrete Mathematics 2*.

Bourbaki, N. (1971). *Topologie générale*. Hermann.

Cheng, X., C. Dale, et J. Liu (2008). Dataset for "statistics and social network of youtube videos". School of Computing Science Simon Fraser University British Columbia, Canada. http ://netsg.cs.sfu.ca/youtubedata/.

Dalud-Vincent, M. (1994). *Modèle prétopologique pour une méthodologie d'analyse des réseaux : concepts et algorithmes*. Ph. D. thesis, Université Claude Bernard - Lyon 1.

Degenne, A. (2004). Entre outillage et théorie, les réseaux sociaux. *Réseaux Sociaux de l'Internet*.

Freeman, L. (1977). A set of measures of centrality based on betweenness. *Sociometry 40*, 35–41.

Largeron, C. et S. Bonnevay (1997). Une méthode de structuration par recherche de fermés minimaux : application à la modélisation de flux de migrations inter-villes. In *5ème rencontres de la Société Française de Classification*, Lyon, France.

Le, T. V., N. Kabachi, et M. Lamure (2008). Pretopology and a homogeneous method for data clustering. In *RIVF'08 conference*, Hochiminh city, Vietnam.

Levorato, V. et M. Bui (2007). Modeling the complex dynamics of distributed communities of the web with pretopology. In *I2CS*, Munich, Germany.

Levorato, V., T. V. Le, M. Lamure, et M. Bui (2009). Classification prétopologique basée sur la complexité de kolmogorov. *Studia informatica universalis 7.1*, 199–222.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, pp. 281–297.

Newman, M., A.-L. Barabási, et D. J. Watts (2006). *The Structure and Dynamics of Networks*. Princeton University Press.

Summary

In this paper, we propose to model social networks by applying the pretopology theory as a generalization of the graph theory. After giving some definitions and examples, we explain how measures used in social network analysis (k-means, betweenness centrality) can be generalized with pretopology theory in order to obtain new interesting results. To argue in this sense, our work will be supported by an example of application obtained on a real Web social network.

L’ontologie NiceTag : les tags en tant que graphes nommés

Alexandre Monnin*, Freddy Limpens**
David Laniado***, Fabien Gandon**

* EXeCO, Université Paris I Panthéon-Sorbonne
DICEN, Conservatoire National des Arts et Métiers
Alexandre.Monnin@malix.univ-paris1.fr
http://execo.univ-paris1.fr/spip.php?article67
** Edelweiss, INRIA Sophia-Antipolis
{freddy.limpens, fabien.gandon}@sophia.inria.fr
http://www-sop.inria.fr/members/Freddy.Limpens/
http://www-sop.inria.fr/members/Fabien.Gandon/wakka.php?wiki=FabienGandon
*** DEI, Politecnico di Milano
david.laniado@elet.polimi.it

Résumé. Notre analyse part du constat selon lequel les modélisations des tags dont nous disposons actuellement ne prennent pas suffisamment en considération leur richesse et leur diversité. Aussi proposons-nous, pour pallier ce défaut, une ontologie dans laquelle les tags seraient assimilés à des graphes nommés. Ceux-ci sont constitués au minimum d'une ressource reliée à un « signe » qui peut lui-même s'apparenter à n'importe quelle ressource accessible en ligne (un concept d'une ontologie, une image, etc.). Ce modèle entend ainsi fournir une caractérisation suffisamment générale et flexible des tags, et, par voie de conséquence, un cadre susceptible de s'appliquer à tous les tags, quelque soit le modèle sur lequel repose leur description (SCOT, CommonTag, etc.).

1 Introduction

Les tags constituent aujourd'hui un dispositif clef du Web social ainsi qu'un nouveau support d'expression permettant de remplir bien des offices : sélectionner, catégoriser ou classer des contenus, commenter, voter, partager, identifier, etc. Le tagging social et les résultats qu’il engendre (les folksonomies), peuvent être perçus comme des opportunités nouvelles à même d'impliquer les utilisateurs dans une nouvelle forme de commerce vis-à-vis des contenus du Web, libérée des contraintes de l’indexation traditionnelle.

S’il est un trait commun aux modélisations actuelles du tag, c’est bien que ceux-ci y sont représentés comme des instances d'une classe « tag » unique. Cette unicité traduit un rapport univoque aux libellés desdits tags et ce alors même que le libellé d'un tag peut ressortir à une multitude d'emplois, et par conséquent, se voir modéliser de bien des manières différentes en fonction des différents actes de tagging concernés. Qui plus est, la relation entre une ressource taguée et le signe employé pour la taguer est modélisée à l'aide de la seule et

unique propriété « has tag » (SCOT[1]) ou « tagged » (CommonTag[2]). Pourtant, la nature des tags n'est pas univoque ce dont rend compte le modèle CommonTag.

Le but du modèle ici proposé est de founir une modélisation des « actes de tagging » (`TagAction`) qui ne soit pas tributaire d'une interprétation univoque et réductrice de ce que doit être le signe employé en guise de tag. Nous proposons donc, pour décrire les tags de la manière la plus flexible qui soit, de les assimiler en priorité à l'ensemble associant une ressource taguée au signe utilisé pour la taguer, celui-ci pouvant dès lors emprunter différentes formes et conceptualisations (qu'il s'agisse d'une image, d'un litéral, d'un concept issu d'une ontologie, etc.). Ces deux entités sont modélisées à l'aide de la classe `rdfs:Resource` (RDF Vocabulary Description Language), de façon à laisser aux usagers la liberté de mobiliser les modèles du tag ou de la ressource de son choix. Ceci posé, le lien entre la ressource taguée et le signe utilisé pour la taguer est représenté par une propriété et l'assertion obtenue capturée dans un graphe nommé. Sachant que la déclaration des graphes nommés n'est pas nativement supportée en RDF, la décision fut prise de pallier ce manque en intégrant le modèle de Carroll et al. (2005) et la déclarations des sources en RDF/XML proposée par Gandon et al. (2007). Ce choix, rétro-compatible avec les recommandations du W3C, offre en retour la possibilité de mobiliser les différentes initiatives visant à formaliser le tagging (SCOT, CommonTag, etc.) autant que nécessaire et d'établir le lien entre elles de la manière la plus efficace possible – ce qui se traduit, au plan opérationnel, par la possibilité de poser des requêtes portant sur un ensemble hétérogène de modèles. De plus, pour répondre aux problèmes d'ambiguité et d'imprécision des tags, les modèles de tagging actuels, tels que MOAT (Passant et Laublet, 2008), proposent d'associer la signification du tag à chaque acte de tagging. L'ontologie NiceTag, quant à elle, apporte une réponse complémentaire à ce problème en permettant de préciser la relation liant la resource taguée et le tag pour chaque acte de tagging.

Cet article est organisé de la manière suivante. La deuxième section est dévolue à une discussion des motifs nous ayant amenés à proposer un nouveau modèle du tag. Sont détaillés, dans la section suivante, notre modélisation des tags et l'impémentation des graphes nommés qui l'accompagne. La quatrième section présente quant à elle des exemples d'annotations ainsi que des requêtes portant sur des données réelles et impliquant de multiples modélisations. Nous concluons à l'occasion de la cinquième et dernière section.

2 Nature et Usage des Tags

Longtemps la nature des tags est demeurée, au moins partiellement, offusquée. Grâce, cependant, aux travaux menés pour résoudre la "crise d'identité" du Web Sémantique, un arsenal théorique a vu le jour qui nous permet aujourd'hui d'aborder cette question de front. De quoi s'agit-il ? L'augmentation progressive d'un Web de documents en un Web dit sémantique s'avéra bien vite source d'ambiguïté. Effectuée à partir de l'infrastructure du premier, les URIs semblaient d'un coup identifier tout aussi bien des contenus consultables en ligne que des entités absentes du réseau (*en tant que telles*, nous ne parlons évidemment pas ici de leurs descriptions) : personnes, concepts abstraits, fictions, etc.

[1] http://scot-project.org/scot/
[2] www.commontag.org

Une des solutions proposées afin de remédier à ce problème le fut par P.Hayes et H. Halpin. Elle consiste à bien dissocier la *référence*, relation que n'affectent pas les règles techniques de fonctionnement du Web Sémantique, de *l'accès*, traduisant la dimension causale des échanges sur un réseau informationnel tel que le Web. La première doit son fonctionnement à des règles sémantiques, la secondes aux spécifications qui fournissent au Web son assise technologique. Dispositifs matériels autant que sémiotiques, les tags exhibent, *mutatis mutandis*, une dualité similaire à celle des URIs. Dans le sillage des auteurs précédemment cités, nous émettons l'hypothèse selon laquelle il est essentiel d'intégrer les conclusions de ces analyses à tout effort visant à modéliser les tags. Confondre accès et référence serait en effet oublier que le lien symbolique usuel entre mots et choses ne nécessite aucunement de se voir implémenté d'une quelconque manière. Nul besoin d'avoir recours à des moyens d'ordres techniques pour qu'un mot atteigne son objet, aucun artefact n'y pourvoira.

Qu'est-ce en effet qu'un tag, *à première vue*, si ce n'est, pour le dire très grossièrement, l'association, au moyen d'une balise HTML `<a>` et de l'élément `href`, d'un *libellé* qui prend la forme d'une suite quelconque de caractères, d'une images, etc. et d'un lien hypertexte (une URI) ? Contrairement aux vedettes matières ou aux descripteurs dont la sémantique est attachée d'une manière contrainte, soit à un modèle spécifique, soit à un lexique intégralement ordonné par des relations de sens en vue d'éliminer toute ambigüité, le libellé d'un tag est un espace vierge susceptible d'accueillir des entités contrastées, linguistiques ou non, déjouant ainsi toute intelligence globale de la sémantique sous-jacente à son utilisation.

Avec, parfois, l'ajout d'un élément supplémentaire destiné à identifier un lien HTML en tant que tag par l'intermédiaire du microformat `rel="tag"`. Il existe d'ailleurs, à cet égard, une autre manière de caractériser les tags, à notre sens discutable, et qu'illustre ce passage tiré des spécifications du microformat cité à l'instant :

> "By adding rel="tag" to a hyperlink, a page indicates that the destination of that hyperlink is an author-designated "tag" (or keyword/subject) for the current page. Note that a tag may just refer to a major portion of the current page (i.e. a blog post). e.g. by placing this link on a page, tech the author indicates that the page (or some portion of the page) has the tag "tech". The linked page SHOULD exist, and it is the linked page, rather than the link text, that defines the tag. The last path component of the URL is the text of the tag, so fishwould indicate the tag "tech" rather than "fish"[3].

Plusieurs raisons nous conduisent à rejeter ce point de vue :

a) Ce modèle ancre le tagging dans une activité d'emblée communautaire puisque le lien auquel vient se greffer un libellé n'est pas le lien de la ressource qui déclenche l'acte de tagger lui-même mais de la page qui liste l'ensemble des ressources taguées au moyen du « libellé » que l'on retrouve à la fin de l'URI de cette même page. L'activité de l'utilisateur appelé à choisir ses propres libellés (ici « fish ») serait niée si l'on en restait là. Un tag n'étant plus dès lors un signe accolé à une ressource mais une partie d'une URI (ici « tech »).

b) Que le libellé du tag soit ainsi reporté sur l'URI viole un des principes sous-jacent de l'architecture REST du Web sémantique, à savoir le principe d'opacité des URIs[4]. Qui plus est, et l'exemple cité l'illustre, il existe un risque de confusion entre l'ancre textuelle du tag (ici « fish ») et le « libellé » directement intégré à l'URI (ici « tech »).

c) La disparition du lien intention(n)el qui justifie qu'un libellé ait été ajouté à une ressource donnée est éminemment problématique. Des études se sont penchées sur la nature

[3] http://microformats.org/wiki/rel-tag

[4] http://www.w3.org/DesignIssues/Axioms.html#opaque

d'un tel lien afin de mettre en lumière les différentes fonctions du tagging. L'assimilation entre un tag et une URI donnant accès à un ensemble de ressources agrégées par un ou plusieurs individus titulaires d'un compte, ou par la communauté entières des utilisateurs, tend à gommer cette dimension pourtant inhérente à chaque acte de tagging. C'est d'ailleurs là un point que souligne le W3C dans la présentation qu'il donne de ces outils :

> Tagging has emerged as a popular method of categorizing content. Users are allowed to attach arbitrary strings to their data items (for example, blog entries and photographs). While tagging is easy and useful, it often discards a lot of the semantics of the data. A folksonomy tag is typically 2/3 of an RDF triple. The subject is known: e.g., the URL for the flickr image being tagged, or the URL being bookmarked in delicious. The object is known: e.g., http://flickr.com/photos/tags/cats or http://del.icio.us/tag/cats. But the predicate to connect them is often missing. Machine-tags lend themselves to RDF more since they better capture the relationship between the subject and the object. Folksonomy providers are encouraged to capture or infer the semantics around their tags and to leverage semantic web technologies such as RDF and SKOS to publish machine readable versions of their concept schemes. (http://www.w3.org/RDF/FAQ)

Toutefois, l'on aurait tort d'imaginer que ces deux visions s'opposent frontalement. Si l'on ne peut partir de la première pour aboutir à la seconde, l'inverse n'est pas vrai. En partant de ces actes singuliers, qui associent par l'intermédiaire d'un nombre limité de relations typées (cf. section 3.2) *une* ressource à *un* libellé selon des critères d'identité très stricts[5], en les explicitant davantage que cela n'a été le cas jusqu'à maintenant, on solutionne le problème soulevé dans l'extrait précité de la FAQ du W3C lié à l’absence de prédicat. Ceci posé, rien n'empêche ensuite d'affaiblir, volontairement, les critères d'identité du tag. En abandonnant, par exemple, ces relations et les diverses contraintes (de cardinalité ou autre) qui les accompagnent. Ceci afin de lier un libellé non plus simplement à *une* ressource mais à *un ensemble* de ressources, collectées tantôt par le ou les titulaires d'un compte ouvert sur un site de social tagging ou, tout simplement, par l'ensemble des utilisateurs de la communauté. De tels tags « collectifs » se conçoivent alors comme des agrégats de tags individuels. Aussi, au lieu d'aboutir à plusieurs définitions contradictoires les concernant, une solution de continuité émerge par l'application plus ou moins stricte des critères en fonction desquelles les tags sont identifiés.

---

[5] Il faut distinguer ici deux relations irréductibles impliquant des termes différents : la relation entre le libellé et la ressource (que l’on peut modéliser à l’aide de la propriété `irw:refersTo`), et la relation, d'une tout autre nature, découlant des spécifications techniques liée à l'architecture du Web Sémantique, entre la ressource qui déclenche l'acte de tagging et l'URI qui l'identifie et y donne accès (`irw:identifies` et `accesses`). Chaque acte de tagging sur le Web, nommés « tag action » dans notre ontologie, est déclenché par la consultation d'une *ressource en ligne* (les « data item » de la citation précédente) et consiste à lui ajouter un libellé qui lui-même renvoie à une ressource qui peut ou non s'identifier à la précédente. Typiquement, en accédant, via une URI d'un site marchand, au descriptif du roman *La vie devant soi*, paru sous le nom d'Emile Ajar, nous pouvons créer un tag « Romain Gary» et une relation de type « a pour auteur » par exemple qui reliera ainsi le référent de mon tag, non à l'article lui-même mais bel et bien au roman qu'il entend décrire. Bien entendu, dans bien des cas, la ressource à laquelle on accède constituera la référence du tag. *La vie devant soi* réfère, comme nom propre, au livre de R. Gary mais également, dans le contexte quasi-propositionnel du tagging, soit, ici, par ajout de la relation « est à propos de », à la page consultée sur le site marchand (à cet égard les grammairiens et philosophes médiévaux, et le fondateur de la sémiotique moderne lui-même, Charles Sanders Peirce, à leur suite, distinguaient différentes manières de signifier. En particulier la *significatio*, d'ordre lexical, un terme étant pris isolément, et *l'acceptio*, en prise direct, à l'inverse, avec le contexte de l'énonciation).

Voilà, pour résumer les trois principaux aspects où l'ontologie NiceTag innove :

a) en proposant de modéliser le tag au niveau de l'acte singulier de tagging, accompli par un individu (voire une machine lorsque le choix des libellés est automatisé) les tags ainsi définis le sont avec une granularité inédite (cf. section 3.1) ;

b) par la prise en compte des fonctions variées que le tagging remplit, directement ancrée dans les usages et qui les motivent (cf. section 3.2) ;

c) en identifiant les tags non seulement au niveau individuel mais également collectif, par l'assouplissement progressif des critères permettant de les individuer. Ainsi s'effectue le passage du tagging à la folksonomie, des actes singuliers de tagging à leurs agrégats communautaires (avec toutes les nuances intermédiaires). En outillant ce passage progressif de l'un à l'autre, un levier est créé au passage qui permet de répondre aux besoins des diverses communautés amenées à employer le tagging, en mettant tantôt l'accent sur la dimension individuelle de gestion de l'information, tantôt sur la possibilité de favoriser graduellement et à la demande les phénomènes de sérendipité.

3 Modéliser les Tags avec l'ontologie NiceTag

3.1 Les Actions de Tagging en tant que Graphes Nommés

La classe `TagAction` tient donc lieu dans notre modèle, nous l'avons dit, d'équivalent de ce que l'on désigne habituellement sous le vocable « tag ». Elle est modélisée à l'aide d'un graphe nommé contenant des triplets. Ceux-ci ont pour fonction de décrire le lien entre une ressource taguée et un signe. La figure 1 présente le modèle le plus simple de la classe TagAction : une rdfs:resource est liée à une autre rdfs:resource par la propriété « `nt:hasSign` ». Par ce biais, notre ontologie peut être associée à diverses façons de modéliser les ressources taguées et les signes utilisés en guise de tags.

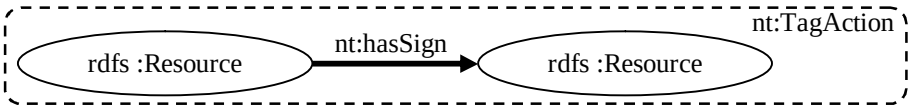

FIG. 1 – *Classe `TagAction` déclarée sous forme de graphe nommé.*

Le graphe nommé « `TagAction` » est déclaré en tant qu'instance de la classe `nt:TagAction`, elle-même une sous-classe de la classe `rdfg:graph` (cf. fig. 2) tirée du modèle de graphe RDF de Carroll et al. (2005). A l'heure actuelle, la syntaxe RDF/XML ne prend pas en charge l'expression des graphs nommés. C'est pourquoi Carroll et *al* ont proposé une nouvelle syntaxe XML pour RDF nommée TriX (Triplets en XML). Gandon *et al.* (2007) ont quant à eux ont proposé une extension rétro-compatible de la syntaxe RDF/XML qui sert à nommer les graphes RDF. Dans la section 3.3 nous décrivons l'implémentation de cette extension afin d'implémenter des instanciations d'actes de tagging.

Pour rendre compte des différentes actions que l'on peut accomplir rien qu'en taguant, nous avons défini plusieurs sous-classes de la classe `TagAction`. Les instances de la classe `TagAction` sont déclarées en tant que graphes nommés résultant d'une action humaine (`ManualTagAction`), pour les différencier de formes plus complexes de tagging telles celles qui impliquent des 'machine tags' (`MachineTagAction`). Agrégées, les actions de tagging peuvent aussi bien revêtir un caractère *collectif* (`CollectiveTagAction`) qu'*individuel* (`IndividualTagAction`). En effet, comme nous l'avons vu, il est souhai-

table de bien distinguer la représentation collective d'un tag, tel que le tag possédant l'adresse `http://delicious.com/tag/improv`, qui renvoie à l'ensemble des signets étiquetés « improv », des représentations individuelles, liées à un compte, de ce même tag, telles que `http://delicious.com/fabien_gandon/improv`, adresse qui pointe vers tous les signets que Fabien Gandon a tagué à l'aide du libellé « improv » (de telles représentations passent par un assouplissement des critères d'identité associés aux tags).

Enfin, la classe `TagAction` est déclarée fille de la classe `sioc:Item` de façon à rendre ainsi compte de la nature partageable des tags. Aussi le tagging peut-il être assimilé à une manière de poster. Ceci nous permet dès lors de décrire l'endroit où les tags sont stockés grâce à la classe `sioc:has_container`, de même que le compte (`sioc:User`) de l'utilisateur (`foaf:Person`) du tag à l'aide de `sioc:has_creator`.

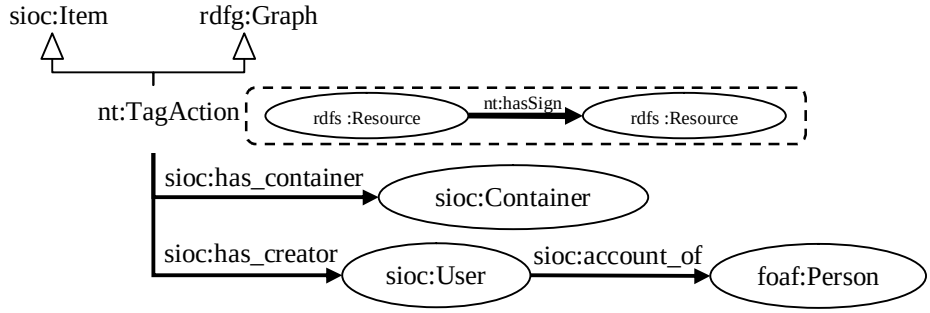

FIG. 2 –*`TagAction` sous-classe de `sioc:Item` et `rdfg:Graph`.*

3.2 Modéliser les usages

Nombre de modélisations actuelles du tag ambitionnent d'associer à celui-ci une signification bien définie ; cette mise en relation est destinée à pallier le problème que représente le fait qu'un terme puisse être doté d'une pluralité de significations selon les contextes ou les communautés qui l'emploient (Passant et Laublet 2008). Seulement, la polysémie est loin d'être la seule source d'ambigüité affectant les tags : une part de leur signification réside en effet dans les types de relations (jusqu'à présent implicites) qui s'établissent entre la ressource et le signe. Par exemple, l'utilisation du tag « blog », l'un des plus populaires sur delicious, pourra sans contredit renvoyer à deux réalités au moins, bien distinctes l'une de l'autre, alors même que le libellé « blog » aura quant à lui été employé conformément une seule et même définition. La ressource peut en effet être *à propos* des blogs ou *être* elle-même un blog (voire les deux à la fois, le tout étant alors de savoir à quoi l'utilisateur fait référence). Qui plus est, certains tags destinés à un usage personnel (idiosyncrasiques) ne font sens que pour un utilisateur précis.

Golder et Huberman (2006) ont identifié pas moins de sept classes de tags à partir des fonctions qu'ils remplissent. Sen *et al.* (2006) ont ramené les classes ainsi mises à jour à trois grandes catégories de tags : factuels, subjectifs et personnels. Des études quantitatives basées sur des applications populaires démontrent qu'une part significative des tags tend à tomber sous les deux dernières catégories (Sen et al. (2006), Al-Khalifa and Davis (2007)). D'autres travaux proposent une classification fonctionnelle fondée sur une première distinction entre tags « liés au sujet » et « non liés au sujet », cette dernière catégorie admettant, à son tour, une subdivision entre tags « affectifs » ou relatifs tantôt à la dimension temporelle, tantôt à

l'accomplissement d'une tâche (Kipp 2008). Les tags liés au sujet sont susceptibles quant à eux d'une caractérisation plus précise qui passe par la distinction entre « liés au contenu » et « liés à la ressource ». Inspirés par ces études, en particulier celle, séminale, de Golder et Huberman, nous avons modélisé les différents usages possibles des tags au moyen de sous-propriétés de la propriété `nicetag:hasSign` (cf. Fig 3.)

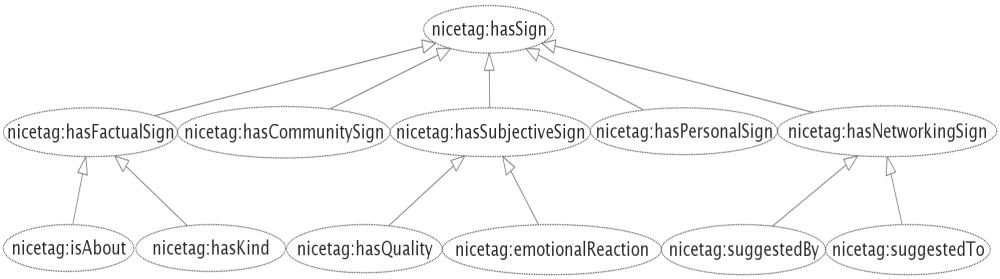

FIG. 3 – *Sous-propriétés de* `nicetag:hasSign`.

La première de ces relations entre signe et ressource est la relation `isAbout`, qui rend compte, sans doute, de l'usage le plus commun du tag, visant à identifier un sujet (le fait d'être *à propos de*). Seconde sous propriété de `hasSign`, `hasKind`, destinée à couvrir l'ensemble des cas où un tag est utilisé pour identifier et caractériser une ressource (forum, vidéo, etc.). La propriété `hasQuality` associe la ressource à un adjectif ou n'importe quel signe exprimant une qualité (joli, idiot, etc.). L'expression des émotions suscitées par une ressource se fait au moyen de la propriété `emotionalReaction` ; typiquement, il s'agira d'exclamations ou d'émoticônes (« wow ! », « ^^ », « :-) »). `hasPersonalSign` couvre l'ensemble des cas où un tag ne fait sens que pour son seul créateur. Ceci inclut tant les classes touchant à l'organisation des tâches de Golder et Huberman que l'ensemble des expressions indexicales ou « token réflexives », qu'elles le soient explicitement (mes_trucs, ma_thèse) ou non (« thèse » pour *ma* thèse, etc.) ; ces dernières trouvant dès lors leur place sans risque d'être confondues avec le sujet ou l'une des caractéristiques quelconques d'une ressource. Dans le même ordre d'idées, la propriété `hasCommunityTag` fut introduite pour rendre compte des tags à destination d'un public ou d'une communauté particulière. A titre d'exemple, nous avons utilisé le tag « #vocampnice2009 » pour partager des ressources concernant le VoCamp où la présente ontologie fut élaborée sur une pluralité de plateformes du Web social. Restent enfin deux propriétés, `suggestedTo` et `suggestedBy`, pour modéliser les actions touchant au *networking*. Plusieurs plateformes ont implémenté pareilles fonctionnalités, notamment delicious en développant pour ce faire une syntaxe spéciale (le double tag « for:username »).

3.3 Utiliser la déclaration de source RDF/XML pour l'implémentation et l'utilisation des graphes nommés

Une requête portant sur une collection de graphes dans SPARQL peut utiliser mot clef GRAPH ou FROM, utilisé pour faire concorder des motifs recherchés (*patterns*) avec des graphes nommés. Le modèle de données RDF se focalise avant tout sur l'expression de triplets dotés d'un sujet, d'un prédicat et d'un objet, cependant ni lui ni sa syntaxe RDF/XML ne fournissent de mécanisme permettant de spécifier la source de chaque triplet. Pour ce faire, il existe une méthode que propose la soumission membre du W3C « RDF/XML Source

Declaration » (Gandon et al. 2007) et qui consiste à associer aux triplets encodés en RDF/XML une URI spécifiant leur origine. Elle requiert l'emploi d'un unique attribut afin de spécifier la source auxquels des triplets exprimés en RDF/XML sont attachés. L'URI de la source d'un triplet est :

1. l'URI source spécifiée par un attribut `cos:graph` associé à l'élément XML encodant ce triplet, si celui-ci existe ; sinon
2. l'URI source du parent de l'élément (obtenu en appliquant les mêmes règles de manière récursive) ; sinon
3. l'URI de base du document.

La portée d'une déclaration de source s'étend de la balise ouvrante dans lequel il apparaît jusqu'à la balise fermante correspondante, à l'exclusion de la portée de toute déclaration interne. Une telle déclaration de source s'applique à tous les éléments et attributs inclus dans son champ. Si aucune source n'est spécifiée, l'URL du document RDF/XML fait alors office de source par défaut. Une seule source peut être déclarée en tant qu'attribut d'un élément donné.

Le code 1.1 montre comment ceci s'applique à un tag capturé dans un graphe nommé. Les lignes 5 à 8 présentent la déclaration du tag en tant que graphe nommé : `http://mysocialsi.te/tag#7182904`. Les lignes 10 à 13 réutilisent le nom de ce graphe pour qualifier le tag en tant que tag créé manuellement par Fabien Gandon le 7 octobre 2009. L'on pourra, à condition de charger au préalable cet ensemble de données RDF dans un entrepôt adéquat, résoudre des requête SPARQL similaires à celle du code 1.2. La ligne 3 correspondant à une recherche portant sur un graphe nommé et le triplet qu'il contient. La ligne 4 permet de s'assurer que ces tags ont été générés manuellement.

Code 1.1. Un tag sous la forme d'un graphe nommé utilisant la syntaxe RDF/XML

```
1  <rdf:RDF xmlns:dc="http://purl.org /dc/elements/1.1/"
2           xmlns:rdf ="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
3           xmlns:cos="http://www.inria.fr/acacia/corese#">
4   <rdf:Resource rdf:about="http://www.yesand.com/"
5                 cos:graph="http://mysocialsi.te/tag#7182904">
6       <nicetag:isAbout>improvisation</nicetag:isAbout>
7   </rdf:Resource>
8   <nicetag:ManualTagAction
9      rdf:about="http://mysocialsi.te/tag#7182904">
10   <dc:creator>Fabien Gandon</dc:creator>
11   <dc:date>2009-10-07T19:20:30.45+01:00</dc:date>
12  </nicetag:ManualTag>
13 </rdf:RDF>
```

Code 1.2. Requêtes SPARQL recherchant des tags créés manuellement.

```
1  SELECT ?t ?a ?g WHERE {
2     GRAPH ? tag { ?t ?a ?g }
3     ? tag rdf : type nt: ManualTagAction }
```

4 Exemples de tags.

Sachant que notre modèle décrit d'abord et avant tout le lien entre la ressource taguée et le signe, sachant également que nous ne posons aucune contrainte tant sur la nature dudit signe que sur celle de la ressource taguée, nous sommes en mesure, par conséquent,

d'exprimer les tags de multiples manières. La figure 4 présente des exemples d'annotations exprimées au moyen de notre modèle. Les actions de tagging font l'objet d'une déclaration sous forme de graphes nommés ainsi qu'expliqué section 3.3 et sont dépeints à l'aide d'une ellipse en pointillés rouges englobant les triplets qu'elle contient. Chaque ellipse représente ainsi une action de tagging et nous avons adopté un code couleur pour distinguer les différentes ontologies mobilisées dans ce schéma. Chaque action de tagging est susceptible de se voir typée à l'aide des sous-classes de `nicetag:TagAction`. Notre exemple provient de données réelles disponibles sur le Web. Les actions de tagging représentées pourraient être typées au moyen de la propriété `ManualTagAction` dans la mesure où elles correspondent toutes à des cas concrets observés chez des utilisateurs de flickr.com ou delicious.com.

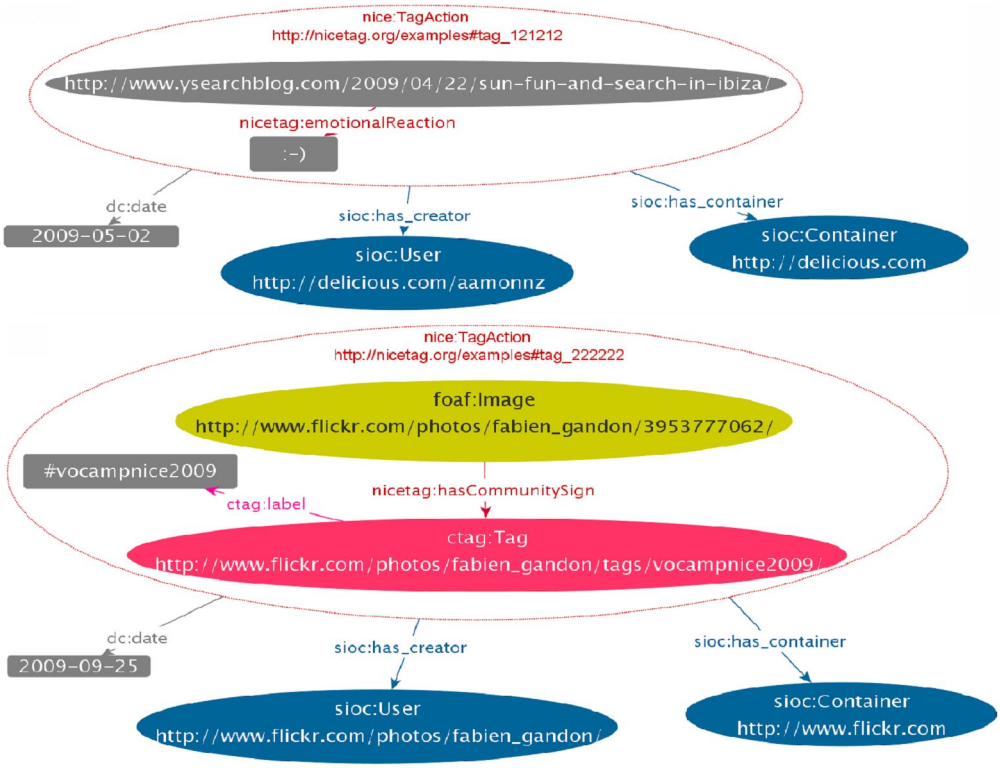

FIG. 4 – *Exemples d'annotations exprimées au moyen de l'ontologie NiceTag.*

Le signe employé pour taguer peut fort bien n'être qu'une suite de caractères telle que « :-) » dans l'exemple de `TagAction`#121212, ou une forme « sémantique » de tagging dans l'exemple de `TagAction`#222222. Rien n'interdit non plus de modéliser les tags à l'aide des instances de la classe `Tag` de l'ontologie CommonTag, de l'ontologie SCOT ou de n'importe quelle ontologie existante ; en un mot, toute `rdfs:Resource` accessible sur le Web. Il convient de noter s'agissant de CommonTag, que la signification du tag « Nice » (donnée par la propriété `ctag:means`) est incluse dans le graphe nommé de l'action de tagging et renvoie à la représentation de la ville de Nice, en France, sur le site geoname.org.

Bien que nous ne contraignions nullement le choix du modèle pour la ressource taguée, nous reconnaissons cependant le caractère essentiel du travail réalisé par Presutti et Halpin

(2009) et présentons des exemples mobilisant la classe `irw:WebResource`, identifiés ici (`irw:identifies`) par l'URL de la page Web de l'événement que constitue le VoCamp qui s'est tenu dans la ville de Nice en 2009. La figure 4 présente des exemples de tagging qui se passent de la classe `irw:Resource`. Dans les deux cas, nous avons recours aux sous-propriétés de `hasSign`. L'action de tagging `#121212` expose par exemple un tag constitué du lien entre une URL et une chaîne de caractère représentant un émoticône (`rdf:Literal` « `:-)` ») mobilisant la propriété « `emotionalReaction` ». Le tag `#222222` présente un cas où une image localisée sur le site flickr.com (modélisée par la classe `foaf:Image`), reçoit un tag `#vocampnice2009` à l'aide de la propriété `hasCommunitySign`. Il s'agit, en l'occurrence, du tag sur lequel se sont accordés les participants du VoCampNice 2009 pour identifier et faire référence à cet événement.

La flexibilité entourant le choix des signes et des ressources permet de récupérer, en une seule requête, tous les actes de tagging exprimés avec notre ontologie, quel que soit le modèle sollicité ou les ressource taguées elles-mêmes. A titre d'exemple, il est possible, en utilisant le moteur RDF CORESE[6], d'écrire les requêtes SPARQL présentées dans le code 1.3. La ligne 2 témoigne de ce que par le biais du mécanisme d'inférence de CORESE, nous sommes en mesure de récupérer tous les types de relations de tagging exprimées à l'aide de `nicetag:hasSign` ou de ses sous-propriétés. La même chose vaut pour tous les types subordonnés de la classe `nicetag:TagAction` (ligne 7). La ligne 3 montre, avec l'assertion OPTIONAL, que notre modèle est capable de récupérer des ressources utilisées pour taguer, fussent-elles typées ou non.

Code 1.3. Requête sur des actions de tagging réparties selon plusieurs modèles.

```
1  SELECT * WHERE {
2   GRAPH ? tagaction {?resource nicetag:hasSign ?sign }
3   OPTIONAL {
4    ?sign rdf:type ?signtype .
5    ?sign rdfs:label ?signlabel .}
6   ?resource rdf:type ?resourcetype .
7   ?tagaction rdf:type nicetag:TagAction }
```

Code 1.4. Requête sur des actions de tagging liées à différents comptes d'utilisateurs.

```
1  SELECT * WHERE {
2    GRAPH ? tagaction {?res nicetag:hasSign ?sign }
3    ?tagaction sioc:has_creator ?user .
4    <http://ns.inria.fr/fabien.gandon/foaf#me> foaf:holdsAccount ?user .
5    ?tagaction rdf:type nicetag:TagAction }
```

Ces exemples présentent également une distinction entre créateurs et containers des actions de tagging. Les utilisateurs sont modélisés en tant qu'instances de la class `sioc:User`. En ajoutant des triplets pour connecter les différents comptes d'une personne (modélisée à l'aide de `foaf:Person`) avec la propriété `foaf:holdsAccount`, il devient loisible de rapporter tous les actes de tagging d'une personne donnée grâce à la requête présentée dans le code 1.4. Le container d'une action de tagging est modélisé au moyen de la classe `sioc:Container` qui permet de poser des requêtes de type « récupérer tous les tags provenant de delicious.com et seulement ces derniers » (cf. code 1.5).

Code 1.5. Requêtes visant à récupérer les actions de tagging de delicious.com

```
1  SELECT * WHERE {
2     GRAPH ?tagaction {?resource nicetag:hasSign ?sign }
```

[6] http://www-sop.inria.fr/edelweiss/wiki/wakka.php?wiki=Corese

```
3     ?resource rdf:type ?resourcetype .
4     ?tagaction rdf:type nicetag:TagAction .
5     ?tagaction sioc:has_container <http://delicious.com > }
```

5 Conclusion.

L'essence du tag telle que la conçoit l'ontologie NiceTag consiste à donner les moyens d'enregistrer la trace d'actions associant une ressource avec un signe grâce à un triplet RDF. Dans l'optique d'intégrer à nos spécifications les fonctions qu'est susceptible de remplir un tag, nous avons créé plusieurs sous-propriétés qui couvrent les divers types de relations qu'entretiennent le tag et la ressource taguée. Cet ensemble est à son tour conçu à la manière d'une instance de la classe `TagAction` et peut en conséquence être enrichi par toutes les propriétés qui lui sont associées, au nombre desquelles figure celle qui spécifie l'utilisateur qui a accompli l'action de taguer, sa date ou son container. Qui plus est, il est possible de définir le genre auquel appartient l'action de tagging (automatique, individuel, collectif…) en choisissant l'une des sous-classes de `TagAction` définies à cet effet. Ce faisant, et en vertu du recours à la déclaration de source pour la syntaxe RDF/XML, qui assigne une URI à une action de tagging, nous obtenons une grande expressivité pour représenter les tags sous une multitude de facettes tout en évitant le fardeau de la réification. Tant les graphes nommés que la déclaration de source pour la syntaxe RDF/XML fournissent une plus-value notable au prix de modifications modestes et rétro-compatibles avec les recommandations officielles relatives au Web Sémantique.

Combiné avec les vocabulaires consacrés au tagging, notre modèle offre un cadre des plus flexibles pour réaliser l'interopérabilité des systèmes de social tagging. Le recours à l'ontologie NiceTag et aux langages de requête SPARQL permet d'agréger et de poser des requêtes sur une variété de sources et de représentations. Quant aux modèles actuels de tagging ils proposent d'enrichir la représentation des actions de tagging en se focalisant sur le signe utilisé pour taguer, qu'il s'agisse d'associer un tag à un concept du Web Sémantique bien définie (avec MOAT ou CommonTag) ou de spécifier les relations morphologiques ou lexicales entre tags (à l'aide de SCOT). L'ontologie NiceTag permet pour sa part de préciser la nature de chaque acte de tagging en précisant la relation entre signe et ressource taguée. Ainsi, devient-il aisé de lever l'ambigüité des tags dès lors qu'un même signe, tout en conservant la même signification, peut être utilisé tant pour spécifier la thématique d'une ressource (`isAbout` « blog ») que son type (`hasKind` « blog »).

Concernant les risques de surcharge cognitive inhérents à notre modèle, nous constatons qu'il est très difficile de naviguer après-coup dans un ensemble de tags dépassant une taille relativement modeste. Les approches (CommonTag, MOAT) offrant de remédier à ce problème en proposant de spécifier la signification des tags (entendus ici comme de simples libellés) nous semblent pour leur part impliquer un effort non négligeable de la part des utilisateurs qui ont ou auront à choisir parmi un vaste ensemble d'entrées possibles. L'alternative que nous proposons nous semble à tous égards plus économique. Elle consiste à choisir parmi un nombre comparativement extrêmement limité de relations pragmatiques, indissociables des usages concrets des tags et du point de vue des utilisateurs. La motivation accompagnant cet effort a toutes les chances d'être plus forte, car tournée d'avantage vers l'usage personnel et circonstancié, que l'action d'accrocher aux tags des « significations » non ambigües - ce qui relève semble relever davantage, disons-le, d'une forme d'altruisme.

Références

Al-Khalifa H. S. et Davis H. C. (2007), Towards better understanding of folksonomic patterns, Hypertext 2007, 163–166.

Carroll J. J., Bizer C., Hayes P. and P. Stickler (2005). Named graphs, provenance and trust. In WWW '05: Proceedings of the 14th international conference on World Wide Web, 613–622, New York, NY, USA: ACM.

Gandon F., Bottolier V., Corby O. et P. Durville (2007), RDF/XML source declaration, w3c member submission. http://www.w3.org/Submission/rdfsource/.

Golder S. A. and Huberman B. A. (2006), Usage patterns of collaborative tagging systems. *J. Inf. Sci.*, 32:2, 198–208.

Halpin H. and Presutti V. (2009), An ontology of resources: Solving the identity crisis. *LNCS*, 5554, 521–534.

Hayes P. J. & Halpin H. (2008), In defense of ambiguity. *Int. J. Sem Web Inf. Sys.*, 4:2, 1–18.

Kim H.-L., Scerri S., Breslin J., Decker S. and H.-G. Kim (2008), The state of the art in tag ontologies: A semantic model for tagging and folksonomies.

Kipp M. E. (2008), @toread and cool: Subjective, affective and associative factors in tagging.

Passant A. et Laublet P. (2008), Meaning of a tag: A collaborative approach to bridge the gap between tagging and linked data. In Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW2008), Beijing, China.

Sen S., Lam S. K., Rashid A. M., Cosley D., Frankowski D., Osterhouse J., Harper F. M. and J. Riedl (2006), tagging, communities, vocabulary, evolution, 181–190.

Wolff C., Heckner M. and S. Mühlbacher (2008). Tagging tagging. Analysing user keywords in scientific bibliography management systems. *Journal of Digital Informaton*, 9:27.

Summary

Current tag modelling does not fully take into account the rich and diverse nature tags, as signs, can take on. We propose an ontology of tags in which tags are modelled as named graphs. These named graphs are made of a resource linked to a "sign" which can be any resource reachable on the Web (an ontology concept, an image, etc.). The purpose of our model is to be able to describe tags in a very general manner, and as an immediate consequence, to describe tags as modelled by other tag models (SCOT, CommonTag, etc.).

Un Wiktionnaire Multilingue et Multiculturel pour les Sciences Sociales et Humaines

L. Khelifa[1,3], N. Lammari[1], H. Fadili[2], J. Akoka[1]

[1] Laboratoire CEDRIC du Conservatoire National des Arts et Métiers de Paris
192, rue Saint Martin, 75141, Paris cedex 3, France
{lammari, akoka}@cnam.fr et lydia-nadia.khelifa@auditeur.cnam.fr

[2] Fondation Maison des Sciences Humaine et Sociales de Paris
54 boulevard Raspail, 75270 Paris cedex 06, France
fadili@msh-paris.fr

[3] Ecole Nationale Supérieur d'Informatique d'Alger (ex INI)
BP 68M Oued Smar, 16309, El Harrach, Alger, Algérie

Résumé. Ce papier est une contribution à la construction d'un Wiktionnaire pour les sciences sociales et humaines (SSH). Ce dernier est une extension du schéma du Wiktionnaire existant afin qu'il puisse prendre en compte aussi bien l'aspect multiculturel des SSH mais aussi pour permettre une représentation de ses entrées selon le standard ISO 1951. Sa construction devrait permettre aux chercheurs des deux rives de la méditerranée d'échanger et de partager des connaissances dans le domaine des sciences sociales et humaines et cela quelque soit leurs lieux géographiques de travail et/ou de résidence. La description conceptuelle de ce dictionnaire en ligne est suivie d'une brève présentation du prototype développé à l'aide de la technologie du Wiki sémantique.

1 Introduction

Selon les définitions simplifiées des dictionnaires, les sciences humaines ont pour objet d'étude tout ce qui concerne les hommes, leur histoire, leurs cultures, leurs réalisations et leurs comportements individuels et sociaux. Les sciences sociales, quant à elles, ont pour objet d'étude les sociétés humaines. Les sciences sociales et humaines (SSH) regroupent de ce fait plusieurs champs disciplinaires hétérogènes tels que, par exemple, la sociologie, l'économie, l'ethnologie, l'anthropologie, la psychologie, l'histoire, la géographie, la démographie, les sciences politiques, l'archéologie, la linguistique, les sciences administratives, les sciences de la religion.

Les sciences sociales et humaines (SSH) jouent un rôle primordial dans la compréhension et l'interprétation du contexte économique, culturel et social des populations. L'évolution de la recherche dans ce domaine passe inévitablement par l'échange et le partage des connaissances entre les chercheurs. Afin de promouvoir les échanges entre les pays du Maghreb et la France dans le domaine des sciences sociales et humaines, un projet de

construction d'un contenu multilingue et multiculturel a été lancé par le FMSH1 en collaboration avec des partenaires des pays du Maghreb et de la France2. Une fois réalisé, ce projet permettra de développer les échanges entre chercheurs maghrébins et leurs partenaires français et de mettre en commun un ensemble de savoir sur les deux cultures et les deux sociétés. Dans ce projet, il est question dans un premier temps de construire un dictionnaire en ligne des SSH franco-maghrébin, inexistant jusqu'à l'heure actuelle. Ce dictionnaire doit non seulement être conforme au standard ISO 1951 (ISO 1951, 2006) mais aussi extensible à plusieurs langues. Il doit aussi reposer sur la technologie Wiki. Une des raisons motivant le choix, du FMSH, pour la technologie Wiki est la facilité et la rapidité de définition, structuration et description de n'importe quelles données, suivant n'importe quel schéma, en utilisant le langage WIKIML (Wiki Markup Language) qui lui est convertible en XML (eXtented Markup Language). De plus, la gestion de l'évolution d'une application de type dictionnaire peut être très compliquée et difficile à mettre en place. Ce problème peut être géré plus facilement sur une plateforme de type Wiki surtout si l'on souhaite changer seulement la structure de la description du contenu.

La fondation Wikimedia héberge un Wiktionnaire. Ce dernier est un dictionnaire ouvert, universel, libre en développement. Il permet, à des personnes autorisées, d'éditer, de publier facilement et rapidement des contenus en ligne et de les faire évoluer via des processus de travail collaboratif par mutualisation de compétences. Il offre aussi une gestion complète des versions, une gestion des historiques des contenus et enfin une gestion des notifications permettant à des personnes intéressées par des thèmes particuliers d'êtres alertées à chaque création, modification ou suppression de contenus en rapport avec leurs thématiques favorites. Cependant, son schéma actuel ne répond pas à tous les besoins fonctionnels du dictionnaire des SSH tel que celui de la recherche d'information par contexte; d'où notre proposition d'étendre le Wiktionnaire actuel.

Le reste du papier est organisé comme suit. La section 2 décrit les spécificités du dictionnaire en ligne des SSH. La section 3 est dédiée à la conception de ce dictionnaire. Le prototype est présenté en Section 4. Enfin, la section 5 conclut ce travail et présente nos perspectives.

2 Description du dictionnaire des SSH

Dans le but de renforcer l'échange et le partage des connaissances entre les chercheurs des deux rives de la méditerranée dans le domaine des sciences sociales et humaines (SSH) et ce quelque soit leurs lieux géographiques de travail et/ou de résidence, le projet de rédaction et de mise en ligne d'un dictionnaire multilingue et multiculturel des SSH a été lancé par la FMSH2. Ce dictionnaire devrait, à court terme, contenir les principaux termes SSH utilisés en France et dans les pays du Maghreb et préciser leurs usages par les deux sociétés et fournir leur traduction d'une langue à une autre. A long terme ce dictionnaire devrait englober les différentes langues du bassin méditerranéen.

La conception de ce dictionnaire doit prendre en compte les faits suivant :

1 Un des acronyme de La Fondation Maison des sciences de l'homme (FMSH), http://www.msh-paris.fr/

2 Les partenaires sont : FMSH, Cnam de Paris, INI (Institut National d'Informatique) d'Alger.

- qu'une entrée A_k dans une langue source peut avoir plusieurs sens et donc plusieurs traductions B_1, ..B_j,…B_m dans la langue cible. Cette même entrée A_k peut être définie avec plusieurs éléments A_1, …, A_i, …A_n du schéma du dictionnaire (synonyme, antonyme, étymologie, expressions figées, hyperonyme, hyponyme, etc.) qui peuvent être à leur tour des entrées dans la même langue source et par conséquent, peuvent avoir plusieurs sens dans cette même langue source et plusieurs traductions dans la langue cible (voir figure 1). Notons, à cet effet que, selon le sens de la traduction, une langue source peut aussi devenir cible et qu'une entrée dans une langue source peut ne pas avoir d'équivalent dans une langue cible.
- la signification attribuée à une entrée du dictionnaire SSH dépend du contexte de définition de cette entrée. Ce dernier est décrit par un ensemble fini et connu de paramètres contextuels qui varient d'une discipline à une autre. Parmi ces paramètres on peut citer les paramètres temporels et géographiques.
- l'ensemble des éléments servant à décrire une entrée fait partie de la norme ISO 1951 (ISO 1951, 2006).

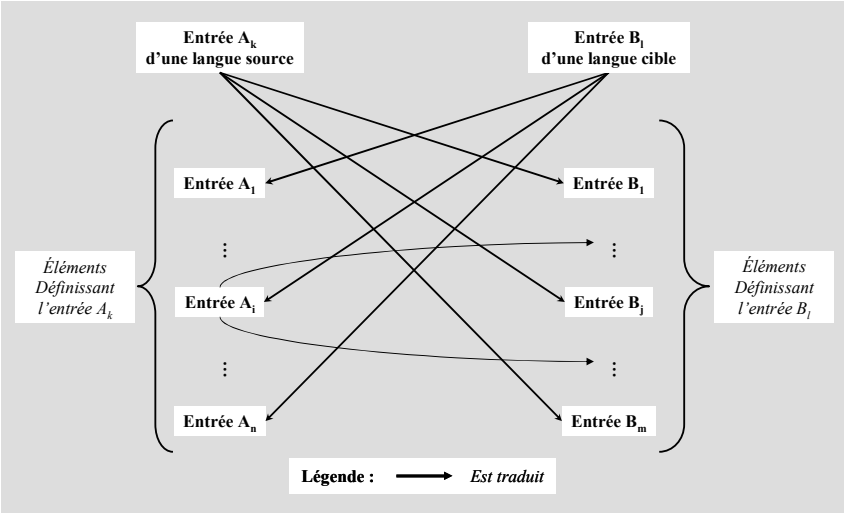

FIG. 1 - *Extrait du schéma du dictionnaire des SSH.*

Il existe, dans la littérature, plusieurs projets de construction de dictionnaires spécialisés. Parmi ces projets, on peut citer le projet PAPILLON (Mangeot, 2006), le projet DHYDRO (Descotte et al., 1999), le projet JMdict/EDICT (Bond et Breen, 2007). et enfin le projet SAIKAM (Ampornaramveth et Aizawa, 2001). PAPILLON utilise le paradigme de construction collaborative de Lunix pour l'édition collaborative de définition. Il offre, parmi les critères de recherche possibles, la restitution d'un terme à partir de sa lecture contextuelle. Dans le projet DHYDRO, un espace terminologique multilingue spécialisé dans le domaine de l'hydrographie a été construit. JMdict/EDICT propose un outil d'édition, à

distance, d'une base terminologique multilingue. SAIKAM est un dictionnaire en ligne dédié à la création de nouveau termes Thai à partir de termes Japonais.

Notons aussi que le W3C propose SKOS (Simple Knowledge Organization System) un modèle de représentation de thésaurus, de taxonomies ou de tout autre vocabulaire contrôlé (SKOS, 2009). C'est un modèle basé sur RDF et RDFS dont l'objectif est le lien, via le Web, de systèmes d'organisation de connaissances.

La réalisation du dictionnaire en ligne des SSH a été, dans le cahier des charges, conditionnée par l'exploitation de la technologie Wiki pour tous les avantages qu'elle offre dont la facilité dans la construction et la maintenance collaborative d'un contenu par des non informaticiens. Cependant, aucun des projets cités ci-avant n'a exploité la technologie Wiki, pour l'élaboration de leurs dictionnaires. Ceci nous a amené à explorer la possibilité d'exploiter le Wiktionnaire actuel hébergé par la fondation WIKIMEDIA. Ce dernier est structuré en articles (Wiktionary, 2009). Chaque article sert à décrire un terme et regroupe :

- une section principale qui sert à décrire le terme dans la langue associée au Wiktionnaire (exemple : section de langue française pour un Wiktionnaire en langue française),
- zéro ou plusieurs sections de langue autre que celle du Wiktionnaire,
- une section catégorie permettant de classer le terme dans une ou plusieurs catégories parmi celles répertoriées
- et enfin une section de liens interwikis permettant de faire des liens vers le même article dans les autres Wiktionnaires. Ces liens se font vers les articles ayant exactement le même titre que l'article, et non vers ses traductions.

La section principale propose :

- un ensemble obligatoire d'éléments de description de base : étymologie, une ou plusieurs sections pour le type de mot (c'est-à-dire ses variations orthographiques, ses abréviations, le ou les termes dérivés, ses synonymes, ses antonymes, ses hyponymes, ses holonymes, ses méronymes, ses traductions, etc)
- et un ensemble d'éléments optionnels : la ou les prononciations, la ou les anagrammes, une section « à voir aussi » qui regroupe les liens en rapport avec le terme de l'article et une section référence permettant de donner les références utilisées lors de la rédaction de l'article.

Une section de langue autre que celle du Wiktionnaire est similaire à la section principale sauf qu'elle ne possède ni de section « Traduction », ni de sections « Hyperonymes », « Hyponymes », « Holonymes » et « Méronymes ».

Le Wiktionnaire actuel ne répond pas aux spécificités du dictionnaire des SSH. D'une part, il ne dispose pas de système automatique de gestion des correspondances qui permettrait de gérer la complexité des renvois entre la langue source et la langue cible. Il est possible, à l'aide du Wiktionnaire actuel, de faire évoluer une entrée indépendamment des autres entrées auquel elle est liée. En d'autres termes, il est possible d'ajouter, dans un Wiktionnaire dédié à une langue A, une traduction d'un terme vers une langue B sans qu'il y est répercussion de ce changement dans le Wiktionnaire dédié à la langue B. De plus les liens interwikis ne peuvent s'établir qu'entre articles de même nom. Cela signifie qu'on ne pourra pas lier deux termes dont l'un est la traduction de l'autre si ces deux termes sont dans des Wikis différents. D'autre part, le schéma du Wiktionnaire actuel ne permet pas une recherche, par contexte, de la signification d'un terme. Cette fonctionnalité s'avère très importante dans le domaine des SSH.

Une autre version du Wiktionnaire existe. Il s'agit de OmegaWiki (OmegaWiki, 2009). Il est basé sur une extension du MediaWiki. OmegaWiki contrairement au Wiktionnaire actuel, réunit dans un même espace tous les Wiktionnaires des différentes langues. Ce qui permet de pallier l'inconvénient du Wiktionnaire actuel concernant la non répercussion des changements d'un Wiktionnaire d'une langue sur celui d'une autre langue. En plus du fait qu'OmegaWiki soit en lecture seulement, il conserve la structure du Wiktionnaire actuel et ne permet donc pas une recherche de termes par contexte.

3 Conception et réalisation du Wiktionnaire des SSH

Tel que mentionné dans la section précédente, une entrée du Wiktionnaire SSH peut avoir plusieurs descriptions. Chacune d'elles est applicable à un contexte donné décrit par un ensemble de paramètres de contexte tels que les paramètres temporel et géographique. De plus, chacune de ces descriptions doit être conforme à la norme ISO 1951. Par conséquent, la conception de notre Wiktionnaire doit reposer sur des correspondances entre les éléments de départ (entrées) et leurs contextes de définition dans la langue source et les éléments d'arrivée (entrées) et leurs contextes de définition dans la/les langue(s) cible(s) selon un schéma qui pourrait contenir les éléments de la norme ISO 1951 suivants : définition, antonyme, synonymes, termes associés, informations orthographiques, prononciation, etc.

Cette description conceptuelle du Wiktionnaire SSH pourrait être représentée à l'aide d'un modèle de classes UML. La figure 2 présente un extrait de ce schéma conceptuel. Ce modèle montre qu'une description d'une entrée (mot) dans une langue donnée est construite par union des variantes de cette description. A chaque variante correspond un contexte défini par la discipline concernée et un ensemble d'éléments de contexte nommés « valeurs des paramètres de contexte ». Chaque discipline a ses propres paramètres de contexte. Chaque entrée décrite à l'aide d'une variante de description donnée peut avoir un synonyme associé à cette variante.

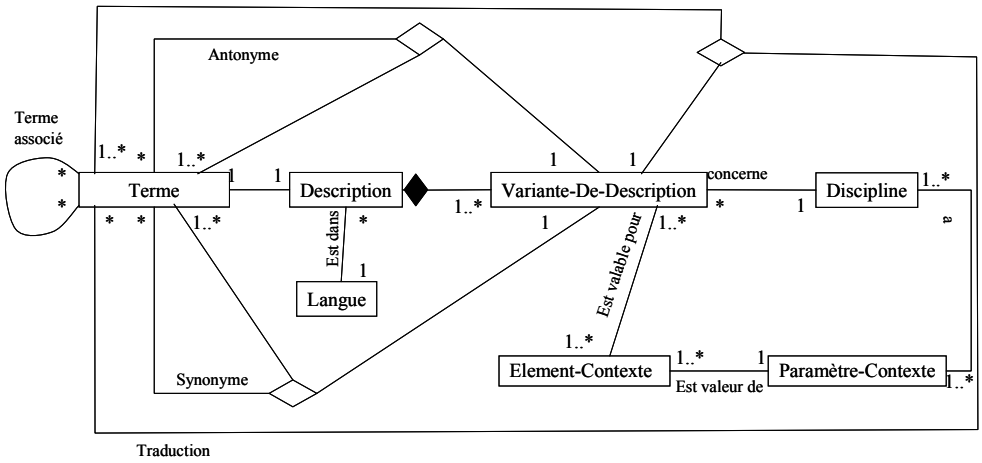

FIG. 2 - Un *extrait du modèle conceptuel du dictionnaire des SSH.*

L'utilisation de la technologie Wiki est une contrainte technique associée à l'élaboration de notre dictionnaire en ligne. Il existe à l'heure actuelle plusieurs Wikis. WikiNi, Wiclear, DokuWiki, MediaWiki et les Wikis sémantiques en sont des exemples. Les Wikis sémantiques tel que KawaWiki (Kawamoto et al., 2006), IkeWiki (Schaffert et al., 2006), SweetWiki (Buffa et al., 2008), Kaukolu (Kiesel, 2006) et le MediaWiki sémantique (Krötzsch et al., 2006) sont des applications du web sémantique aux Wikis. KawaWiki permet la création de pages Wikis, en utilisant des modèles en RDF, ainsi que l'interrogation à l'aide du langage SPARQL. IkeWiki est un outil pour une construction formalisée et collaborative de contenus. Il offre des possibilités d'annotation de liens et de raisonnement. SweetWiki annote sémantiquement les ressources d'un Wiki. Il supporte le tagging social et utilise des ontologies pour décrire le domaine et la structure du Wiki. Il dispose aussi d'un éditeur WYSIWYG. Kaukolu est un Wiki sémantique base sur JSPWiki. Il permet l'annotation, la création et l'édition de pages Wiki. Pour favoriser la création de nouvelles pages, il transforme les URIs en alias. Le MédiaWiki sémantique est une extension du MédiaWiki. Il hérite des avantages du MédiaWiki tels que la facilité d'édition de documents collaboratifs (minimum de pré-requis techniques) et l'évolutivité. Il permet aussi d'annoter les pages Wikis, leurs contenus et les liens entre elles et cela à l'aide de métadonnées compréhensibles par une machine. De plus, pour des objectifs de navigation, les MédiaWikis sémantiques et les Wikis sémantiques en général, à travers l'utilisation intensive des hyperliens donne la possibilité, à un futur utilisateur, d'avoir une vue globale sur une page et de « zoomer », en cas de besoin, sur une partie de son contenu.

Notre étude de l'état de l'art et sa confrontation avec les spécificités de notre Wiktionnaire des SSH, nous a permis de retenir, pour la réalisation de notre Wiktionnaire, la technologie du mediaWiki sémantique.

Les concepts associés aux MédiaWiki sémantiques sont représentés à travers le méta-modèle de la figure 3. Un MédiaWiki sémantique, comme le montre la figure 3, est un ensemble de pages Wikis que l'on peut annoter. Une page Wiki peut être reliée à une autre page Wiki à travers des hyperliens externes. Les hyperliens peuvent aussi être utilisés à l'intérieur d'une page. Les hyperliens peuvent aussi être annotés.

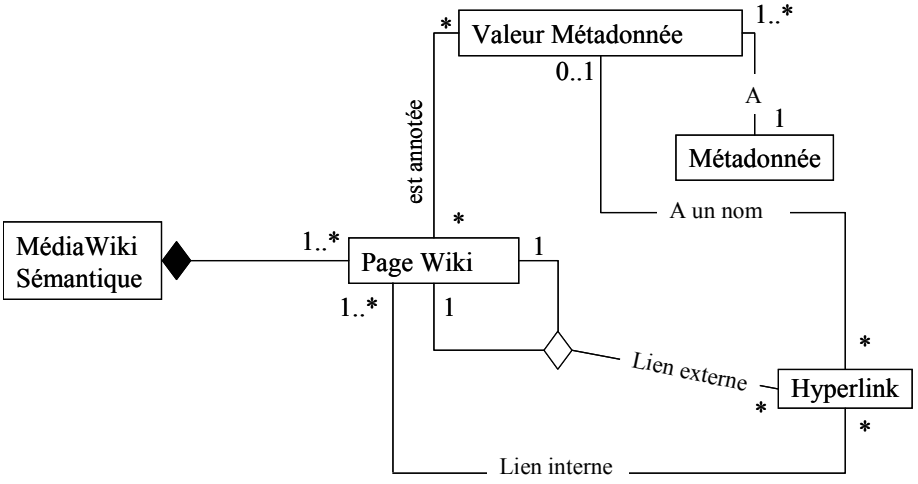

FIG. 3 - *Le métamodèle du MédiaWiki sémantique.*

Les correspondances entre les concepts du MediaWiki sémantique et ceux de notre Wiktionnaire (figure 2) sont représentées dans la table 1.

Concepts du Wiktionnaire des SSH	Concepts du MédiaWiki Sémantique
Description	Page Wiki
Variante de description	Page Wiki
Elément de contexte	Valeur de la métadonnée du paramètre de contexte
Langue	Métadonnée
Discipline	Métadonnée
Paramètre de contexte	Métadonnée
Antonyme	Hyperlien
Terme associé	Hyperlien
Synonyme	Hyperlien
Traduction	Hyperlien

TAB 1. *Correspondance entre les concepts du Wiktionnaire des SSH et ceux du MédiaWiki sémantique.*

Cette table montre que les différentes descriptions d'une entrée (variantes) sont transformées dans un MédiaWiki sémantique en pages Wikis. Ceci s'applique aussi pour une description complète d'une entrée. Les concepts « langue », « discipline », « paramètre de contexte » sont considérés comme des métadonnées du MédiaWiki sémantique. Un élément du contexte du Wiktionnaire est une valeur d'une métadonnée dans le MédiaWiki sémantique. Tous les autres concepts (antonymes, termes associés, synonymes, traductions, etc.) sont transformés en des liens Wikis.

Enfin, pour assurer l'extensibilité de notre Wiktionnaire à plusieurs langues (tel que l'Amazigh) et aux dialectes des pays du Maghreb, nous proposons la construction d'un Wiki par langue. L'exemple de la figure 4 illustre la structure de notre Wiktionnaire des SSH.

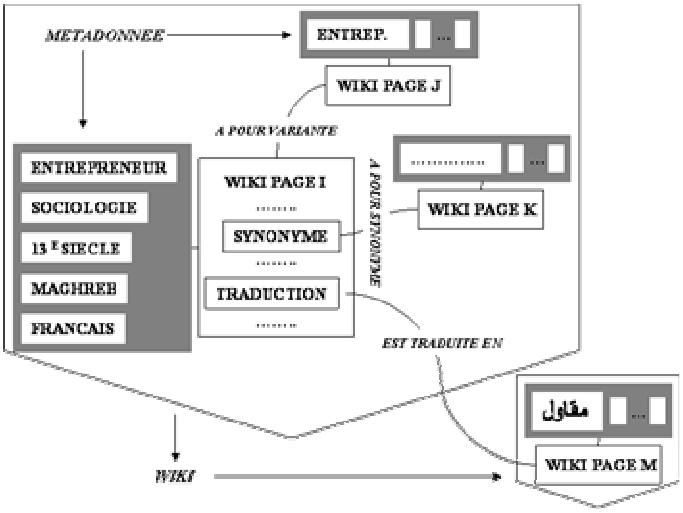

FIG. 4 -. *Illustration de la structure du Wiktionnaire des SSH à travers un exemple.*

Cette figure décrit une page Wiki pour une variante de la description du mot « Entrepreneur » qui est une entrée en Français du Wiktionnaire. Cette page est annotée par les valeurs de métadonnées suivantes :

- «entrepreneur» associé à la métadonnée «terme»
- «sociologie» qui correspond à une valeur de la métadonnée «discipline»
- «Français» qui correspond à une valeur de la métadonnée « langue »
- Les valeurs «13^{ième} siècle» et «Maghreb» qui sont les valeurs respectives des paramètres temporel et géographique. Ces deux paramètres représentent les éléments du contexte du paramètre de contexte « discipline ».

Cette page Wiki associée à une variante de description du mot «entrepreneur» est reliée à d'autres variantes via le lien «a pour variante». De plus cette variante contient un hyperlien «est traduite en» qui relie cette page à sa traduction en arabe pour le même contexte.

4 Le prototype

Après transformation du schéma conceptuel de notre dictionnaire des SSH en un schéma logique respectant la technologie du MédiaWiki sémantique, nous sommes passés à sa réalisation. Pour ce faire, nous avons choisi, pour la réalisation de notre Wiktionnaire des SSH, de construire un Wiki par langue et d'établir des liens entre eux. Un tel choix nous offre la possibilité de réaliser, dans un premier temps un Wiktionnaire franco-arabe extensible, par la suite, à d'autres langues et dialectes pratiqués dans le bassin méditerranéen.

L'éditeur de notre Wiktionnaire (figure 5) intègre, à l'heure actuelle, un sous ensemble des éléments de la norme ISO 1951. Son extension à l'ensemble des éléments de cette norme ou uniquement à celui utile au domaine des SSH, est quelque chose de possible.

L'utilisateur, via cet éditeur, peut annoter une page Wiki, associée à une entrée du Wiktionnaire, en utilisant les métadonnées de son contexte de description. Il peut aussi compléter la description d'une entrée en utilisant des annotations associées aux éléments du

schéma de la norme ISO 1951. Avant de saisir une description (dans une langue donnée) d'une entrée, l'utilisateur doit fournir le contexte de définition de son entrée. En d'autres termes, il doit fournir la discipline, la langue concernée (champ renseigné automatiquement), les autres éléments de contexte qui rendront valide et spécialiseront sa description. Selon le contexte fourni, le système propose soit de modifier une ancienne version de la description (dans le cas où l'entrée existe déjà sous le même contexte) ou encore de la créer. Durant la création ou la modification d'une description, l'utilisateur aura à utiliser les tags proposés pour ajouter éventuellement de nouveaux synonymes, antonymes, etc. Le MédiaWiki sémantique se chargera, par la suite, de traduire ces tags en RDF.

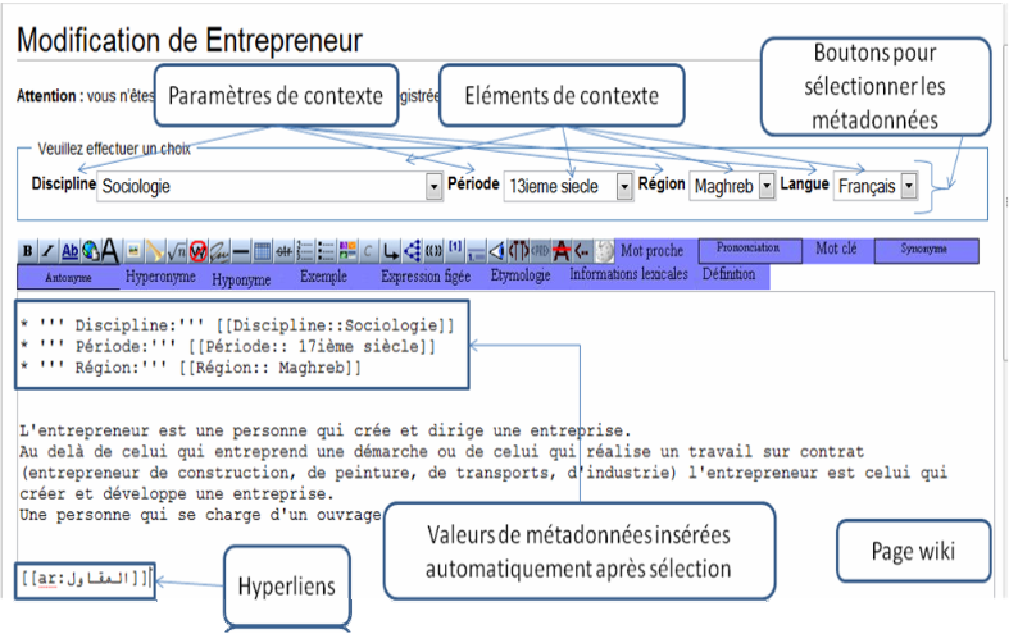

FIG. 5 - *Interface d'édition.*

Notons que compte tenu de l'aspect multiculturel du Wiktionnaire des SSH, une entrée peut ne pas avoir de correspondant dans une langue cible. Notons aussi qu'une description globale d'une entrée peut être obtenue de façon automatique, en rassemblant, dans une seule page Wiki, les différentes variantes d'une entrée. L'utilisateur peut aussi consulter une description pour un contexte donné. Le système, dans ce cas, lui fournira une description dans laquelle les hyperliens, vers les synonymes, les antonymes, les termes associés et sa traduction, apparaissent. Par exemple, l'interface de la figure 6 est fournie à un utilisateur qui souhaite obtenir la description en Français du terme «entrepreneur» pour un contexte décrit à travers les valeurs fournies des métadonnées.

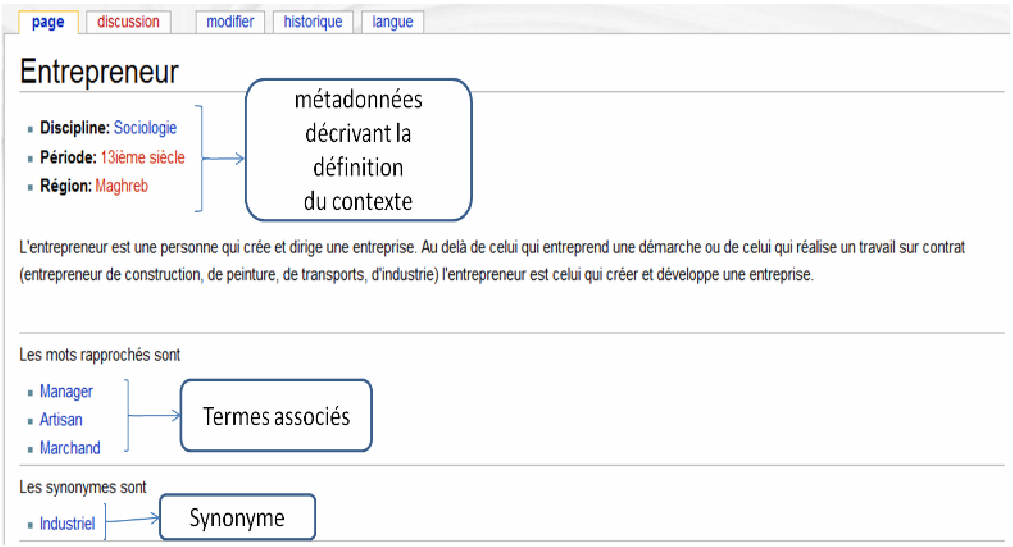

FIG. 6 - *Un exemple de consultation.*

5 Conclusion

Nous avons décrit dans ce papier le dictionnaire en ligne des SSH. Pour sa réalisation nous avons utilisé, tel que imposé dans le cahier des charges, la technologie Wiki pour une édition facile et collaborative de son contenu.

Lors de son alimentation, ce Wiktionnaire contribuera au développement des échanges entre chercheurs du bassin méditerranéen et à la mise en commun d'un ensemble de savoir sur les deux cultures et les deux sociétés.

Après une présentation des spécificités de notre dictionnaire, nous les avons conceptualisés sous forme d'un diagramme des classes UML. La prise en compte de la contrainte technique, nous a amené à opter vers une implémentation de type MédiaWiki sémantique. Une première version de ce prototype est présentée dans ce papier.

La prochaine version du prototype devrait dans un premier temps prendre en charge la gestion des accès (accès libre en consultation mais réservé en gestion aux chercheurs impliqués dans son alimentation) puis la langue Amazigh avec ses symboles graphiques. Une autre perspective de ce travail de recherche et l'extension du Wiktionnaire à une architecture pair à pair.

Reference

Ampornaramveth, V. et Aizawa A. (2001). *Saikam: Collaborative japanese-thai dictionary development on the internet.* The Asian Association for Lexicography (ASIALEX) Biennial Conference. Korea.

Bond, F. et J. Breen (2007). *Semi-automatic refinement of the JMdict/EDICT Japanese-English dictionary*. 13th Annual Meeting of The Association for Natural Language Processing. Kyoto.

Buffa, M., G. Crova, F. Gandon, C. Lecompte, et J. Passeron (2008). *SweetWiki: A semantic wiki*. Journal of Web Semantics, 6: 84-89.

Descotte, S., J. L. Husson, L. Romary, M. Van Campenhoudt, et N. Viscogliosi (1999). *From specialised lexicography to conceptual databases: which format for a multilingual maritime dictionary*. The 2d International Conference on Maritime Terminology. Finland.

ISO 1951 (2006). *ISO TC 37/SC 2/N 323: Presentation/Representation of Entries In Dictionaries.*

Kawamoto, K, Y. Kitamura, et Y. Tijerino (2006). KawaWiki: A SemanticWiki Based on RDF Templates. Workshop of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. United States.

Kiesel, M. (2006). *Kaukolu: Hub of the Semantic Corporate Intranet*. Workshop From Wiki to Semantics of the 3rd European Semantic Web Conference. Montenegro.

Krötzsch, M., D. Vrandecic, et M. Völkel. *Semantic MediaWiki*. The 5th International Semantic Web Conference. United States: Spinger Verlag

Mangeot, M., (2006). *Papillon project: Retrospective and Perspectives*. International Workshop Acquiring and Representing Multilingual, Specialized Lexicons: the Case of Biomedicine. LREC Conference. Italy: Pierre Zweigenbaum.

OmegaWiki (2009). *http://www.omegawiki.org/Meta:Main_Page*.

Schaffert, S. (2006). *IkeWiki: A Semantic Wiki for Collaborative Knowledge Management*. 15th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises. United Kingdom: IEEE Computer Society.

SKOS (2009). *http://www.w3.org/2004/02/skos/.*

Wiktionary (2009). *http://fr.wiktionary.org/wiki/Wiktionnaire:Structure_des_articles #Structure.*

Summary

This paper presents our contribution to a construction of a human and social sciences (HSS) Wiktionary. The latter is an extension of the existing Wiktionary in order to take into account the multicultural aspect of the HSS domain and to allow the representation of the one-line dictionary entries using the ISO 1951 standard. The HSS Wiktionary will allow researchers of the two banks of the Mediterranean Sea to exchange and to share their knowledge in the field of human and social sciences. After a conceptual description of the HSS Wiktionary, the paper gives an overview of the prototype that has been developed using a semantic Wiki technology.