Le Web Social 2011

En conjonction avec
11ème Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC 2011).

Organisé par :

Hakim Hacid
Alcatel-Lucent Bell Labs
France

Cécile Favre
Laboratoire ERIC, Université Lyon 2
France

Ludovic Denoyer
LIP 6, Université Pierre et Marie Curie
France

Brest, France, 25 Janvier 2011

Préface Atelier Web Social - EGC 2011

Préambule

Avec l'avènement du Web 2.0, l'utilisateur est désormais 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 dynamique et la plus intéressante du 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 aujourd'hui effectués dans le cadre du Web social. Les recherches 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. 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 valorisée par les utilisateurs et les fournisseurs de services actuellement.

Après le succès de la première édition de cet atelier en 2010, nous proposons cette année une seconde édition qui vise une fois de plus à permettre la rencontre entre chercheurs et jeunes chercheurs issus à la fois du monde académique et industriel, autour des problématiques liées au Web social en général, particulièrement l'extraction de connaissances à partir du Web social. Cette année, les organisateurs ont décidé d'élargir les thématiques du workshop à des domaines connexes, le but recherché étant de confronter les idées issues de différentes communautés de recherche afin d'établir une vision plus claire et plus commune des éléments qui entourent l'émergence du Web social. Cette rencontre sera l'occasion de faire un état des lieux des avancements des différentes pistes et perspectives de recherche afin de tenter de faire ressortir les verrous scientifiques et industriels à court, moyen et long termes.

Les grandes thématiques d'intérêt pour cet atelier incluent :

- L'analyse de réseaux sociaux.
- L'apprentissage et les réseaux sociaux.
- La recherche d'informations dans les réseaux sociaux.
- La sociologie des réseaux sociaux
- Les technologies des réseaux sociaux.

Remerciements

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

- les membres du comité de programme dont le travail d'évaluation était crucial pour assurer la qualité de l'atelier ;
- les auteurs pour la qualité de leurs contributions constituant la base essentielle de discussions fructueuses ;
- les deux conférenciers invités, Mr. Alain Cupcic (Orange Labs) et Mme Christine Largeron (Laboratoire Hubert Curien – Université Jean Monnet Saint-Etienne)
- pour avoir accepté de partager leur expérience dans le cadre de cet atelier ;
- les organisateurs d'EGC qui ont mis en place l'environnement et les moyens pour la réussite des ateliers.

Comité de programme

- Frédéric Amblard, Université Toulouse 1
- Amel Bouzeghoub, Télécom & Management SudParis
- Johann Daigremont, Alcatel-Lucent Bell Labs France
- Jérôme David, INRIA Grenoble Rhône-Alpes
- Patrick Gallinari, Université Paris 6
- Jean-Gabriel Ganascia, Université Paris 6
- Fabien Gandon, INRIA Sophia Antipolis
- Eric Gaussier, Université Grenoble 1
- Karim Hebbar, Alcatel-Lucent Bell Labs France
- Rushed Kanawati, Université Paris 13
- Luigi Lancieri, Orange Lab
- Christine Largeron, Université Saint-Etienne
- Nicolas Lumineau, Université Lyon 1
- Nathalie Pernelle, Université Paris 11
- Mathieu Roche, Université Montpellier 2
- Fatiha Sais, Université Paris 11
- Yacine Sam, Université de Tours
- Julien Velcin, Université Lyon 2

Les organisateurs de l'atelier Web Social – EGC 2011
Hakim Hacid, Alcatel-Lucent Bell Labs France
Cécile Favre, Laboratoire ERIC - Université de Lyon (Lyon 2)
Ludovic Denoyer, LIP 6, Université Pierre et Marie Curie (Paris 6)

Table des matières

The policy of the telecom operator regarding Data Retention

Alain Cupcic

Orange Research & Development
2 avenue Pierre Marzin
22300 Lannion, France
alain.cupcic@orange-ftgroup.com

[Conférencier invité]

Résumé. I would like to introduce to the participants of this workshop two ways of reflection : (i) the legal obligations of a Telecom Operator and, specifically, the Data Retention, (ii) the project VIGIEs. The legal obligations of a Telecom Operator and, particulary, the Data Retention. With regards to conflicting laws, with regards to technical stakes, and with regards to the operator ethical strategy. The conflicting needs of privacy and security are jointly growing in unprecedented proportions. As the same time, the electronical ways to communicate are exploding in the whole world. And their traces are a big temptation for governments to control and operators to analyze the behaviours of their users.The ANR project VIGIEs gathers different partners around the 2006/24/EC European directive to research solutions to store, analyze, normalize huge data of heterogeneous electronical communications.

Etude comparative d'outils de fouille et d'analyse de réseaux sociaux

Christine Largeron

Université Jean Monnet
Institut Supérieur d'Economie, d'Administration et de Gestion (ISEAG)
2, rue Tréfilerie
42023 Saint-Etienne cedex 2, France
Chrsitine.Largeron@univ-st-etienne.fr

[Conférencier invité]

Résumé. Social networks have seen widespread development since the appearance of web 2.0 platforms. This has led to a renewed interest in social network analysis (SNA) and given rise to a new field of research: the social network mining. This has also led to a growing need for SNA methods and tools. Indeed, in various applications, these tools are useful to do deeper analysis of the network. For this reason, several new tools have been developed oin recent years. The purpose of this article is to compare some of these tools which implement algorithms dedicated to social network mining.

Les Modèles de Markov Cachés pour la prédiction des intentions de recherches dans les *folksonomies*

Ch. TRABELSI, B. MOULAHI et S. BEN YAHIA

Faculté des Sciences de Tunis
{chiraz.trabelsi, sadok.benyahia}@fst.rnu.tn

Résumé. L'essor des sites collaboratifs sur Internet relevant du mouvement participatif que l'on désigne souvent du nom de "Web 2.0" a permis la naissance de nouvelles formes d'indexations des contenus du Web créées librement par les usagers et partagés au sein de réseaux sociaux, baptisées sous le nom de *folksonomies*. Considérées comme source de données, ces dernières s'avèrent d'un grand intérêt pour la recherche d'information. Cependant, la démarche de recherche dans les folksonomies diffère des stratégies de recherche de la traditionnelle médiation des moteurs de recherches dans la mesure où elle prend en considération l'aspect social et comportemental des usagers. Ainsi, afin d'assister les usagers et de leur permettre l'accès le plus pertinent à l'information, nous proposons une approche basée sur le couplage des Modèles de Markov Cachés et l'analyse formelle des concepts pour la prédiction des intentions de recherches dans *les folksonomies*. Les premiers résultats obtenus sur une *folksonomie* réelle s'avèrent prometteurs et ouvrent de nombreuses perspectives.

1 Introduction et positionnement

Le tagging social s'est récemment imposé dans le paysage du web social et collaboratif (Web 2.0) comme un support à l'organisation des ressources partagées, en permettant aux utilisateurs de catégoriser leurs ressources en leurs associant des mots clefs, appelés *tags*. La structure ainsi créée, appelée *folksonomie*, peut être considérée comme un hypergraphe tripartite d'utilisateurs, de tags et de ressources. Cette structure spécifique aux *folksonomies* a rendu leur exploitation d'un grand intérêt pour la recherche d'information. Cependant, la démarche de recherche dans les *folksonomies* diffère des stratégies de recherche de la traditionnelle médiation des moteurs de recherches. En effet, les usagers des *folksonomies* disposent généralement d'outils simples permettant d'explorer les folksonomies afin d'en extraire les ressources couvrant leurs besoins. Cependant, la pertinence des résultats retournés par de tels outils dépend étroitement de la précision des requêtes formulées par les usagers dans la mesure où des algorithmes du domaine de l'extraction des connaissances des bases de données sont exploités. Spécifiquement, une fois qu'une requête est soumise, toutes les ressources indexées par au moins l'un des termes de la requête sont extraites ; chaque ressource sera alors classée en considérant des méthodes de classement spécifiques. Ces dernières s'avèrent en effet inadaptées au contexte du web collaboratif dans la mesure où elles ne prennent pas en compte l'aspect social et comportementale des usagers des folksonomies. Le domaine de la recherche d'information

dans les folksonomies est apparu avec la promesse de permettre aux usagers l'accès le plus pertinent à l'information (Deuff (2007)).

Plusieurs approches ont alors été alors proposées. Certaines tentent d'étendre les approches de recherches traditionnelles par l'intégration de modules supplémentaires capables de garder la trace sociale et comportementale des usagers à travers la proposition de solutions prometteuses reflétées par les systèmes de recommendations (Garg et Weber (2008), Karen et al. (2008) et Amer-Yahia et al. (2008)). En effet, l'objectif de ces systèmes est de filtrer un flux entrant d'information (ressources) de façon personnalisée pour chaque usager, tout en s'adaptant en permanence au besoin d'information de chacun. Pour cela, les moteurs de ces systèmes gèrent des profils d'utilisateurs permettant de choisir les ressources à transmettre à chacun, et adaptent ces profils au cours du temps en exploitant au mieux le retour de pertinence que les utilisateurs fournissent sur les ressources reçues.

Tandis que d'autres travaux proposent d'adapter les techniques de prédiction de requêtes sur le web, généralement basées sur les techniques de fouille des fichiers logs (Jones et al. (2006), Fonseca et al. (2005)), à la prédiction des requêtes dans les folksonomies et ce, en exploitant l'hypothèse de dualité entre l'activité d'annotation des ressources par des tags et celle de la recherche des ressources via les tags. En effet, Krause et al. (2008a), Benz et al. (2010), Bischoff et al. (2008) et Mei et al. (2007) ont mis en exergue que le comportement de tagging d'un usager de *folksonomie* reflète étroitement son comportement de recherche. Ainsi, si un utilisateur annote une ressource R avec un tag T, ce dernier choisira d'accéder à la ressource R si elle apparaît dans le résultat de la recherche obtenu par la soumission de T comme requête. D'une façon similaire, si un utilisateur accède à une ressource R retournée à la suite de la requête Q, ce dernier choisira un tag identique ou similaire à Q pour annoter R. Ainsi, étant définie par un ensemble d'assignements, *i.e.,* des triplets (Ressources, Utilisateurs, Tags), estampillés par la date et l'heure, les *folksonomies* peuvent être assimilées à des fichiers logs, baptisés par Krause et al. (2008b) sous le nom de *logsonomies* et dans lesquelles les tags représentent les requêtes, les utilisateurs correspondent aux identifiants de sessions (IDs) et les ressources annotées définissent les ressources accédées.

À l'heure actuelle, à notre connaissance, aucun travail n'a exploité ces *logsonomies* pour la recommandation des ressources ou la prédiction des requêtes dans les *folksonomies*.

De manière générale, nous nous proposons de répondre à deux principaux défis auxquels doit faire face la recherche d'information dans les folksonomies, à savoir : *(i)* l'identification et la modélisation des intentions de recherches des usagers des *folksonomies* ; et *(ii)* la prédiction des intentions de recherches de ces usagers afin de leur suggérer des requêtes ou leur recommander des ressources couvrant au mieux leur besoin d'information.

Dans ce papier, nous avons défini les intentions de recherches des usagers, comme étant l'intérêt commun porté par une communauté d'utilisateurs pour un ensemble de ressources donné, préalablement retourné à la suite de leur soumission d'un ensemble de requêtes, *i.e.,*tags. Nous avons par la suite modélisé ces intentions de recherches par les concepts triadiques issus de l'analyse formelle des concepts (AFC). En effet, les concepts triadiques permettent de découvrir les sous ensembles d'utilisateurs de la *folksonomie* partageant la même conceptualisation et ce, sur les mêmes ressources. L'algorithme TRIAS, proposé par Jäschke et al. (2006), a été alors utilisé pour l'extraction des concepts triadiques à partir de la *logsonomie*.

Par ailleurs, étant donné que les intentions de recherches des usagers peuvent être vues comme des états cachés de la *logsonomie* et que d'un autre côté elles présentent une structure

tripartite spécifique, nous avons fait le choix d'utiliser les Modèles de Markov Cachés MMCs pour la prédiction des intentions de recherches des usagers. En effet, considérés comme des outils de modélisation puissants, les MMCs ont prouvé leur pertinence dans de nombreux domaines notamment pour la reconnaissance de la parole (Rabiner (1989)) ou l'analyse des séquences en bioinformatique (Durbin (1999)).

Ainsi, afin d'assister les usagers dans leur recherche sur les folksonomies et de leur permettre l'accès le plus pertinent à l'information, nous introduisons dans cet article une approche, appelée PREINTRECH, basée sur le couplage des MMCs et de l'AFC pour la prédiction des intentions de recherches dans les *folksonomies*.

Dans la section 2, nous présentons les définitions préliminaires utilisées dans la suite de l'article. Nous décrivons par la suite dans la section 3 notre approche, PREINTRECH, composée de trois phases à savoir, une phase d'extraction, une phase de construction du MMC et une phase de matching et de prédiction. Nous réservons la section 4 pour étayer à travers un exemple, basé sur un jeu de données réelles, le déroulement de PREINTRECH. L'étude expérimentale de notre approche est illustrée dans la section 5. La section 6, conclut cet article et dresse les perspectives de recherche.

2 Définitions préliminaires

Nous introduisons dans cette section les principales définitions utilisées dans la suite de cet article, *i.e., logsonomie*, concept triadique et intention de recherche.

Définition 1 (LOGSONOMIE)*(Benz et al. (2010)) Une logsonomie est un ensemble de tuples* $\mathcal{L} = (\mathcal{ID}, \mathcal{Q}, \mathcal{RS}, \mathcal{G})$*, avec :*

- $\mathcal{ID}$*,* $\mathcal{Q}$*,* $\mathcal{RS}$ *représentent respectivement, l'ensemble des identifiants utilisateur, l'ensemble des requêtes soumises et l'ensemble des ressources accédées correspondantes,*
- $\mathcal{G} \subseteq \mathcal{ID} \times \mathcal{Q} \times \mathcal{RS}$ *représente une relation triadique, dont chaque* $g \subseteq \mathcal{G}$ *peut être représenté, par un triplet :*

$$g = \{(id, q, rs) \mid id \in \mathcal{ID}, q \in \mathcal{Q}, rs \in RS\}$$

ce qui signifie que l'utilisateur identifié par id*, a accédé à la ressource* rs *à la suite de sa soumission de la requête* q*.*

Considérée comme un graphe tripartite de sessions utilisateurs, de requêtes et ressources accédées, la *logsonomie* peut être représentée par un contexte triadique (Lehmann et Wille (1995)).

Exemple 1 *La Figure 1 illustre un exemple d'une logsonomie* $\mathcal{L}$ *avec* $\mathcal{ID} = \{id_1, id_2, id_3, id_4, id_5, id_6, id_7\}$, $\mathcal{Q} = \{q_1, q_2, q_3, q_4, t_5\}$ *et* $\mathcal{RS} = \{rs_1, rs_2, rs_3\}$*.*

Notons que chaque $\times$ *représente une relation triadique entre une session utilisateur appartenant à* $\mathcal{ID}$*, une requête appartenant à* $\mathcal{Q}$ *et une ressource accédée appartenant à* $\mathcal{RS}$*.*

Nous présentons également une adaptation des définitions d'un tri-set et d'un concept triadique proposées par Jäschke et al. (2008).

Définition 2 (TRI-SET (FRÉQUENT)) *Soit* $\mathcal{L} := (\mathcal{ID}, \mathcal{Q}, \mathcal{CL}, \mathcal{G})$ *un contexte triadique. Un tri-set de* $\mathcal{L}$ *est un triplet* (A, B, C) *avec* $A \subseteq \mathcal{ID}$, $B \subseteq \mathcal{Q}$, $C \subseteq \mathcal{CL}$ *tels que* $A \times B \times C \subseteq \mathcal{G}$*. Un tri-set* (A, B, C) *de* $\mathcal{L}$ *est dit fréquent si* $|A| \geq minsupp_{id}$, $|B| \geq minsupp_q \times$ *et* $|C| \geq minsupp_{cl}$ *avec* $minsupp_{id}$, $minsupp_q$ *et* $minsupp_{cl}$ *définissent les supports minimaux spécifiés par l'utilisateur.*

$\mathcal{ID}/\mathcal{RS}-\mathcal{Q}$	rs_1					rs_2					rs_3				
	q_1	q_2	q_3	q_4	q_5	q_1	q_2	q_3	q_4	q_5	q_1	q_2	q_3	q_4	q_5
id_1		×	×	×			×	×	×			×	×	×	
id_2		×	×	×		×	×	×	×		×	×	×	×	
id_3		×	×	×		×	×	×	×		×	×	×	×	
id_4						×			×		×			×	
id_5		×	×	×	×		×	×	×	×		×	×	×	
id_6				×	×				×	×					
id_7	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×

FIG. 1 – *Un exemple d'une logsonomie.*

Définition 3 (CONCEPT TRIADIQUE (FRÉQUENT)) *Un concept triadique (ou tri-concept) d'un contexte triadique $\mathcal{K}$ = ($\mathcal{E}$, $\mathcal{I}$, $\mathcal{C}$, $\mathcal{Y}$) est un triplet (U, T, R) avec $U \subseteq \mathcal{E}$, $T \subseteq \mathcal{C}$, et $R \subseteq \mathcal{I}$ avec $U \times T \times R \subseteq \mathcal{Y}$ tels que le triplet (U, T, R) est maximal, i.e., pour $U_1 \subseteq U$, $T_1 \subseteq T$ et $R_1 \subseteq R$ avec $U_1 \times T_1 \times R_1 \subseteq \mathcal{Y}$, les inclusions $U \subseteq U_1$, $T \subseteq T_1$, et $R \subseteq R_1$ impliquent que (U, T, R) = (U_1, T_1, R_1). Un tri-concept est dit fréquent s'il est un tri-set fréquent. L'ensemble de tous les tri-concepts de $\mathcal{K}$ est égal à $\mathcal{TC}_\mathcal{K}$ = {$\mathcal{TC}_i$ | $\mathcal{TC}_i$ = (U, T, R) $\in \mathcal{Y}$ est un tri-concept, i = 1...n}.*

À la lumière de ces deux dernières définitions, nous proposons dans ce qui suit la définition d'une intention de recherche :

Définition 4 (INTENTION DE RECHERCHE) *Une intention de recherche dans une logsonomie $\mathcal{L} = (\mathcal{ID}, \mathcal{Q}, \mathcal{RS}, \mathcal{G})$, est un tri-concept $\mathcal{IT}$ = (U', T', R') avec $U' \subseteq \mathcal{ID}$, $T' \subseteq \mathcal{Q}$, et $R' \subseteq \mathcal{RS}$ avec $U' \times T' \times R' \subseteq \mathcal{G}$.*

Ainsi, une intention de recherche $\mathcal{IT}$ représente l'intérêt commun porté par la communauté d'utilisateurs U' pour l'ensemble de ressources R' accédées à la suite de leur soumission de l'ensemble de requêtes Q'.

Dans ce qui suit, nous allons introduire notre approche, appelée PREINTRECH, pour la prédiction des intentions de recherches des utilisateurs dans les *folksonomies*.

3 PREINTRECH : Approche de prédiction des intentions de recherche dans les *folksonomies*

À l'instar de ce qui ce fait dans le domaine de la suggestion de requêtes, nous proposons une démarche axée sur le contexte des requêtes utilisateurs soumises par ces derniers, *i.e.,* la séquence des requêtes précédemment soumises. L'architecture générale de PREINTRECH, schématisée par la figure 2, se compose de trois phases, à savoir : *(i)* **une phase d'extraction** se déroulant hors ligne et correspondant à la fois à l'extraction des séquences de requêtes utilisateurs et à l'extraction des intentions de recherches à partir de la *logsonomie* ; *(ii)* **Une phase de construction** qui se déroule également hors ligne et durant laquelle le MMC sera construit ; et enfin *(iii)* **une phase de matching et de prédiction** se déroulant en ligne et qui correspond au matching du contexte courant de la requête utilisateur avec l'un des états du MMC afin d'identifier les intentions de recherches. Ainsi, à la suite d'une requête soumise par un utilisateur, PREINTRECH lui retourne deux listes de candidats répondant au mieux à son intention de recherche, à savoir : les requêtes suggérées, *i.e.,* des requêtes alternatives et les ressources recommandées ; Chaque candidat étant classé suivant une distribution de probabilité générée

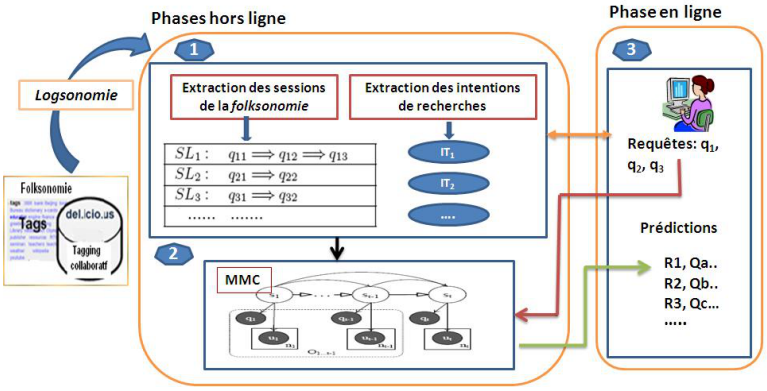


FIG. 2 – *Architecture générale de* PREINTRECH

suite à l'apprentissage du MMC. Les différentes phases de PREINTRECH sont explicitées dans ce qui suit :

3.1 Phase d'extraction

La phase d'extraction se déroule en deux étapes distinctes : *(i)* **une étape d'extraction des séquences de requêtes utilisateurs** ; et *(ii)* **une étape d'extraction des intentions de recherches** à partir de la *logsonomie*.

3.1.1 Etape 1 : Extraction des séquences de requêtes utilisateurs

Avant de présenter notre méthode d'extraction de séquences des requêtes utilisateurs à partir d'une *logsonomie*, on a besoin d'identifier en premier lieu les sessions utilisateurs que nous définissons comme suit :

Définition 5 (SESSION UTILISATEUR)
Une session S_i, relative à l'utilisateur idu_i, est définie par :

$$S_i := \{\{\text{Requêtes utilisateur } q_{S_i,p}\}, \text{Heure et date de soumission } hd_{S_i,t}, rs_{S_i,r}\}.$$

avec $rs_{S_i,r}$:= La ressource rs accédée par l'utilisateur idu_i dans la session S_i, $q_{S_i,p}$:= La p-ème requête soumise dans S_i ; et $hd_{S_i,t}$:= L'heure et la date de soumission dans S_i.

La méthode que nous avons utilisée pour l'extraction des séquences de requêtes utilisateurs à partir d'une *logsonomie* s'articule autour de deux étapes explicitées dans ce qui suit :

1. **Étape** 1 **:** ***Extraction des sessions utilisateurs :*** En se basant sur la définition précédemment énoncée, il s'agit dans cette étape d'extraire, pour chaque identifiant utilisateur, les sessions utilisateurs correspondantes. Ainsi, on aura plusieurs sessions par identifiant utilisateur. Un exemple de sessions utilisateurs est présenté dans le tableau 1.
2. **Étape** 2 **:** ***Filtrage des sessions utilisateurs :*** Dans cette étape, mises à part les séquences de requêtes, toutes les autres informations, contenues dans les sessions utilisateurs préalablement extraites, sont élaguées. Un exemple de séquences de requêtes utilisateurs

$S_1 := \{\{q_{1,1}, q_{1,2}, q_{1,3}\}, hd_{1,1}, rs_{1,a}\}; \{\{q_{1,1}, q_{1,4}\}, hd_{1,2}, rs_{1,b}\}; \{q_{1,4}, hd_{1,3}, rs_{1,c}\}$
$S_2 := \{q_{2,1}, hd_{2,2}, rs_{2,d}\}; \{q_{2,3}, q_{2,4}, hd_{2,2}, rs_{2,e}\}$
$S_3 := \{q_{3,1}, hd_{3,1}, rs_{3,f}\}; \{\{q_{3,2}, q_{3,3}\}, hd_{3,2}, rs_{3,g}\}; \{q_{3,3}, hd_{3,3}, rs_{3,d}\}; \{q_{3,4}, q_{3,5}, hd_{3,4}, rs_{3,g}\}$

TAB. 1 – *Un exemple de sessions utilisateurs*

obtenues à partir du Tableau 1 est donné par : $SL_1 : q_{1,1} \Longrightarrow q_{1,2} \Longrightarrow q_{1,3}$, $SL_2 : q_{1,1} \Longrightarrow q_{1,4}$, $SL_3 : q_{2,1} \Longrightarrow q_{2,2}$, $SL_4 : q_{2,3} \Longrightarrow q_{2,4}$, $SL_5 : q_{3,2} \Longrightarrow q_{3,3}$ et $q_{3,4} \Longrightarrow q_{3,5}$. Notons que SL_i, décrit la séquence de requêtes obtenue à partir du filtrage de la session utilisateur S_i.

3.1.2 Extraction des intentions de recherches

Lors de la soumission d'une requête q par un utilisateur idu_i, PREINTRECH commence par capturer le contexte de q, qui représente la séquence de requêtes précédemment soumises par idu_i juste avant qu'il ne soumette q et explore par la suite la *logsonomie* pour chercher quelles sont les requêtes soumises par les autres utilisateurs juste après qu'ils ne soumettent q en considérant le même contexte. Cependant, différents utilisateurs peuvent soumettre différentes requêtes pour décrire le même besoin d'information. Par ailleurs, deux utilisateurs peuvent soumettre la même requête mais accèdent à différentes ressources. Ainsi, si on modélise les intentions de recherches par les requêtes individuelles et les ressources correspondantes alors non seulement on va augmenter considérablement le nombre d'intentions et par conséquent la complexité du modèle mais aussi perdre la relation sémantique entre les requêtes soumises et les ressources accédées relatives au même besoin d'information.

Ainsi, nous avons fait le choix de modéliser les intentions de recherches par les concepts triadiques qui offrent la possibilité de regrouper dans un même concept, l'ensemble des identifiants utilisateurs, l'ensemble des requêtes soumises par ces derniers ainsi que l'ensemble des ressources accédées correspondantes. Une adaptation de l'algorithme TRIAS, proposé par Jäschke et al. (2006) a été par conséquent utilisée. TRIAS prend en entrée la *logsonomie* $\mathcal{L} := (\mathcal{ID}, \mathcal{Q}, \mathcal{RS}, \mathcal{G})$ ainsi que les valeurs de support minimales à savoir, $minsupp_{id}$, $minsupp_q$ et $minsupp_{rs}$ et retourne l'ensemble des concepts triadiques. L'exécution de TRIAS sur la *logsonomie* illustrée par la figure 1, pour $minsupp_{id}$ =2, $minsupp_q$ = 2 et $minsupp_{rs}$ = 2, permet d'obtenir les intentions de recherches IT_1, IT_2 et IT_3 avec IT_1 = {(id_1, id_2, id_3, id_5, id_7), (q_2, q_3, q_4), (rs_1, rs_2, rs_3)}, IT_2 = {(id_5, id_6, id_7), (q_4, q_5),(rs_1, rs_2)} et IT_3 = {(id_2, id_3, id_4, id_7),(q_1, q_4),(rs_2, rs_3)}.

Ainsi, l'intention de recherche IT_2 par exemple, signifie que la communauté d'utilisateurs identifiés par (id_5, id_6, id_7) ont tous porté un intérêt commun aux ressources (rs_1, rs_2) à la suite de leur soumission des requêtes (q_4, q_5).

3.2 Phase de construction du Modèle de Markov Caché

Durant cette phase, le MMC sera construit en considérant les séquences des requêtes utilisateurs de la *logsonomie* et les intentions de recherches préalablement extraites durant la phase d'extraction. Un MMC est défini par deux composantes principales : les observations qui représentent les sessions observables et les états cachés. Ainsi, nous avons modélisé les observations par les séquences des requêtes utilisateurs observables de la *logsonomie* et les états cachés par les intentions de recherches (inobservables). À chaque état caché du MMC, représentant une intention de recherche, est associé une distribution de probabilité d'émission

de symboles, *i.e.,* de requêtes ou de ressources observables. À partir de la suite observée de tels symboles, il est possible d'apprendre les paramètres du MMC susceptibles de produire cette suite avec une forte probabilité.

Ainsi, étant donnés un ensemble d'états cachés $S = \{s_1,\ldots,s_{ns}\}$, un ensemble de requêtes $Q = \{q_1,\ldots,q_{nq}\}$, un ensemble de ressources accédées $RSs = \{rs_1,\ldots,rs_{nrs}\}$ et un ensemble d'identifiants utilisateurs $IDus = \{idu_1,\ldots,idu_{nidu}\}$ avec ns := Le nombre total des séquences de requêtes SL_i ; nq := Le nombre total de requêtes contenues dans SL_i ; nrs := Le nombre total de ressources associées aux requêtes q_{nq} ; et $nidu$:= Le nombre total des identifiants utilisateurs ayant soumis les requêtes q_{nq}. Le MMC est un modèle probabiliste, noté $\lambda = (A, B, B', \pi)$, défini par :

- $A = [\ldots a_{ij} \ldots]$, la matrice de distribution des probabilités de transitions entre les états, avec $a_{ij} = \{P(s_j \mid s_i)\}$, représente la probabilité qu'un utilisateur transite de l'état s_i vers l'état s_j. La somme de tous les a_{ij} d'une ligne de la matrice est égale à 1.
- $B = [\ldots b_j(q) \ldots]$, la matrice de distribution des probabilités d'émission des requêtes avec, $b_j(q) = \{P(q \mid s_j)\}$, désigne la probabilité pour qu'un utilisateur se trouve à l'état s_j et soumette la requête q.
- $B' = [\ldots b_k(rs) \ldots]$, la matrice de distribution des probabilités d'accès aux ressources avec $b_k(rs) = \{P(rs| s_k)\}$, désigne la probabilité pour que l'utilisateur se trouve à l'état s_j et accède à la ressource rs.
- $\pi = [\ldots \pi_i \ldots]$, la matrice de distribution des probabilités initiales, la valeur $\pi_i = P(s_i)$ désigne la probabilité pour que l'utilisateur soit initialement à l'état s_i.

Les différentes probabilités sont obtenues comme suit :

1. $\pi_i = P(s_i) = \frac{|\varphi(s_j)|}{|SL_c|}$ avec :
   - $SL_c = \cup_{i\in 1,\ldots,ns} E_i$ = l'ensemble total des séquences des intentions de recherches candidates.
     - E_i=$\cup_{t\in 1,\ldots,N_q} SC_{i,t}$ = l'ensemble des séquences des intentions de recherches candidates relatives à la session SL_i.
     - $SC_{i,t}$= $\{s$ / $P(q_t| s) \neq 0, q_t \in SC_{i,t}\}$.
   - $\varphi(s_j)$= les séquences dans SL_c commençant par s_j.
2. $b_j(q) = P(q| s_j) = \frac{\sum_{rs\in RS_j} CP(rs,q)}{\sum_{q\in Q_j} \sum_{rs\in RS_j} CP(rs,q)}$ avec :

   $CP(rs, q)$ = le nombre de fois où la ressource rs a été accédée suite à la soumission de la requête q.
3. $b_k(rs) = P(rs| s_k) = \frac{\sum_{q\in Q_k} CP(rs,q)}{\sum_{q\in Q_k} \sum_{rs\in RS_k} CP(rs,q)}$.
4. $a_{i,j} = P(s_j \mid s_i) = \frac{CS(s_i,s_j)}{NC}$ avec :
   - NC = le nombre d'occurrences de s_j dans SL_c.
   - $CS(s_i, s_j)$= le nombre de fois où l'état s_i est suivi par l'état s_j dans SL_c.

Une fois que la construction du MMC est terminée, PREINTRECH procède à la phase suivante, à savoir la phase de matching et de prédiction explicitée dans la prochaine section.

3.3 Phase de matching et de prédiction

La phase de matching et de prédiction se déroule en deux étapes à savoir, ***(i)*** le matching du contexte courant de la requête utilisateur avec l'un des états du MMC ; ***(ii)*** la prédiction de

l'état de transition de ce dernier. Ainsi, à la soumission d'une requête q par un utilisateur idu_i, PREINTRECH commence par chercher l'état s_{MS} le plus probable de représenter q. Ceci est effectué en calculant, pour chaque état s_i du MMC, la valeur de $Mat_i = \Pi_i \times b_i(q)$. Ainsi, l'état s_{MS} correspond à l'état ayant la plus grande valeur de Mat_i.

Une fois l'état s_{MS} identifié, PREINTRECH procède à la prédiction de l'état s_{NextS}, *i.e.,* qui décrit l'intention de recherche de l'utilisateur idu_i, représentant l'état le plus probable auquel l'état s_{MS} pourrait transiter. Ceci est effectué en calculant la valeur de l'indice $NextS$ comme suit : $NextS = argmax_j\{a_{\{MS,j\}} \times b_j(q)\}$ avec q désigne une requête appartenant à l'état s_j successeur de s_{MS} dans le MMC.

Par conséquent, s_{NextS} désigne l'intention de recherche de l'utilisateur idu_i après avoir soumis la requête q.

4 Exemple illustratif de PREINTRECH sur un jeu de donné réel

Considérons le tableau 2 représentant 5 intentions de recherches, *i.e.*, IT_1, IT_2, IT_3, IT_4 et IT_5 extraites par l'algorithme TRIAS à partir d'un petit jeu de données réelles collecté à partir de *del.icio.us* [1].

IT_1 = {{*css css3 webdesign standards tutoriel inspiration*}, {*rs1=<http ://www.cssreboot.com/>*}, {*IDuser1 IDiuser2 IDuser3 IDuser4*}},
IT_2 = {{*javascript tutoriel webdev tools api ajax web programming*}, {*rs2=<http ://24ways.org/advent/edit-in-place-with-ajax>, rs3=<http ://javascript.internet.com/>, rs4=<http ://www.programmableweb.com/>*}, {*IDuser1 IDiuser3 IDuser4 IDuser10 IDuser11*}}
IT_3 = {{*google howto maps googlemaps google.maps add mashups archives*}, {*<http ://stuff.rancidbacon.com/gmaps-standalone/>, <http ://googlemapsmania.blogspot.com/>* }, {*IDuser4 IDiuser5 IDuser6 IDuser13*}}
IT_4 = {{*media search google video searchengine technology web audio*}, {*rs5=<http ://video.google.com/>, rs6=<http ://searchenginewatch.com/>, rs7=<http ://www.lexxe.com/>* }, {*IDuser4 IDiuser5 IDuser6* }}
IT_5 = {{*audio download fun google mp3 music tools* }, {*rs8=<http ://www.music-map.com/>, rs9=<http ://sploitcast.com/>*}, {*IDuser5 IDiuser6 IDuser8 IDuser10 IDuser12*}}

TAB. 2 – *Intentions de recherches*

Chaque intention de recherche est représentée par un triplet, *i.e.,* l'ensemble des requêtes fréquemment utilisées par un ensemble d'utilisateurs pour la recherche d'un ensemble de ressources. La matrice de transitions correspondante, *i.e.,* A, ainsi que les distributions des différentes probabilités d'observation (de ressources et de requêtes) sont obtenues par PREINTRECH par le calcul des probabilités comme indiqué dans la section 3.2. Le MMC à cinq états correspondant est illustré par la figure 3. La distribution des probabilités initiales obtenue est définie par : $\pi = \begin{pmatrix} 0,2 & 0,2 & 0,2 & 0,2 & 0,2 \end{pmatrix}$.

1. http://www.delicious.com

$$A = \begin{pmatrix} 0,4 & 0,6 & 0,0 & 0,0 & 0,0 \\ 0,2 & 0,5 & 0,3 & 0,0 & 0,0 \\ 0,0 & 0,0 & 0,3 & 0,2 & 0,5 \\ 0,0 & 0,0 & 0,0 & 0,3 & 0,7 \\ 0,0 & 0,0 & 0,0 & 0,6 & 0,4 \end{pmatrix}$$

Ainsi, partant de l'intention de recherche représentée par l'état s_1, les utilisateurs ont une probabilité de 0,4 de garder cette même intention et une probabilité de 0,6 de passer à l'intention de recherche représentée par l'état s_2. Ainsi, en supposant qu'un utilisateur soumette la requête *"audio"*, PREINTRECH commence par chercher l'état ayant la plus grande probabilité de représenter cette requête et ce, en calculant pour chacun des cinq états, la quantité Mat_i = $\pi_i \times b_i(audio)$ à savoir :

$Mat_1 = \pi_1 \times b_1(audio) = 0,2 \times 0 = 0$;

$Mat_2 = Mat_3 = 0$;

$Mat_4 = \pi_4 \times b_4(audio) = 0,2 \times 0,05 = 0,01$;

$Mat_5 = \pi_5 \times b_5(audio) = 0,2 \times 0,2 = 0,04$.

Par conséquent, l'état s_5 représente l'intention de recherche de l'utilisateur pour la requête *"audio"*. Ainsi, les ressources candidates (*rs8 :<www.music-map.com/>*) et (*rs9 :<www.sploitcast.com/>*), avec les probabilités respectives de 0,6 et 0,4, sont recommandées à l'utilisateur. Par ailleurs,

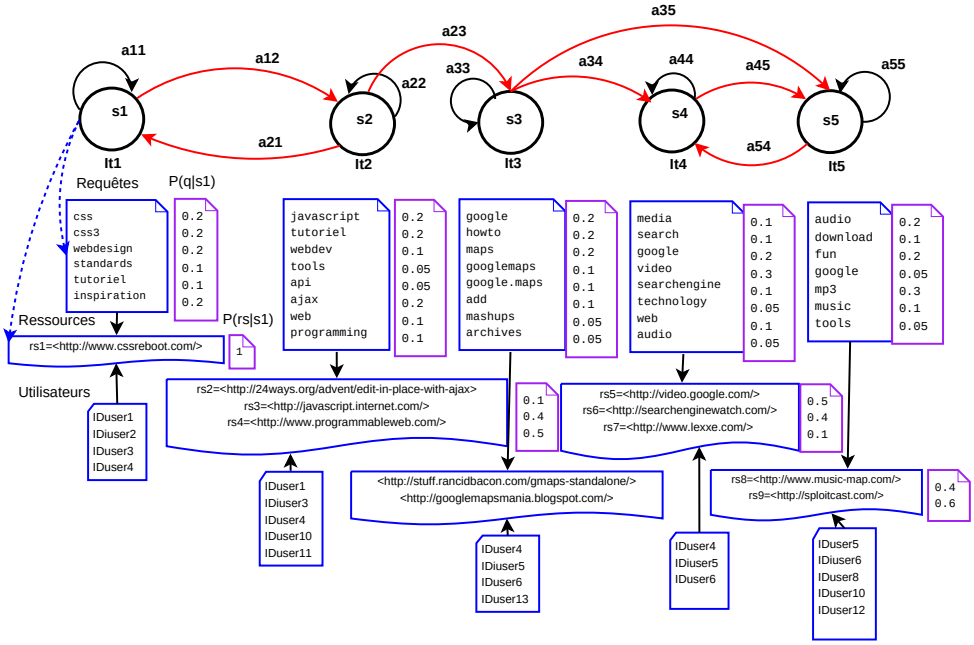


FIG. 3 – *Exemple illustratif de* PREINTRECH

les états de transitions possibles à partir de l'état s_5 sont soit s_4 ou s_5. Ainsi, les requêtes prédites candidates retournées par PREINTRECH, suite à la requête *"audio"*, sont égales à $argmax_j\{a_{\{5,j\}} \times b_j(q)\}$ avec $j \in \{4,5\}$ et q est une requête appartenant aux intentions de recherches représentées par les deux états s_4 et s_5.

En effet, étant donné que d'une part : $Max(P(q)|s_5) = 0,3$ et $Max(P(q)|s_4) = 0,3$ et que d'autre part $argmax_j\{a_{\{5,5\}} \times b_5(q), a_{\{5,4\}} \times b_4(q)\} = argmax_j\{0,12, 0,36\} = 4$ alors l'intention de recherche prédite est représentée par l'état d'indice 4. En se basant sur l'intention de recherche prédite, les requêtes {"vidéo", "media", "google",... }, appartenant à l'intention de recherche représentée par l'état s_4, peuvent être suggérées à l'utilisateur.

5 Etude expérimentale

Le problème de l'évaluation de notre approche est assez complexe étant donné que d'une part aucun outil de référence standard n'est disponible et que d'autre part aucune autre approche n'a exploité les MMCs pour la prédiction des intentions de recherches dans les *folksonomies*. Ainsi, afin de souligner la portée de notre approche aux différentes applications de prédiction des intentions de recherches tels que la recommendation de ressources ou la suggestion de requêtes, nous allons nous basés sur les métriques usuelles issues de la recherche d'information (Baeza-Yates et Berthier (1999)), *i.e.,* les mesures de *Rappel* et de *Précision*.

5.1 Description du jeu de données

Pour mener notre étude expérimentale, nous avons considéré une *folksonomie* collectée à partir du site *del.icio.us*. En effet, mis en ligne depuis le mois de Septembre 2003, *del.icio.us* est un site de partage collaboratif à la fois populaire et en pleine expansion permettant aux usagers de partager des marques pages internet et de librement les annoter avec un ou plusieurs tags ou mots clés. Nous avons ainsi exploité un jeux de données [2] tiré à partir de *del.icio.us* et constitué d'un ensemble daté d'annotations effectuées par 1518 utilisateurs sur 12813 ressources avec 5621 tags. Nos expérimentations ont alors été conduites sur la *logsonomie* correspondante, *i.e,* les tags représentent les requêtes, les utilisateurs correspondent aux identifiants de sessions (IDs) et les ressources annotées définissent les ressources accédées.

5.2 Evaluation de PREINTRECH

Pour l'évaluation de PREINTRECH, nous avons utilisé la même méthode que celle utilisée pour une tâche d'apprentissage supervisé. Ainsi, nous avons divisé l'ensemble des données existantes, *i.e.,* la *logsonomie* $\mathcal{L}$, en deux ensembles de données, *i.e.,* un ensemble d'apprentissage contenant un ensemble de requêtes Q_{T-1} et un ensemble de test contenant un ensemble de requêtes Q_T. Ainsi, toutes les sessions utilisateurs enregistrées entre le 1 décembre 2005 et le 15 janvier 2006 ont été utilisées en tant qu'ensemble d'apprentissage pour la construction du MMC. Le reste des sessions utilisateurs, *i.e.,* enregistrées entre le 16 janvier 2006 et le 20 décembre 2007, a été considéré comme une base de test.

Les résultats obtenus par PREINTRECH sur la base de test, ont été par la suite comparés avec les requêtes représentant les intentions de recherches réelles des utilisateurs. Supposons que suite à la requête $q_T \in Q_T$, PREINTRECH a retourné la liste de requêtes Q_R alors les mesures de *rappel* et de *précision* sont données comme suit :

Rappel = $\frac{|QR \bigcap \{Q_T \backslash q_T\}|}{|Q_T \backslash q_T|}$ et *Précision* = $\frac{|QR \bigcap \{Q_T \backslash q_T\}|}{|QR|}$.

La figure 4 montre les valeurs moyennes de *précision* et de *rappel* en fonction du nombre k des requêtes prédites. Ainsi, on peut relever que la *précision* moyenne augmente considérablement quand le nombre de requêtes prédites diminue et inversement, le *rappel* moyen augmente en fonction de l'augmentation du nombre de requêtes prédites. En effet, pour un $k = 1$, le

2. `http://data.dai-labor.de/corpus/delicious/`

rappel moyen est égal à $0,28$ alors que pour un $k = 5$, le *rappel* est égal à $0,43$ soit une augmentation du *rappel* moyen de 53%. De même, pour un $k = 1$, la *précision* moyenne est de $0,34$ alors que pour un $k = 5$, on a une *précision* moyenne de $0,19$, *i.e.,* une diminution de la *précision* moyenne de 54%. Par ailleurs, PREINTRECH a pu donner des prédictions à 74% des requêtes appartenant à l'ensemble d'apprentissage Q_T et ce, malgré l'absence d'une phase de pré-traitement, *i.e.,* lémmatisation et filtrage des tags de la *logsonomie* considérée.

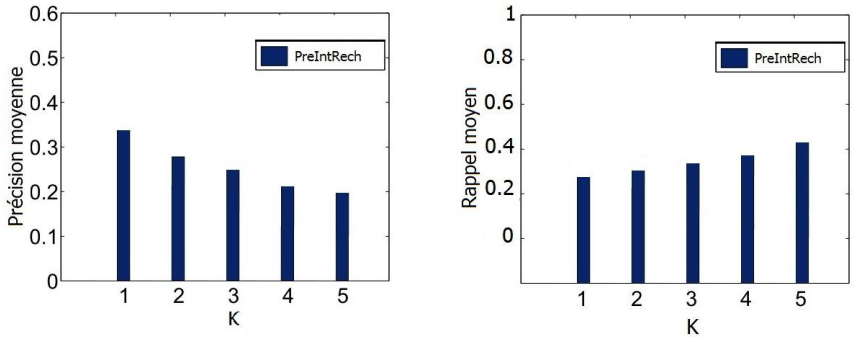


FIG. 4 – *Rappel et précision moyens des requêtes prédites*

6 Conclusion et perspectives

Dans ce papier, et afin de permettre aux usagers l'accès le plus pertinent à l'information à travers la prédiction de leur intention de recherche, nous avons exploité la structure des *folksonomies*, *i.e.,* les triplets (Ressources, Utilisateurs, Tags), pour introduire une nouvelle approche de prédiction basée sur l'utilisation des *MMCs*, appelée PREINTRECH. Par ailleurs, étant donné le nombre considérable des triplets dans les *folksonomies* et le nombre d'intentions de recherches pouvant être extraites, nous avons utilisé les concepts triadiques. L'étude expérimentale menée sur la *folksonomie del.icio.us* a mis en exergue que PREINTRECH apporte des predictions pour 74% de l'ensemble des requêtes soumises. Les perspectives de prolongement du présent travail sont comme suit : *(i)* L'intégration d'un nouveau module de stémmatisation et de désambiguïsation sémantique des tags et ce notamment, par l'exploitation des ontologies en ligne et des ressources lexicales externes, *i.e.,* WORDNET et WIKIPÉDIA ; *(ii)* La considération des requêtes qui ne présentent pas de matching avec les tags de la *folksonomie*. Ceci revient à inclure un processus d'auto-complétion notamment en y incluant un module d'indexation de requêtes ; et *(iii)* L'analyse de la possibilité de modifier la requête de l'utilisateur dans le cas où les besoins d'informations de ce dernier ne sont pas entièrement satisfaits. Pour se faire, nous comptons initialiser le MMC par la réestimation des probabilités d'émission des candidats, *i.e.,* les requêtes suggérées et des ressources recommandées, acceptées par l'utilisateur.

Références

Amer-Yahia, S., A. Galland, J. Stoyanovich, et C.Yu (2008). From del.icio.us to x.qui.site : recommendations in social tagging sites. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008*, Vancouver, pp. 1323–1326. ACM Press.

Baeza-Yates, R. et R.-N. Berthier (1999). *Modern Information Retrieval*. Boston, MA, USA : Addison-Wesley Longman Publishing Co., Inc.

Benz, D., A. Hotho, R. Jäschke, B. Krause, et G. Stumme (2010). Query Logs as Folksonomies. *Datenbank-Spektrum, Volume 10*(1), pp. 15–24.

Bischoff, K., C. S. Firan, W. Nejdl, et R. Paiu (2008). Can all tags be used for search ? In *Proceedings of the 17th ACM conference on Information and knowledge management, CIKM 2008*, Napa Valley, California, pp. 193–202. ACM Press.

Deuff, O. L. (2007). Folksonomies et communautés de partage de signets : Vers de nouvelles stratégies de recherche d'informations. In *Proceeding of the workshop Collaborer, Echanger, Inventer : Expériences de réseaux, h2ptm 2007*, Tunisia.

Fonseca, B., P. Golgher, B. Pôssas, B. Ribeiro-Neto, et N. Ziviani (2005). Concept-based interactive query expansion. In *Proceedings of the 14th ACM international conference on Information and knowledge management, CIKM 2005*, Bremen, Germany.

Garg, N. et I. Weber (2008). Personalized tag suggestion for flickr. In *Proceedings of the International Conference on World Wide Web, WWW 2008*, Beijing, pp. 1063–1064. ACM Press.

Jäschke, R., A. Hotho, C. Schmitz, B. Ganter, et G.Stumme (2008). Discovering shared conceptualizations in folksonomies. *Web Semantics : Science, Services and Agents on the World Wide Web, Volume 6*, pp. 38–53.

Jäschke, R., A. Hotho, C. Schmitz, B. Ganter, et G. Stumme (2006). TRIAS - an algorithm for mining iceberg tri-lattices. In *Procedings of the 6th IEEE International Conference on Data Mining, ICDM 2006, IEEE Computer Society, Hong Kong*, pp. 907–911.

Jones, R., B. Rey, O. Madani, et W. Greiner (2006). Generating query substitutions. In *Proceedings of the 15th International Conference on World Wide Web, WWW 2006*, Edinburgh, Scotland, pp. 387–396.

Karen, H., T. Sutter, L. B. Marinho, et L. Schmidt-Thieme (2008). Tag-aware recommender systems by fusion of collaborative filtering algorithms. In *In ACM symposium on Applied computing, SAC 2008*, New York, USA, pp. 1995–1999. ACM Press.

Krause, B., A. Hotho, et G. Stumme (2008a). A comparison of social bookmarking with traditional search. In *Proceedings of the 30th European Conference on IR Research, Advances in Information Retrieval, ECIR 2008*, Volume 4956, pp. 101–113. Springer.

Krause, B., R. Jäschke, A. Hotho, et G. Stumme (2008b). Logsonomy - social information retrieval with logdata. In *Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia, HT 2008*, New York, NY, USA, pp. 157–166.

Lehmann, F. et R. Wille (1995). A triadic approach to formal concept analysis. In *Proceedings of the 3rd International Conference on Conceptual Structures : Applications, Implementation and Theory*, pp. 32–43. Springer.

Mei, Q., J. Jiangy, H. Suz, et C. Zhai (2007). Search and tagging : Two sides of the same coin ? In *Technical Report No. 2919, University of Illinois at Urbana-Champaign (UIUCDCS-R-2007-2919)*.

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications inspeech recognition. In *Proceedings of the IEEE*, Volume 77(2), New Orleans, Louisiana, pp. 257–286.

Summary

The development of collaborative tagging systems on the Internet under the participatory movement often called "Web 2.0", has enabled the emergence of new forms of indexing web content created by users and freely shared within social networks, these structure are called folksonomies. Thus, to allow access to the most relevant information, we propose a new context based approach, by learning the tags previously queried by users. Specifically, we adopted Hidden Markov Models and formal concept analysis to predict user's search intents based on a a real *folksonomy*. Conducted experiments emphasize the relevance of our proposal and open many issues.

Vers un système de recommandation sémantique et social dans un réseau professionnel

Maria Malek, Hubert Kadima, Dalia Sulieman

LARIS-EISTI
PRES Université de Cergy
prenom.nom@eisti.fr

Résumé. Nous présentons dans cet article un système de recommandation de talents dans un réseau social professionnel. Le réseau est composé d'un ensemble de personnes ayant des liens professionnels. Selon la demande d'un acteur d'origine X, le système doit proposer un ou plusieurs acteurs Z répondant au mieux aux critères demandés. Nous proposons deux algorithmes de recommandation qui utilisent les trois types d'information suivants : le premier type correspond à l'information stockée sur la personne d'une façon décentralisée au niveau de chaque nœud qui consiste à la description sémantique du profil utilisateur. Le deuxième type d'information est celui qui utilise la structure du réseau même. Notre contribution consiste à utiliser l'arbre couvrant du graphe pour améliorer la recherche. Le troisième type correspond à l'information stockée dans les mesures sociales liées aux acteurs intermédiaires passant par les chemins retenus. Cette heuristique donne plus d'importance aux chemins ayant des acteurs plus prestigieux. Nous appliquons et évaluaons notre algorithme sur un graphe de collaborations scientifiques entre auteurs appartenant à la même communauté scientifique.

1 Introduction

L'analyse des réseaux sociaux est définie comme étant l'étude des entités sociales (les personnes dans les organisations qu'on appelle acteurs) ainsi que leurs interactions et leurs relations. Ces interactions et relations peuvent être représentées par un graphe ou un réseau, dans lequel chaque nœud représente un acteur et chaque lien est une relation. Nous pouvons étudier les propriétés (Freeman (1979), Everett et Borgatti (1999)) de la structure et son rôle ainsi que la position et le prestige de chaque acteur social (Newman (2003)). Nous pouvons rechercher aussi les différents types de sous-graphes comme par exemple les communautés formées par des groupes d'acteurs ayant des intérêts communs, en isolant le groupe d'individus ayant une densité élevée. Le réseau social peut être aussi une source permettant l'élaboration de recommandations : trouver un expert dans un domaine donné, suggérer des produits à vendre, proposer un ami, etc. Cette élaboration peut être fondée sur des algorithmes d'exploration de chemins, d'analyse de degrés (RSL (2009), Ereteo et al. (2009)),etc.

Les réseaux sociaux peuvent être modélisés par des graphes (Fournier (2006); Berge (1983)) ayant des propriétés spécifiques.

Nous présentons dans cet article un système de recommandation de talents dans un réseau social professionnel. Le réseau est composé d'un ensemble de personnes ayant des liens professionnels. Selon la demande d'un acteur d'origine X, le système doit proposer (recommander) un ou plusieurs acteurs Z répondant au mieux que possible aux critères demandés (exemple : recherche d'une personne ayant des compétences données pour un poste, etc.). Nous proposons un algorithme de recommandation qui utilise les trois types d'information suivants :

- L'information stockée sur la personne (l'acteur ou le nœud du graphe) d'une façon décentralisée au niveau de chaque nœud. Cette connaissance peut être représentée en utilisant une ontologies décrivant les profils utilisateurs.
- L'information décrite par la structure du réseau même. Autrement dit, en explorant les liens partants de l'acteur origine x et en utilisant des algorithmes d'exploration de graphe comme les chemins les moins coûteux, nous pouvons délimiter le champ de recherche de l'ensemble Z. Notre contribution consiste à utiliser l'arbre couvrant du graphe.
- L'information stockée dans les mesures liées aux acteurs intermédiaires passant par les chemins retenus. Cette heuristique donne plus d'importance aux chemins ayant des acteurs plus prestigieux.

Le reste de cet article est organisé ainsi : nous décrivons dans la suite notre système de recommandation, nous proposons ensuite une réalisation du système pour la recommandation d'auteurs dans le domaine de la bibliographie scientifique. Nous exposons finalement les résultats, les perspectives et la conclusion.

2 Description du système de recommandation

Notre but est de proposer un algorithme de recommandation efficace dans notre réseau social professionnel composé d'un ensemble de personnes ayant des liens professionnels. L'application d'une telle approche peut être dans la recherche d'une expertise ou bien d'une collaboration scientifique ou professionnel.

Le principe du système est de proposer (recommander) un ou plusieurs acteurs répondant au mieux aux critères demandés à partir d'une requête posée par un utilisateur X qui est lui même acteur dans le réseau social.

2.1 Idée de l'algorithme

l'idée étant de proposer un algorithme de recherche qui combine la sémantique, la structure & les propriétés des réseaux sociaux :

La partie Sémantique consiste en : l'information stockée sur la personne (l'acteur) d'une façon décentralisée au niveau de chaque nœud : *le profil utilisateur*.

La partie Structure est constituée de : l'information décrite par la structure du réseau même : nous utilisons la technique de l'arbre couvrant minimum (ou maximum).

La partie Propriétés du réseau consiste à utiliser : l'intermédiarité des acteurs passants par les chemins retenus : *acteurs prestigieux*.

2.1.1 La partie sémantique

La partie sémantique consiste à calculer la mesure de similarité entre la requête posée par l'acteur X et le profil utilisateur stocké en unnœud donné :

- R_X est la requête posée par le sommet X sous forme d'un ensemble de termes T_i : $R_X = \{T_1, T_2.., T_n\}$. T_i étant un terme donné, P_i le poids associé.
- Pro_Z est le profil associé à un sommet donné Z donné également par un ensemble de termes pondérés : $Pro_Z = \{(T_1, P_1), (T_2, P_2)..,(T_m, P_m)\}$.
- Nous définissons la similarité (*la pertinence*) entre la requête R_X et le profil du sommet Pro_Z par :

$$sim(R_X, Pro_Z) = \frac{\sum\limits_{j \in inter(R_X, Pro_Z)} Pro_Z.P_j}{\sum\limits_{j=1}^{m} Pro_Z.P_j + |R_X \text{ diff } Pro_Z|}$$

.

- avec : $inter(R_X, Pro_Z) = \{k \in \{1, ..m\}, Pro_Z.T_k \in R_X\}$
  la fonction $inter$ calcule les termes en commun entre la requête et le profil.
  La fonction $diff$ correspond à la différence ensembliste. Cette mesure est une adaptation de la mesure de similarité de Jaccard dans un contexte de termes pondérés.

2.1.2 La partie structure

Il s'agit de trouver l'arbre couvrant maximum à partir d'un graphe valué en utilisant une version adaptée de l'algorithme de Kruskal.

Le but est d'améliorer la recherche en effectuant une navigation optimisée dans l'arbre couvrant au lieu d'explorer le graphe ou une partie du graphe. L'arbre couvrant maximum sera l'arbre couvrant le plus représentatif dans notre graphe (voir 2.2.1)

2.1.3 Intermédiarités des nœuds

- Si deux nœuds *non adjacents* k & j qui se communiquent et si le nœud i se trouve sur le chemin de communication alors *i est un acteur intermédiaire*.
  - $C_B(i) = \sum\limits_{j<k} \frac{p_{jk}(i)}{p_{jk}}$
  - p_{jk} le nombre des chemins les plus court entre j et k,
  - $p_{jk}(i)$ le nombre des chemins les plus court entre j et k passant par i.

L'utilisation de la mesure de l'intermédiarité va nous permettre de privilégier certains chemins de recherche par rapport à d'autres.

2.2 L'algorithme de recommandation

Nous proposons un algorithme ayant l'entrée et la sortie suivantes :

Entrée : une requête posée par l'auteur X formulée par une suite de mots (termes) clés.

Sortie : une suite pondérée d'auteurs $\{(Z_1, P_1), (Z_2, P_2)..,(Z_n, P_n)\}$ correspondants au mieux à la requête *ainsi que* : la chaîne sémantique reliant les deux auteurs : une chaîne sémantique reliant deux auteurs X, Z_i est constituée de la liste de mots (termes) clefs se trouvant dans la suite des sommets reliant X à Z_i.

2.2.1 Etapes de l'algorithme

1. Trouver l'arbre couvrant maximum par rapport aux poids des arêtes. Dans notre cas, il s'agit de trouver l'arbre couvrant le plus représentatif.
2. Calculer et stocker les intermédiarités des nœuds.
3. Extraire de l'arbre couvrant une liste de sommets triée à recommander en utilisant l'algorithme exhaustif ou l'algorithme guidé détaillés par la suite.

2.2.2 Algorithme exhaustif pour la recherche de toutes les solutions

1. Recherche dans A (l'arbre couvrant) des sommets Z_i à recommander à X à partir de sa requête (voir figure 1), et ceci en effectuant un parcours en largeur dans A : trouver un ensemble $\{Z_1, Z_2, .., Z_n\}$ tel que $sim(R_X, Pro_{Z_i}) >= seuil$.
2. Nous associons à chaque Z_i proposé un poids "rating" qui exprime l'importance de la recommandation ; ce poids dépend de la similarité entre la requête et le profil de Z_i ainsi que de l'intermédiarité des nœuds se trouvant sur le chemin de la solution :
   - Soient $[Y_1, Y_2, .., Y_l]$ la liste des sommets se trouvant sur la chaîne reliant X à Z_i.
   - P_i étant le *rating* à associer au sommet Z_i, P_i est calculé par :

     $P_i = sim(R_X, Pro_{Z_i}) * \frac{\sum_{j=1}^{l} intermediarite(Y_j)}{l}$ **si** $l > 1$

     $P_i = sim(R_X, Pro_{Z_i})$ **sinon**

2.2.3 Algorithme guidé pour la recherche d'une solution

Nous proposons une deuxième version moins coûteuse permettant de trouver une solution, en trouvant rapidement le chemin de la recherche dans l'arbre couvrant au lieu d'effectuer un parcours en largeur.

Nous utilisons une heuristique permettant de choisir le sommet à visiter parmi un ensemble de sommets candidats et nous appliquons ensuite un algorithme de type A*, permettant de passer à chaque étape par le sommet Y maximisant l'heuristique :

$$h(Y) = (seuil - sim(Pro_X, Pro_Y)) * intermediarite(Y)$$

jusqu'à ce qu'on arrive à un sommet Z à recommander pour lequel nous avons :

$$sim(X, Z) >= seuil.$$

Nous démontrons théoriquement que notre heuristique est monotone, nous démontrons également qu'elle reconnaît la solution, car elle prend la valeur nulle pour la solution.

D'un autre côté, nous montrons par l'expérimentation que cette version converge plus rapidement vers la solution et permet d'explorer de 11 à 49 % de l'arbre couvrant en comparaison avec la version exhaustive.

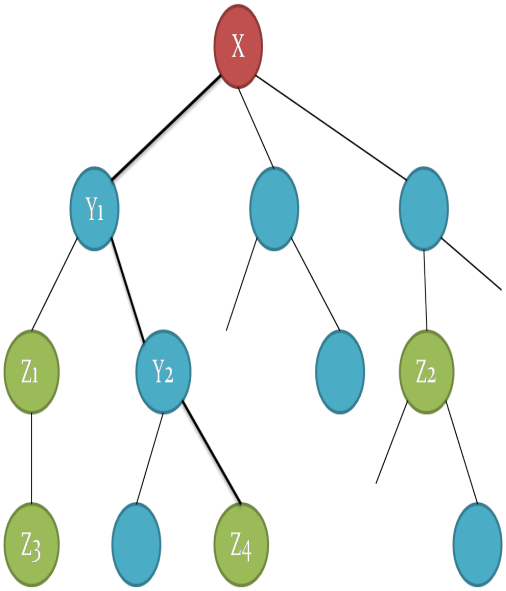


FIG. 1 – *Exemple de déroulement de l'algorithme de recommandation dans un arbre couvrant : X étant l'auteur qui soumet la requête ; une recherche des sommets similaires est effectuée, soient* Z_4, Z_3, Z_1, Z_2 *cette liste : un poids qui dépend de la mesure d'intermédiartié des auteurs sur le chemin* (X, Z_i) *est affecté à chaque sommet retenu ; la liste de sommets à recommander est* $[(Z_4, P_4), (Z_3, P_3), (Z_1, P_1), (Z_2, P_2)]$.

3 Etapes de réalisation du système

Le but de cette section est de montrer l'application de notre approche sur un réseau professionnel composé d'un ensemble d'auteurs de références bibliographiques. Le réseau est un graphe de similarité dont les nœuds sont les auteurs. Un lien est créé entre deux auteurs s'ils sont "structurellement" similaires. Deux auteurs sont "structurellement" similaires s'ils citent un certain nombre d'articles en commun *et/ou* s'ils sont cités par une certain nombre d'articles. Nous analysons le site de références bibliographique *libra.msra.cn* : *communauté datamining*.

Nous procédons tout d'abord par l'extraction du *graphe de citations* entre auteurs et ensuite le graphe de similarité structurelle entre auteurs est extrait à partir du graphe de citations.

3.1 Graphe de citations

Par simplification, nous nous limitons à l'ensemble des publications effectuées à partir de l'année 2005.A partir de l'analyse des citations, le graphe de citations est extrait.

Le graphe de citations entre auteurs est un graphe dirigé dans lequel : les nœuds sont les auteurs et les liens dirigés sont les citations entre auteurs pondérées par leur nombre.

Nous stockons au niveau de chaque acteur-auteur un vecteur *pondéré* de mots clefs qui constituera le profil utilisateur. Ce vecteur est extrait (actuellement) en effectuant une simple analyse de texte constituant les titres des articles.

3.2 Extraction du graphe de similarité structurelle

Le graphe de de similarité structurelle est un graphe non dirigé extrait à partir du graphe de citations précédent : les nœuds de ce graphes sont les auteurs. Une arête entre deux auteurs exprime la similarité structurelle entre deux auteurs.

Nous rappelons que deux auteurs sont "structurellement" similaires s'ils citent un certain nombre d'articles en commun *et/ou* s'ils sont cités par une certain nombre d'articles. Le graphe de similarité structure est construit à partir de deux mesures qui sont : le couplage bibliographique ainsi que la co-citation détaillés à la suite.

3.2.1 Les mesures de citations bibliographiques

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

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

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

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

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

Le couplage bibliographique Le couplage bibliographique est une mesure de similarités entre deux documents qui exprime le fait que si les papiers i et j citent le papier k alors ils sont liés. De même, si les papiers i et j citent plusieurs papiers alors ils sont *similaires*. Les citations sont modélisées par une matrice L appelée la matrice de citation ; le terme L_{ij} vaut 1 si i cite j, sinon il vaut 0. Le couplage bibliographique est une mesure qui est définie par :

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

Remarquer bien que la mesure du couplage bibliographique est symétrique.

Le graphe de similarité entre auteurs Nous définissons le graphe de similarité à partir de la somme des deux matrices C et B qui représentent simultanément la co-citation et le couplage bibliographique. Une relation de similarité structurelle est crée entre les deux auteurs i et j ssi $[B + C][i][j] >= seuil$ (voir figure 2).

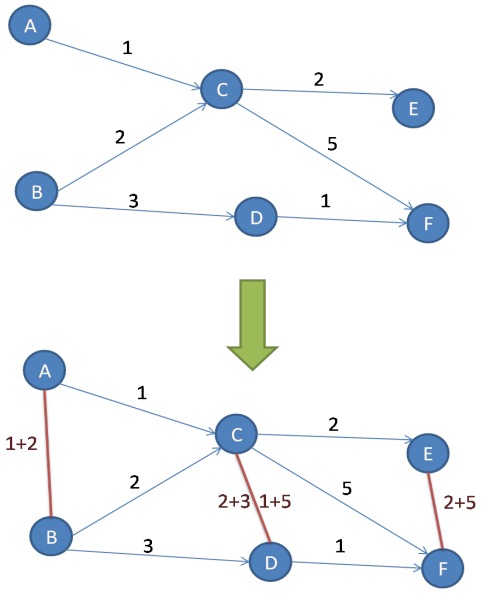


FIG. 2 – *Du graphe de citations vers le graphe de similarité*

4 Expérimentations

Nous démontrons quelques résultats visuels du graphe et de l'arbre couvrant obtenu (voir figure 3).

Nous avons par ailleurs étudié la loi de distribution des degrés du graphe de similarité et nous avons démontré qu'elle obéit à la loi de distribution en puissance ce qui confirme encore la nature sociale de ce graphe.

Nous présentons un exemple d'une requête : nous supposons que l'auteur "Francesco Masulli " soumet une requête composée des trois termes : T_1= Ranking, T_2= Clustering, T_3= Data mining : en appliquant l'algorithme exhaustif les résultats mentionnés dans le tableau 1 montrent une liste d'auteurs recommandés triés selon leurs rangs.

Nous avons également évaluer la version guidée par rapport à la version exhaustive en procédant ainsi sur dix expériences : nous avons élaboré un ensemble de requêtes à tester par un auteur X (qui devient la racine de l'arbre couvrant) en utilisant les termes trouvés dans la communauté. Ensuite, pour chaque requête nous avons appliqué les deux versions de l'algorithme et relevé les mesures suivantes :

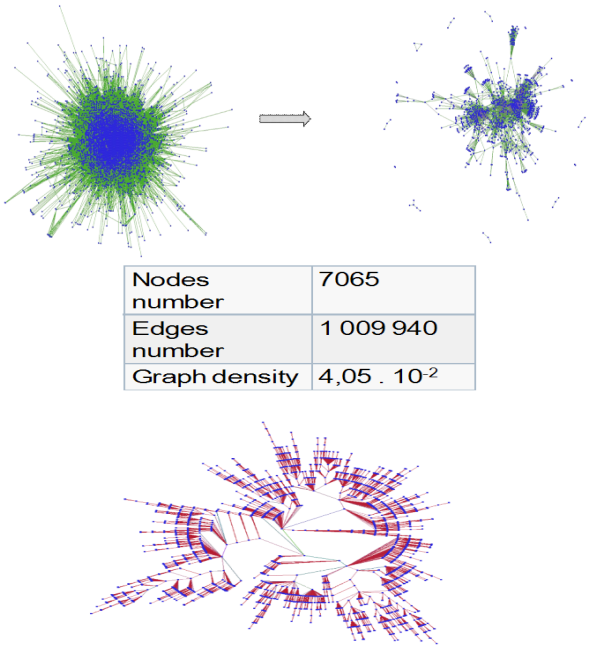


FIG. 3 – *Cette figure montre deux versions du graphe de similarité extraits du site microsoft, ainsi que quelques statistiques démontrant que le graphe est social, l'arbre couvrant est ainsi présenté en bas.*

Author	Rating	Distance
Mikolaj Morzy	0.0072	2
Steven Warner	0.0005	2
Bob Garcia	0.00037	2
Wendy Gersten	0.00029	3
Manuel Lozano	0.00027	2
Matthias Schonlau	9e-005	2
Lyane T Watson	7.89e-005	2
Carl Wunsch	6.78e-005	2
Yang Seok Kim	3.38e-005	2
David W Aha	2.39e-005	2

TAB. 1 – *Résultats de la requête : auteurs à recommander triés par leurs rangs ainsi que la distance séparant chaque auteur recommandé de l'auteur racine.*

Le rang de l'auteur trouvé par l'algorithme guidé par rapport à l'algorithme exhaustif. Nous rappelons que la version exhaustive propose pour la même requête une liste triée d'au-

teurs à recommander.

Le nombre de sommets parcourus par l'algorithme guidé.

Le temps de calcul .

Nous avons trouvé que pour 8 expériences sur 10 le rang numéro 1 a été trouvé par la version guidée tandis que pour les 2 expériences restantes le rang numéro 2 a été trouvé (voir tableau 2). L'arbre couvrant n'a pas été recherché en totalité par la version guidée, l'espace de recherche a été réduit de 11% à 49%, ce qui a réduit considérablement le temps de calcul.

N	L'algorithme exhaustif			L'algorithme A*		
	Auteur recommandé	**Rang**	**Temps de calcul**	**Auteur recommandé**	**Temps de calcul**	**Graphe exploré**
1	Andrew Emili	0.00064	159,41s	Andrew Emili (1)	109,27s	39.25%
2	G V Belle	0.00141	159,35s	G V Belle (1)	17,45s	21.13%
3	Hans A Kestler	0.00060	150,41s	*Yuichi Asahiro (2)*	11,66s	13.86%
4	Jimin Pei	0.00002	160,61s	Jimin Pei (1)	32,52s	20.02%
5	John F Canny	0.00003	159,99s	John F Canny (1)	21,77s	11.77%
6	C Wang	0.00010	157,37s	C Wang (1)	233,99s	49.13%
7	J Michael Brady	0.00001	162,68s	J Michael Brady (1)	118,74s	41.14%
8	Peter G Neumann	0.00022	160,72s	*Elizabeth J O neil (2)*	40,49s	24.88%
9	Peter Eades	0.00004	153,95s	Peter Eades (1)	54,47s	30.95%
10	Liang Chen	0.00019	161.71s	Liang Chen (1)	14,14s	16.67%

TAB. 2 – *Comparaison entre la recherche en largeur et l'algorithme A* : chaque ligne est le résultat d'une requête envoyée par l'auteur racine. Les auteurs recommandés par A* sont indiqués avec leurs rangs trouvés par l'algorithme exhaustif. Nous remarquons que pour 8 requêtes sur 10 l'auteur trouvé par A* est celui ayant eu le rang numéro 1 par l'algorithme exhaustif.*

5 Travaux reliés

Les algorithmes de recherche dans les graphes ont été utilisés pour la recommandation d'experts dans les réseaux sociaux. Nous citons les stratégies suivantes(Zhang et Ackerman (2005)) :

Exploration en largeur qui diffuse la requête pour chaque acteur dans le réseau social en suivant une exploration par largeur.

Recherche aléatoire qui choisit au hasard un voisin à qui de se propager la requête.

Recherche de meilleure connexion proposé par Adamic et al. (2003) et qui fait usage de la distribution des degrés au sein du réseau social.

Les algorithme de faibles et de forts liens qui sont basés sur le fait que les liens entre les deux individus peuvent avoir différentes degrés de forces. La force du lien varie et n'est pas toujours symétrique.

La recherche fondée sur la distance de Hamming qui choisit parmi les voisins ceux qui ont les moindre d'amis en commun avec l'acteur actuel.

La recherche fondée sur l'information qui choisit l'acteur dont le profile est le plus similaire à la requête.

La recherche d'expertise dans les réseaux sociaux ont été abordé dans les travaux de Zhang et Ackerman depuis 2005 (Zhang et Ackerman (2005)).

Les stratégies de recherche dans un graphe ont été appliquées et évaluées sur les données des mails (Campbell et al. (2003)) et plus particulièrement la base de Enron(Zhang et Ackerman (2005)). Les critères d'évaluation étaient : le nombre de personnes trouvées par requête, la profondeur de la chaîne de recherche, etc. Les expériences ont montré que la recherche fondée sur l'information n'est pas plus performante que les stratégies fondées sur les liens sortants comme le parcours en largeur ou bien la distance de Hamming. Les stratégies de liens faibles peuvent être très utiles pour trouver les nouvelles informations.

Dans (Zhang et al. (2007)) la recommandation est formalisée comme étant un problème de tri dans un réseau social hétérogène. La recherche aléatoire est utilisée pour naviguer dans ce type de réseau. D'un autre côté, Newman (Newman (2004)) a traité le problème de la collaboration scientifique et a étudié les relation entre les auteurs. Il a également étudié les caractéristiques de ce réseau ainsi que sa structure.

Dans Grossman (2002) les auteurs ont étudié la structure du réseau social constitué des papiers publiés dans le domaine des mathématiques. les nœuds étant les auteurs et liens correspondent aux papiers entre eux. L'évolution de ce type de réseau a été également analysé.

6 Intégration de l'approche sémantique

Notre objectif est d'intégrer la représentation d'une ontologie de domaine dans notre système de recommandation. Nous souhaitons utiliser cette ontologie de domaines dans l'élaboration des requêtes posées afin d'aider l'utilisateur à re-formuler sa requête ou à la compléter ainsi que dans la représentation ontologique du profil utilisateur ou des ses préférences. Nous allons dans un premier temps expérimenter une taxonomie décrivant un domaine autrement dit une ontologie limitée à la relation IS-a.

Nous enrichissons notre algorithme de recommandation en ajoutant les parties suivantes :

- Un algorithme qui utilise l'ontologie de domaine pour aider l'utilisateur à re-formuler sa requête ou à la compléter.
- Un algorithme qui extrait la représentation du profil et de préférences des utilisateurs en utilisant la terminologie de l'ontologie du domaine. Autrement dit, il s'agit de trouver une algorithme d'annotation automatique du profil utilisateur par les termes de l'ontologie .
- Proposition d'une nouvelle mesure de similarité qui prend en compte la structure de l'ontologie du domaine. Cette mesure prend en compte le niveau de termes ontologique dans l'arbre et elle se base sur des algorithmes de calcul d'ancêtres commun.

Nous allons dans un premier temps expérimenter notre approche sur des systèmes de recommandation de type sur des données du site Amazon.

Les données disponibles nous permet actuellement d'extraire :

- l'ontologie décrivant la taxonomie des catégories des différents livres.
- L'annotation de chaque livre acheté par un ensemble de catégories appartenant à l'ontologie.

– Un graphe de similarité entre les utilisateurs, par rapport à leurs goûts similaires dans le choix des livres ou les DVDs.
– les préférences de chaque utilisateur par rapport à des livres ou CD ou DVD achetés exprimés en termes d'instances mais aussi en termes de concepts appartenant à l'ontologie (ou la taxonomie.

7 Conclusion et perspectives

Nous avons présenté dans cet article un système de recommandation de talents dans un réseau social professionnel. Notre réseau est composé d'un ensemble de personnes ayant des liens professionnels. Selon la demande d'un acteur d'origine x, le système doit proposer (recommander) un ou plusieurs acteurs Z répondant au mieux que possible aux critères demandés. Nous avons appliqué notre algorithme sur la recommandation d'auteurs dans un graphe de similarités entre auteurs.

La recommandation dépend d'un côté, de la similarité entre les profils des auteurs et la requête soumise et d'un autre côte de l'intermédiarité des nœuds-auteurs se trouvant sur les chemins de la solution. Pour effectuer une recherche dans le graphe, l'arbre couvrant le plus représentatif est extrait et ensuite il est exploré.

Le premier algorithme est exhaustif, il est fondé sur la recherche en largeur dans l'arbre couvrant, jusqu'à ce qu'on trouve un auteur à recommander. Le deuxième algorithme utilise l'approche A* pour explorer l'arbre couvrant. Nous définissons une heuristique admissible qui dépend de la similarité entre un profil et la requête ainsi de l'intermédiarité des nœuds-auteurs. Les expériences ont montré que la version guidée trouve souvent la meilleure recommandation tout en améliorant les performances de l'exploration. En comparant les deux algorithmes nous remarquons que L'arbre couvrant n'a pas été recherché en totalité par la version guidée, l'espace de recherche a été réduit de 11% à 49%.

Nous travaillons actuellement sur une élaboration plus fine du profil utilisateur en utilisant une représentation ontologique du domaine. Nous souhaitons également étendre l'algorithme pour des recommandations inter-communautés. Nous réfléchissons également sur l'utilisation de l'arbre couvrant pour des fins sémantiques qui peuvent ramener à la découverte des *rapprochements sémantiques* entre les communautés.

Références

(2009). Ecole d'été web intelligence. In *WI09*. Université de Lyon.

Adamic, L. A., O. Buyukkokten, et E. Adar (2003). A social network caught in the web. *First Monday 8*(6).

Berge, C. (1983). *Graphes*. Gauther-Villars.

Campbell, C. S., P. P. Maglio, A. Cozzi, et B. Dom (2003). Expertise identification using email communications. In *CIKM*, pp. 528–531.

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

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

Fournier, J.-C. (2006). *Théorie de Graphes et applications*. Lavoisier.

Freeman, L. C. (1979). Centrality in social networks : Conceptual clarification. *Social Networks 1*(3), 215–239.

Grossman, J. (2002). The evolution of the mathematical research collaboration graph. *Congressus Numeratium*.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review 45*, 167–256.

Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Science of the United States (PNAS) 101*, 5200–5205.

Zhang, J. et M. S. Ackerman (2005). Searching for expertise in social networks : a simulation of potential strategies. In *GROUP*, pp. 71–80.

Zhang, J., M. S. Ackerman, et L. Adamic (2007). Expertise networks in online communities : structure and algorithms. In *WWW '07 : Proceedings of the 16th international conference on World Wide Web*, New York, NY, USA, pp. 221–230. ACM.

Summary

This paper proposes skills recommendation algorithm in a professional social network. This network consists of a set of persons with professional weighted ties. To answer the request of an actor, the system recommends a list of other actors that match the best requested criteria.

We propose two recommendation algorithms based on three types of knowledge: The first type deals with information concerning the person. This information is stored in the actor vertex level and constitutes the user profiles description. The second type of information is computed from the network structure itself. Actually, this consists of exploring the links starting from the initial actor exploring the maximum spanning tree whose the root is the initial actor. We can thus reduce the search space of target actors. While the third type of information is based on the betweenness centrality measure associated to each actor. This measure enables to estimate the control of an actor over other pairs of actors. We use this measure to extract the best paths from the previous spanning tree.

Capitalisation des échanges informels en entreprise via les réseaux sociaux

Étienne Deparis*,**, Marie-Hélène Abel*
Juliette Mattioli**

*Université de Technologie de Compiègne
UMR CNRS 6599 Heudiasyc, Centre de Recherches de Royallieu – BP 20529
60205 Compiègne Cedex
{etienne.deparis, marie-helene.abel}@utc.fr
http://www.hds.utc.fr/
**Decision Technologies & Mathematics Laboratory
Thales Research & Technology, Campus Polytechnique – 1, avenue Augustin Fresnel
91767 Palaiseau Cedex
{etienne.deparis, juliette.mattioli}@thalesgroup.com

Résumé. Les attentes et les pratiques autour de la gestion de connaissances ont évolué au cours des deux dernières décennies, au point que l'on peut parler de génération de knowledge management (KM). La première génération était surtout axée sur la création de bases de connaissances ou de "bonnes pratiques" de l'entreprise. Les informations revêtant un caractère critique pour l'entreprise étaient ainsi capturées, identifiées et stockées dans ces systèmes, dans le but d'être adressée facilement par les collaborateurs qui en auraient besoin. De nombreux systèmes de gestion documentaire ont alors vu le jour en ce sens.

La seconde génération de KM – KM 2.0 en quelque sorte – bien que s'appuyant toujours sur les bases de connaissances de la génération précédente, tente d'appréhender dans le même temps le côté social et informel de la connaissance. Les technologies employées vont de ce fait se tourner davantage vers la collaboration et les mises en relations d'employés, de clients, car c'est dans ces mises en relation, dans ces échanges que leur expertise va pouvoir s'approfondir, et de nouvelles connaissances émerger pour l'entreprise. La question se pose donc de la capitalisation de ces nouvelles formes de connaissances, que nous essayerons ici d'aborder au travers des outils de réseautage social. Nous tenterons alors de définir dans les grandes lignes ce qu'un tel support pourrait apporter à l'entreprise, avant de présenter des outils actuels approchant notre problématique, et nous conclurons sur la piste de solution envisagée pour répondre plus finement à cette problématique.

Introduction

Les entreprises ont de tout temps eu besoin de stocker, de classer, d'organiser leurs connaissances internes. Pour ce faire elles ont rapidement adopté différents outils. La mission première

de ces outils était l'archivage. Les données étaient ainsi stockées de manière sécurisée face au temps et aux intrusions.

Malheureusement, cette méthode d'archivage, qu'elle soit physique (bibliothèques, entrepôts) ou numérique, ne supportait un accès rapide aux données que suite à une phase de détermination, de classification. Pour des besoins de robustesse cette classification restait généralement très statique dans l'entreprise, n'autorisant que rarement l'ajout de nouvelles catégories.

Parallèlement, la montée en puissance du web s'est faite en deux parties. Dans un premier temps, de ses prémices jusqu'au début des années 2000, internet est resté un média classique, un outil de lecture de plus. Les utilisateurs du réseau n'avaient qu'un rôle passif vis-à-vis du contenu.

Au début du siècle, de nouvelles formes de sites internet ont vu le jour. Allant des wikis aux blogs, ils ont permis aux internautes de devenir acteurs du web, d'avoir un rôle actif sur le contenu. Ces nouvelles fonctionnalités sont à l'origine de la grande démocratisation du web, dès lors plus cantonné aux seuls techniciens.

De nouvelles formes de productions sont apparues sur le web, comme les contenus multimédia. La participation des visiteurs ont amené la création de communautés d'intérêts (Wenger, 1998). Ces communautés ont utilisé tous les moyens à leur disposition pour exister : d'abord les newsgroups, puis les forums et les blogs, et plus dernièrement les réseaux sociaux.

Devant l'engouement provoqué par les nouveaux outils dit « web 2.0 » et les réseaux sociaux en particulier, les entreprises se sont intéressées au phénomène, et ont tenté d'adapter à leurs besoins ces nouvelles technologies et usages (Koch et Richter, 2007) et (McAfee, 2009). L'idée principale de cette adoption est de favoriser les échanges entre ses collaborateurs, afin d'être source d'innovation (McAfee, 2006).

Ces échanges représentent un nouveau vecteur source de connaissances au sein de l'entreprise. Or, celle-ci est très rarement armée pour une bonne capitalisation de ces derniers, comme nous le verrons tout d'abord. Nous essayerons alors de définir dans les grandes lignes ce qu'un tel support pourrait apporter à l'entreprise, avant de présenter des outils actuels approchant notre problématique, et nous conclurons sur la piste de solution envisagée pour répondre plus finement à cette problématique.

1 Problématique

L'entreprise innovante a un souci constant de collaboration de la part de ses employés. Un projet ne se fait pas seul, et pour le mener à bien ses membres vont devoir créer des documents, en consulter d'autres ou s'en échanger.

Cette collaboration va être source de productions et d'échanges pouvant prendre de multiples formes comme des articles, des plans, des commentaires sur des photographies récentes, des dialogues entre membres, mais aussi des mises-à-jour de profil.

Ces échanges vont bien souvent avoir lieu au sein d'outils de travail collaboratif mis en place par l'entreprise. Or, au sein de ces outils, les fragments intéressants vont se retrouver noyés au milieu d'autres informations, qui, selon le contexte et le besoin, ne seront pas pertinents vis-à-vis des connaissances portées par ces fragments. Nous constatons d'ailleurs que cette notion de pertinence va elle-même varier dans le temps. Parfois même, ces connaissances seront éparpillées dans différents outils non reliés entre eux.

La capitalisation de ces fragments va donc dépendre de leur forme et probablement de leur contexte. Nous pensons pouvoir partager ces nouvelles formes d'expression de connaissances en deux parties :

- les fragments dont la forme est similaire aux anciens documents, que nous nommerons la forme classique ;
- les fragments issus des réseaux sociaux et des traces laissées par les utilisateurs, que nous nommerons les fragment sociaux.

Les fragments classiques

Les fragments classiques représentent à l'heure actuelle la majeure partie de la production de l'entreprise. Qu'il s'agisse d'articles publiés au sein d'une ferme de blogs d'intranet, ou des pages de manuels publiées au sein d'un wiki, ces documents sont encore très similaires aux anciennes productions de l'entreprise.

Elles peuvent aisément s'introduire dans les systèmes de capitalisation de connaissances actuels. Un article de blog par exemple va se caractériser par un titre, un contenu, des mots-clés qui seront facilement indexés dans un système de gestion de connaissances. Les utilisateurs peuvent rapidement retrouver du contenu par de simples requêtes dans le système.

Ce type de contenu a tout de même évolué ces dernières années pour se tourner vers d'autres formats. Il n'est pas rare aujourd'hui de publier sur des intranets d'entreprise, ou au sein de laboratoires des vidéos de démonstrations ou d'interviews, ou même du contenu interactif, via des appliquettes java ou flash. Bien que plus difficilement exploitables via les outils d'indexation classiques, ces formats restent caractérisables par le biais de meta-données qui permettront aux utilisateurs de retrouver ces fragments dans le système.

Les fragments sociaux

Les outils web 2.0 ont apporté de nouveaux usages sur le web. Le commentaire d'un article en est un exemple. Auparavant des critiques pouvaient êtres produites, mais chaque critique, réponse etc. était assimilable à un document unique. Les commentaires sont pour leur part directement rattachés aux articles qu'ils annotent, comme autant de post-it auparavant appliqués sur les épreuves papier.

Les réseaux sociaux en ligne ont également créé de nouveaux modes de communication. Par l'introduction des messages de statut, devant décrire originellement un état d'esprit ou une localisation, les réseaux sociaux ont permis la récente explosion autour du microblogging, portée par une génération d'utilisateurs adeptes des SMS [1].

Un dernier vecteur de connaissances pour l'entreprise va se trouver dans les relations qui peuvent se créer au sein des réseaux sociaux. Si deux collaborateurs se relient, quelle importance donner à cet évènement : cela peut être synonyme d'une collaboration fructueuse, d'une amitié, d'une relation commerciale, etc.

Le problème de la capitalisation de ces fragments sociaux apparaît. De natures très diverses, se prêtant peu ou pas aux méthodes d'indexation classiques, leur intégration dans les bases de

1. Le nombre de SMS échangés en France s'élève à plus de 19,3 milliards au dernier trimestre 2009 d'après les résultats définitifs publiés par l'Arcep. Cela représente une hausse de 67% par rapport au dernier trimestre 2008 où 11,6 milliards de SMS avaient été échangés. Ce nombre est en hause constante depuis l'introduction de ce mode de communication.

connaissances habituelles soulève de nombreuses interrogations. De plus, ces fragments sont au cœur de la problématique de leurs contextes de production. Un commentaire seul sorti de son contexte va perdre toute son essence. Il est donc essentiel de se poser la question du support à utiliser pour mettre en œuvre la capitalisation de ces fragments sociaux, et pour cela l'étude de quelques outils actuels va pouvoir guider notre réflexion.

2 Définition d'un support spécifique

Dès lors que nous avons mis en évidence la problématique de capitalisation de ces nouveaux supports de connaissances, le besoin de redéfinir un support de capitalisation se fait sentir. La capitalisation va permettre de transformer les connaissances tacites de l'entreprise en connaissances explicites, exploitables par tous les collaborateurs (Nonaka et Takeuchi, 1995).

Les fragments sociaux sont assimilables à des connaissances tacites de l'entreprise. De fait, ces fragments échapaient jusqu'alors aux méthodes d'indexation habituelles d'informations, et se plaçaient même en dehors des systèmes d'informations. Or, nous constatons que ces fragments sociaux vont de plus en plus être produits au sein d'outils rédactionnels divers – commentaires sur un blog, messages dans un forum, chat, etc. – conduisant donc à la production d'une trace écrite.

Nous pensons donc que c'est par le biais de cette production d'informations, que ces connaissances tacites vont en quelque sorte s'expliciter au sein des systèmes d'informations, pour peu que nous prenions la peine d'indexer ces productions (figure 1). Nous pouvons d'ailleurs constater que nous touchons ici du doigt un des principes de l'entreprise 2.0, qui est la remise à plat des hiérarchies au sein d'un réseau, en plaçant sur un pied d'égalité les fragments classiques et sociaux vis-à-vis de leur indexation.

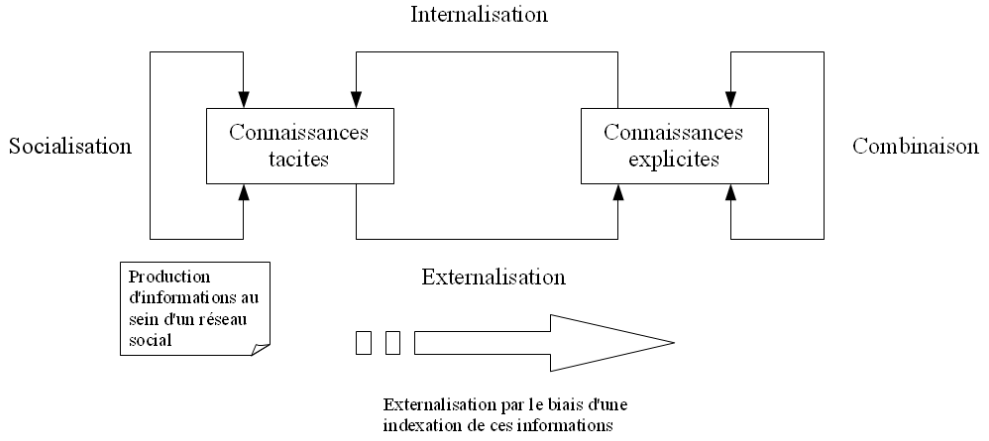


FIG. 1 – *Le cercle vertueux de la connaissance (Nonaka et Takeuchi, 1995)*

Comme toute connaissance tacite, ces fragments, nous l'avons vu, dépendent du contexte dans lequel ils ont été produits. Il paraît donc logique de les considérer dans leurs contextes,

de ne pas les y en extraire. Le système source, au sein duquel ces fragments ont été produits, est donc à considérer comme un support de connaissances.

Si nous prenons l'exemple d'un réseau social d'entreprise, les ressources fournies par les utilisateurs ne peuvent pas être extraites du réseau et insérées de manière brute dans un système de capitalisation de connaissances : sous cette forme, elles n'apporteraient rien à l'entreprise, et elles seraient perdues. Leur apport va se situer dans les liens que ces ressources vont tisser entre elles au sein du réseau. Un commentaire est indissociable de l'article qu'il commente, considéré comme son contexte, tandis que deux notes de service peuvent être considérées indépendamment l'une de l'autre.

Pendant longtemps les informations de l'entreprise, supports de la connaissance, ont pu être rassemblées dans des entrepôts de données, triées et rangées, car leur forme leur permettait d'être contextuellement indépendantes deux à deux. Un système de référence était mis en place pour relier deux documents différents entre eux, mais ce système de référence était rendu possible par la nature artificielle du classement des documents. Le contexte originel de la production était effacé au profit d'une hiérarchie de catégorie conventionnée, de même que ce contexte n'apparaissait pas dans les liens potentiels entre informations, même si des classements multidimensionnels permettaient une liaison.

L'apparition des réseaux sociaux et la création de liens contextuellement forts entre les fragments de connaissances issus des productions des collaborateurs au cours de leurs travaux conduit à repenser cette vision artificiellement hiérarchisée des connaissances, et à la conception d'un support innovant permettant de dresser des ponts entre les anciennes méthodes d'indexation et de stockage des informations, et les nouveaux usages de production de savoir de l'entreprise.

Lorsque les utilisateurs vont ajouter des liens entre leur profil social et des documents issus de la base de connaissances historique de l'entreprise, ils vont enrichir le patrimoine de connaissances de l'entreprise, sans pour autant ajouter ces nouvelles informations dans la base originale. Un bon support semble donc être un outil permettant aux collaborateurs de se définir socialement les uns par rapport aux autres, et leur permettant d'échanger toutes sortes d'informations, issues des bases de connaissances de l'entreprise.

Il s'agit donc en quelque sorte de redéfinir la sémantique permettant de relier deux fragments de connaissances entre eux. Les simples relations d'identité – « est un » article, « appartient à » la catégorie X, etc. – ne semblent plus suffisantes : de même que les relations entre individus sont complexes – relations de groupe comme la famille, l'amitié, le travail, mais aussi croisée, comme les relations associatives et de travail etc. – le fragment de connaissance va de plus en plus avoir besoin de liens complexes pour se définir. Une cartographie sociale des connaissances va se dessiner.

Tout ceci conduit à l'isolation d'un besoin qui semble fondamental pour ces organisations – entreprises ou laboratoires –, à savoir la capitalisation de nouvelles formes d'information, liées aux nouveaux usages drainés par le web 2.0 jusque dans l'escarcelle de l'organisation. Ce besoin, auquel nous allons essayer de répondre en proposant une piste de solution par la suite, entraîne un raffinement de notre problématique de départ.

Si le besoin primordial de ces organisations est la capitalisation des fragments sociaux, il ne faut pas pour autant négliger les fragments classiques. En effet l'existence parallèle de deux systèmes de gestion de connaissances, un pour chaque type de fragment, est abscons. Si l'utilisateur, lors d'une fouille de données doit jongler entre différentes bases de connaissances,

différents systèmes de gestion de connaissances, cela va représenter un effort cognitif important qui conduira certainement à l'abandon de l'un ou de l'autre, voire des deux, de ces supports d'informations, pourtant nécessaires à la sauvegarde des connaissances de l'entreprise.

Ce besoin ne va pouvoir se combler que par la réalisation d'un nouveau type de support d'informations, acceptant des ressources très hétérogènes (Boughzala, 2008). Nous pouvons constater que ce support se démarque de ce qui se fait actuellement, et va donc demander une réflexion sur sa mise en œuvre.

3 Quelques projets pouvant répondre à ce besoin

La mise en œuvre d'un nouveau support de capitalisation de la connaissance, comme nous venons de le suggérer précédemment, dans des structures comme les laboratoires ou les entreprises n'est pas forcément synonyme d'efforts inconsidérés. En effet ces nouveaux supports se placent dans le prolongement de ce que le web 2.0 a pu être pour le web 1.0, c'est-à-dire plus une évolution qu'une révolution (Mielnik et Félix, 2008). Ainsi, ces supports vont pouvoir s'appuyer sur nombres de projets, de systèmes actuels qui pris séparément n'y répondent pas complètement, mais dont la présentation peut être source d'inspiration à différents niveaux de la définition de ce nouveau support.

3.1 eMEMORAe 2.0

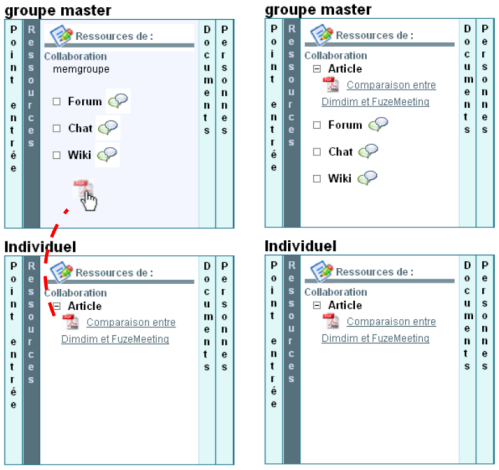
EMEMORAe 2.0 est un projet de recherche mené au sein de l'Université de Technologie de Compiègne (UTC) et conduit par Marie-Hélène Abel. Il s'agit à la base d'un projet permettant la conception d'une mémoire organisationnelle pour le e-learning (Abel et al., 2004).

Le principe de eMEMORAe 2.0 est de permettre à l'enseignant de définir un certain nombre de concepts liés à son cours. Chaque concept est ensuite placé précisément au sein d'une ontologie. La navigation au sein du cours peut donc se faire comme au sein d'une cartographie, en avançant de concept en concepts liés, et non plus linéairement. Chaque concept peut indexer des ressources, fragments divers d'informations, supports de connaissances, allant du cours à des forums, agendas sémantiques aux pages de wiki.

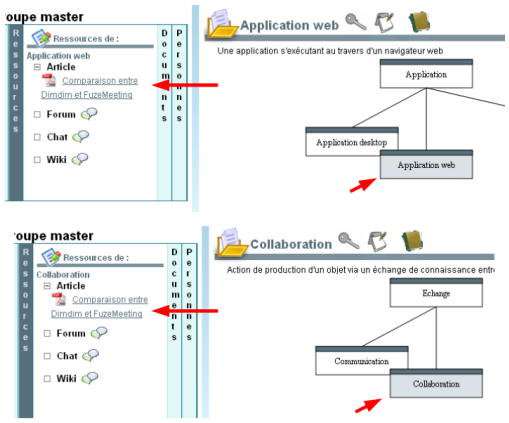
Chaque utilisateur du système va pouvoir naviguer dans l'ontologie, en piochant parmi les concepts ceux pouvant l'intéresser, afin de les compulser dans une vue personnelle. Une autre vue partagée avec son groupe de travail lui permettra de partager des concepts comme autant de points d'entrée dans l'ontologie, mais aussi d'autres types de ressources avec ses camarades (figure 2 (a) page 7).

Dans l'optique d'un outil permettant la collaboration d'entités – entreprises ou laboratoires – différentes, ce principe de vues partagées est très intéressant car il autorise le partage d'informations entre entité, tout en ménageant la possibilité à ces entités de conserver un accès exclusif à certaines ressources. Ainsi, pour un acteur à un instant donné, il ne pourra voir qu'une partie des ressources stockées dans le système, en fonction du concept particulier qu'il observe, de ses droits d'accès à la base de connaissances et de son statut au sein de l'entité. Cette approche conceptuelle autorisant de multiples points de vue sur les ressources stockées au sein de ce système, se rapproche énormément du besoin de contextualisation ressenti dans la capitalisation des fragments sociaux.

Toute l'originalité de eMEMORAe 2.0 tient dans la méthode d'accès aux ressources proposées. Cet accès n'est pas statique et lié à un classement unique. Il ne se fait pas en parcourant un arbre figé dont les ramifications ne se croisent pas. Au contraire, les ressources sont placées au sein d'un réseau constitué par les concepts de l'ontologie décrivant l'entité ou ses activités. Chaque ressource pouvant venir illustrer différents concepts, leurs chemins d'accès au sein de ce graphe conceptuel peuvent emprunter des voies très diverses (figure 2 (b)).



(a) Partage de ressources dans eMEMORAe 2.0



(b) Les ressources peuvent être découvertes par différents chemins

FIG. 2 – *Aperçus d'écrans de l'application eMEMORAe 2.0*

Le manque ressenti à l'utilisation de eMEMORAe 2.0 dans son état actuel provient de la faible intégration du concept d'utilisateur au sein du système. Il manque en effet la possibilité de pouvoir naviguer au sein des utilisateurs comme au sein du corpus, en fonction par exemple de leurs compétences ou de leur profil social. Cette possibilité permet de considérer les acteurs de l'organisation comme porteurs d'une forme de fragment de connaissance sociale, et les re-

présenter de la même manière que les autres ressources autorisent de plus grandes interactions entre collaborateurs lors de phases de recherche d'informations.

De même, la notion de groupe est actuellement sous exploitée par eMEMORAe 2.0. Comme nous l'avons relevé, l'intérêt de la constitution de ces groupes va porter sur la création conjointe de vues sur la base de connaissances, matérialisées par l'affichage plus ou moins restreint de ressources par concept de l'ontologie. Or, actuellement ces groupes sont fondés de manière autoritaire par le responsable du cours au sein du système, et aucun mécanisme n'autorise la création spontanée de groupe au sein des utilisateurs. Il n'est pas possible pour ces utilisateurs de se regrouper par eux-mêmes autour d'une compétence, d'un projet, ou toute autre chose. La création spontanée de communautés d'intérêt n'est pas possible.

3.2 SAP Social Network Analyser et Microsoft Pivot

Social Network Analyser (SNA) est une expérimentation menée par SAP [2] autour des réseaux sociaux d'entreprise. Le but de ce démonstrateur est d'illustrer les usages prometteurs des nouvelles technologies d'interfaces riches dans la manière de manipuler des données, et en l'occurence un annuaire d'entreprise.

Cet exemple de réseau social est très intéressant dans la façon dont il met en scène la navigation au sein des équipes de l'entreprise. Le passage d'une équipe à l'autre, d'une thématique à l'autre se fait très simplement, en cliquant sur différents mots-clés mis en évidence dans les profils utilisateurs, conduisant à une recomposition de l'affichage. Des statistiques sur l'unité visible s'affichent également en temps réel, offrant par ce biais des chemins d'entrées différents dans le graphe social.

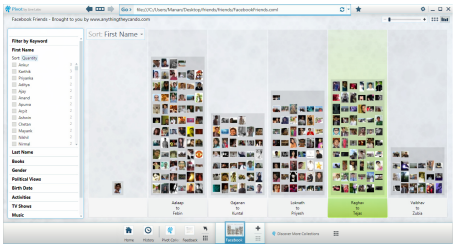
SNA n'offre aucune liaison avec de quelconques ressources de l'entreprise : il se concentre sur les relations sociales en se basant exclusivement sur les « relations de voisinage » en équipe. En effet, SNA seul ne sert à rien en l'état, vu qu'il ne s'agit que d'une démonstration sans aucune vocation à être commercialisée, et qu'il ne gère, de fait, absolument pas la dimension documentaire d'une structure. Mais la manière dont il présente les collaborateurs composant cette structure semble pouvoir convenir dans le cadre d'une remontée d'informations neutre du point de vue de l'objet considéré – utilisateur ou document.

Cette navigation naturelle au sein des collaborateurs est rendue possible grâce aux avancées réalisées dans le domaine des interfaces utilisateurs et plus précisément les travaux réalisés sur les interfaces dites riches. SNA utilise la dernière version du framework de développement Flex d'Adobe [3], mais d'autres innovations sont apparues ces dernières années pour améliorer la manipulation visuelle de grandes quantités de données, comme le concept Pivot développé par Microsoft.

Pivot [4] est un projet de recherche initié par Microsoft dans le cadre de ses Live Labs [5]. Il s'agissait pour eux de tester un nouveau mode de représentation des données issues du web, en permettant à l'utilisateur de naviguer au sein de ces données, et non plus simplement de les faire défiler séquentiellement, comme le permet la navigation hypertexte.

2. http://timoelliott.com/blog/2010/02/sap-businessobjects-social-intelligence-prototype-v2-launches.html
3. http ://www.adobe.com/fr/products/flex/
4. http://www.microsoft.com/silverlight/pivotviewer/
5. http://livelabs.com/

(a) Aperçu d'écran de l'application Social Network Analyser de SAP



(b) Aperçu d'écran de l'application Pivot

Cette expérimentation est elle-même basée sur une autre expérience de Microsoft, appelée Seadragon [6], étudiant les interactions possibles de l'utilisateur sur des données grâce à un outil de zoom. L'idée étant d'utiliser le zoom et le dé-zoom comme outil de raffinement et de navigation au sein d'un cluster de données, un peu à la manière de Prézi [7], qui réutilise ce concept pour innover dans le domaine des présentations.

Ces travaux réalisés sur la présentation des données autorisent désormais à penser à des applications de capitalisation de connaissances vraiment agréables à utiliser, non plus basées sur de simples pages de formulaires à remplir ou consulter, telles les classiques fiches de bibliothèques, mais bien la construction et l'alimentation de graphes interconnectés permettant de se promener au sein des connaissances amassées par une organisation d'une manière très naturelle, passant d'un domaine à l'autre, d'une équipe à l'autre souplement, sans avoir à sans cesse recommencer les mêmes manipulations.

3.3 Diaspora

Diaspora, au même titre que Movim ou Gnu Social [8], est une initiative de développement de réseau social distribué. Ils sont partis d'un constat montrant que les principaux sites sociaux d'aujourd'hui – Flickr, Facebook, Twitter, etc. – n'offrent pas de garanties suffisantes au regard du respect de la vie privée, du fait de leur organisation centralisée.

La meilleure façon de conserver sa vie privée étant de rester chez soi, de nombreuses initiatives ont vu le jour afin de permettre aux internautes d'utiliser des instances de sites sociaux chez eux. Seul l'internaute décide de se connecter à d'autres nœuds du réseau et il garde ainsi la main sur la totalité des données qu'il partage.

Les avancées dans ce domaine sont encourageantes, car une bonne normalisation de ce type de logiciel permettrait à chacune des entités de déployer sa propre instance de réseau social en interne, sans crainte de fuites éventuelles, et pourrait ponctuellement se connecter aux ressources provenant d'une entité tierce : il s'agit d'une implémentation évoluée du principe de mémoire partagée, telle que nous l'avons vu avec eMEMORAe 2.0.

6. http://www.seadragon.com/
7. http://prezi.com/
8. Respectivement https://joindiaspora.com/, http://www.movim.eu/ et http://social.foocorp.net/software/social/

4 Piste de solution

Nom	Type d'outil considéré	Forme des fragments	Paradigme de navigation utilisé	Utilisation de liens entre utilisateurs
	Annuaire d'entreprise		Hypertexte Accès directs	
SAP Network Analyser	Annuaire d'entreprise	Sociale	Cartographique. Nombreux chemins possibles	Forte. Les liens entre utilisateurs forment la colonne vertébrale du système.
	Réseau social d'entreprise	Sociale	Hypertexte et cartographique. Nombreux points d'entrées et chemins possibles	Forte. Les liens entre utilisateurs forment la colonne vertébrale du système.
	Base de connaissances	Classique	Hypertexte Accès directs	
Microsoft Pivot	Navigateur dans une base de connaissances	Classique et sociale	Graphique. Nombreux points d'entrées et chemins possibles	Tout objet peut être considéré indépendamment de sa signification dans l'ontologie de référence.
eMEMORAe 2.0	Base de connaissances	Classique et sociale (via l'indexation des forums ou des pages de wiki par concept)	Hypertexte et cartographique. Nombreux points d'entrées et chemins possibles	Faible. Bien que chaque ressource soit reliée à un utilisateur, il n'existe pas encore de mode de visualisation par utilisateur.

FIG. 3 – *Résumé des possibilités offertes par différents outils actuels*

Nous l'avons vu, aucun de ces produits seul n'apporte une solution complète au problème que nous nous sommes posés. Néanmoins, chacun apporte, par son lot d'innovations propres, une facette du prisme pouvant composer une réponse adéquate.

Nous pensons que eMEMORAe 2.0 est la solution permettant une approche juste du problème. Basée autour d'une ontologie centrale, eMEMORAe 2.0 va permettre de lier les informations de l'entreprise à différents concepts. Ces informations pourront se transformer en connaissances lorsque l'utilisateur va se les approprier, lors de sa navigation au sein de l'ontologie. Une même ressource pouvant être accédée via différents chemins au sein de la cartographie, le système permet pour un même fragment une multitude de points d'entrée possibles dans le système.

Sa conception très modulaire autorise dès à présent d'utiliser un grand nombre de types de ressources, englobant les notions de fragments classiques et sociaux : documents externes – PDF, traitement de texte, etc. –, messages de forum, chat, pages de wiki, etc. L'ajout de nouveaux types de ressources ne devrait pas poser de problèmes techniques. Inclure dans ce

fonctionnement une meilleure représentation des utilisateurs va permettre d'imaginer de nouveaux modes de recherche d'informations.

Les systèmes actuels de gestion de la connaissance ne sont conçus que pour remonter lors d'une requête des documents liés aux critères de la recherche. Rien n'existe à la base pour permettre de rechercher dans le même temps des individus. Or, dans la conception d'une réponse d'appel d'offre, le fait d'identifier les personnes clés de l'entreprise sur la problématique liée est au moins aussi important que le fait de retrouver les archives correspondantes à de précédentes réponses.

Cette recherche d'individus pourra s'effectuer à la fois par le biais des concepts les décrivant que les utilisateurs auront sélectionnés, mais aussi via les concepts liés aux productions qu'ils auront réalisées lors de travaux collaboratifs au sein du système – annotations d'un document, réponse sur un forum, rédaction d'une notice sur un wiki ou d'un article sur un blog – sans même s'en rendre compte.

Lorsqu'un utilisateur sera dans une phase de rassemblement d'informations, il pourra retrouver différents fragments, qu'ils soient classiques – documents ajoutés à la main dans le système – ou sociaux – directement produit sur eMEMORAe 2.0, par l'usage qu'en auront eu les autres utilisateurs ou lui-même.

L'autre force de eMEMORAe 2.0, nous l'avons vu, est de permettre nativement de répartir ses connaissances entre différentes mémoires, partagées ou non. Dans le cadre d'une collaboration entre diverses entités, il s'agit d'un point très important car cela permettra d'encore multiplier les points d'entrées possibles vers la connaissance, à savoir si celle-ci est issue de la propre structure de l'utilisateur, ou d'une organisation partenaire. La base de connaissances se trouve ainsi centralisée, ce qui simplifie sa mise en place, tout en garantissant la confidentialité potentielle des connaissances des partenaires.

Conclusion

Nous avons vu qu'il pouvait être intéressant de classifier les supports de connaissances en deux formes bien distinctes, les fragments classiques et sociaux. Cette séparation impose du coup de devoir repenser nos modes d'accès aux connaissances que peut capitaliser une organisation. Cette remise en question permet d'imaginer de nouveaux systèmes de capitalisation des connaissances prenant en compte les connaissances sociales de l'entreprise, qui auront pu être produites au sein de ce système de manière transparente pour l'utilisateur.

Cette tâche demandera au préalable d'avoir définitivement défini ce que sont les fragments classiques et sociaux, et les caractéristiques des échanges entre individus à l'origine des fragments sociaux, sur une plateforme telle que eMEMORAe 2.0. En effet cette définition plus précise sera la source des modèles permettant de développer de nouveaux systèmes de capitalisation de connaissances innovant, prenant à la fois en compte ces deux formes de fragments ou permettant directement leur production.

Nous pensons que eMEMORAe 2.0 peut être utilisé pour initier un prototype permettant de mesurer l'utilisation d'un tel système de capitalisation de la connaissance. Nous avons en effet mis en évidence le fait qu'il ne lui manquait que peu de chose en l'état actuel de nos travaux pour qu'il joue ce rôle de démonstrateur. Son principal défaut étant qu'il ne met pas suffisament en relief ses capacités sociales, qui pourraient être améliorées. Dans le cadre de ma thèse, je vais donc être amené à ajouter à eMEMORAe 2.0 les moyens d'appréhender complètement

les fragments sociaux, et en particulier renforcer le rôle de la notion d'utilisateur sur cette plateforme.

Références

Abel, M.-H., A. Benayache, D. Lenne, C. Moulin, C. Barry, et B. Chaput (2004). Ontology-based organizational memory for e-learning. *Educational Technology & Society volume 7*(4), 98 – 111.

Boughzala, I. (2008). Ingénierie de la collaboration pour le km. In A. Dudezert et I. Boughzala (Eds.), *Vers le KM 2.0. Quel management des connaissances imaginer pour faire face aux défis futurs*, Chapter 1.1, pp. 21 – 33. Paris : Vuibert.

Koch, M. et A. Richter (2007). *Enterprise 2.0 : Planung, Einführung und erfolgreicher Einsatz von Social Software in Unternehmen* (1 ed.). München : Oldenbourg.

McAfee, A. P. (2006). Enterprise 2.0 : The dawn of emergent collaboration. *MIT Sloan Management Review 47*(3), 21 – 28.

McAfee, A. P. (2009). *Enterprise 2.0 : new collaborative tools for your organization's toughest challenges*. Harvard Business School Press. Harvard Business Press.

Mielnik, J.-C. et E. Félix (2008). Quel partage des connaissances en entreprise à l'heure du web 2.0 et de l'intelligence collective ? In A. Dudezert et I. Boughzala (Eds.), *Vers le KM 2.0. Quel management des connaissances imaginer pour faire face aux défis futurs*, Chapter 1.2, pp. 35 – 54. Paris : Vuibert.

Nonaka, I. et H. Takeuchi (1995). *The knowledge-creating company : how Japanese companies create the dynamics of innovation*. New York : Oxford University Press.

Wenger, E. (1998). *Communitites of Practice : Learning, Meaning and Identity*. Boston : Cambridge University Press.

Summary

Expectations and practices around Knowledge Management (KM) has evolved during the last two decades, allowing us to introduce the idea of KM generations. The first generation worked principaly on knowledge or good practices bases for enterprise. The sensitive datas were captured, identified and stored into these systems in order to be easily retrieved by collaborators who need them. Many documents management systems are born because of that.

The second generation of KM – which we could call KM 2.0 – despite its filiation with the knowledge bases from the first generation, try to capture in the same time the social and informal aspect of knowledge. The used technologies push the collaboration and links between coworkers or customers because their expertise will grow by these links, and new knowledge will be created for the enterprise. The question is about the capitalization of these new kind of knowledge, which we purpose to solve by using social networking tools. We will try to show what could bring such a tool to enterprises. Then we will present some current tool which can be useful for our problematic. Finally, we will conclude with a potential track to answer the question.

Étude d'une ontologie socioculturelle

Papa Fary Diallo, Seydina Moussa Ndiaye
Moussa Lô

Laboratoire d'Analyse Numérique et d'Informatique (LANI),
UFR SAT – Université Gaston Berger,
Saint-Louis – BP 234, Sénégal.
perfary@hotmail.com
seydina.ndiaye@ugb.edu.sn
moussa.lo@ugb.edu.sn

Résumé. Dans ce papier nous proposons un processus de développement d'ontologie socioculturelle dans le but de vulgariser et de pérenniser la culture d'un pays à travers un partage des coutumes et l'histoire des différentes localités du pays. Ce processus peut être assimilé à la construction d'une plateforme qui serait à cheval entre une « mémoire d'entreprise » et un « réseau social », mais appliquée au contexte d'un pays. Ce processus s'articule autour de la théorie du psychologue russe Lev Vygotsky appelée «Vygotskian Framework».

1 Introduction

La méconnaissance des territoires africains s'aggrave de jour en jour avec l'expansion de la télévision (mondialisation des contenus avec les chaînes câblées de plus en plus nombreuses) et l'émergence de l'Internet dans les pays du Sud.

Même les productions réalisées par les chaînes de télévision locales ont tendance, de plus en plus, à s'intéresser à l'Occident au détriment de la découverte de leurs espaces nationaux.

Ainsi pour rafraîchir la mémoire de nos concitoyens et redonner vie aux nombreux récits qui accompagnent la création et la vie au quotidien des différents terroirs africains, nous avons initié la mise en place d'une encyclopédie socioculturelle en ligne. Cela répond à un besoin de disposer d'un cadre de partage de connaissances sur les communautés sénégalaises.

Notre objectif est de développer une infrastructure distribuée qui permettra aux communautés sénégalaises de partager leurs connaissances socioculturelles, touristiques, économiques, scolaires, agricoles, etc. L'infrastructure que nous souhaitons développer peut être assimilée à une plateforme qui serait à cheval entre une « mémoire d'entreprise » (ou « mémoire d'organisation ») et un « réseau social », mais appliquée au contexte d'un pays. Techniquement, il s'agit de réaliser une plate-forme web sémantique qui permettra à des communautés de partager leurs connaissances socioculturelles, économiques, etc.

Nous proposons d'établir un nouveau point de vue de la notion de communauté dans le cadre du Web social où classiquement la communauté représente un ensemble d'individus partageant de mêmes aspirations. Notre approche est moins portée sur les individus (qui sont classiquement les points centraux) que sur les croyances et connaissances que ceux-ci partagent. Cette translation de point de vue nous permet d'aborder une communauté précise comme une entité atomique et de s'intéresser cette fois-ci au partage de connaissances entre communautés.

En représentant sémantiquement les ressources manipulées dans notre « réseau social », nous disposons d'une couche sémantique qui permet l'accès aux différentes informations contenues dans le réseau. Par ailleurs, le Web sémantique ouvre la voie à une approche sémantique de l'analyse des réseaux sociaux, ce qui permet également d'en extraire de nouvelles connaissances.

La représentation sémantique s'appuie sur une ontologie socioculturelle dont le développement constitue l'objectif de cet article.

Le terme "ontologie" a été emprunté à la philosophie où il désigne l'étude de l'être en tant qu'être, sur l'être en soi. Mais en ingénierie des connaissances la définition la plus populaire reste sans doute celle de (Gruber, 1993) : "une ontologie est une spécification explicite d'une conceptualisation". Une ontologie dans le contexte de l'ingénierie de connaissances selon (Uschold, 1998) "peut prendre différentes formes, mais elle inclura nécessairement un vocabulaire de termes et une spécification de leur signification. Cette dernière inclut des définitions et une indication de la façon dont les concepts sont reliés entre eux, les liens imposant collectivement une structure sur le domaine et contraignant les interprétations possibles des termes". Une telle caractérisation rend compte d'objets divers tels des glossaires, des terminologies, des thesaurus et des ontologies (au sens strict), mis en œuvre par différents professionnels (ingénieurs de la connaissance, bibliothécaires, traducteurs) et se distinguant suivant que l'accent est mis sur les termes ou leur signification.

La construction de l'ontologie socioculturelle s'articule autour du processus «Vygotskian Framework » proposé par le psychologue Russe Lev Vygotsky (Ivic, 1994). Ce processus examine la relation entre le savoir et le développement d'une société suivant trois axes : les humains (subject), les objets (bâtiment, parc, etc.), et les artefacts (outils mentaux, outils matériels, images, dessins visuels, etc.).

Ce papier continue par un état de l'art dans lequel nous mentionnons les travaux du web social et du web sémantique qui nous ont guidés dans notre démarche. La troisième section présentera l'ontologie socioculturelle en faisant une présentation de la théorie de Vygotsky, des concepts et relations qui composent notre ontologie et nous proposons une approche d'analyse du réseau social. Nous terminerons par une conclusion et des perspectives pour ce travail.

2 Web sémantique, Web social, Web socio sémantique

Nous nous plaçons dans le cadre de la mise en place d'une plate-forme web sémantique basée sur une ontologie socioculturelle en vue de permettre à des communautés de partager leurs connaissances comme le feraient des individus au sein d'un réseau social.

Nous nous intéressons ici au processus de développement de l'ontologie socioculturelle. Il n'existe pas, à notre connaissance, une telle ontologie. Dans cette section, nous allons présenter les travaux effectués dans le domaine du Web social et du Web sémantique qui ont guidé nos choix méthodologiques.

2.1 Le Web social

Il existe plusieurs définitions du Web social. Mais pour notre étude nous considérons la définition de (Gruber, 2007), qui définit le Web social comme un écosystème de la participation, où la valeur est créée par le regroupement de plusieurs contributions des utilisateurs individuels.

Dans notre cas, les contributions que peuvent introduire les utilisateurs seront certainement les nouvelles structures qui sont nouvellement installées dans une localité (création d'une nouvelle instance) et les relations qu'elles peuvent avoir avec celles existant. Il y'a également les événements socioculturels qui vont s'y produire.

Une fois qu'un réseau social est constitué, il peut être analysé en faisant l'étude des entités sociales ainsi que leurs interactions et leurs relations. Ceci est appelé analyse des réseaux sociaux. Une telle analyse se rapporte à la théorie des réseaux sociaux qui conçoit les relations sociales en termes de nœuds et de liens. Les nœuds sont habituellement les acteurs sociaux dans le réseau mais ils peuvent aussi représenter des institutions, et les liens sont les relations entre ces nœuds, cette représentation est appelée sociogramme et elle a été proposée dans (Moreno, 1933).

Parmi les indicateurs d'un réseau social nous pouvons citer la densité et la centralité.

La densité indique le nombre de liens au sein d'un réseau et permet de définir la cohésion d'un réseau social. Selon (Scott, 2000) cette mesure peut être utilisée dans l'optique d'une analyse socio-centrée ou égo-centrée.

La centralité met en lumière les acteurs les plus importants du réseau. (Freeman, 1979) propose trois définitions de la centralité : (i) la centralité de degré considère comme centraux les nœuds qui possèdent les degrés les plus élevés du graphe c'est-à-dire ceux qui ont plus de liens dans le réseau ; (ii) la centralité de proximité indique le degré auquel un nœud est près de tous les autres nœuds d'un réseau social (directement ou pas), elle reflète la possibilité d'accéder à l'information à la source dans le réseau social ; (iii) la centralité d'intermédiarité se concentre sur la capacité d'un nœud à servir d'intermédiaire dans un graphe. Toutefois, Freeman ne considère que les graphes non orientés. Or dans un réseau social, l'orientation des relations contient à elle seule beaucoup de sémantique. La prise en compte de la direction des relations nous amène à la notion de prestige qui constitue une mesure plus raffinée que la centralité. Nous distinguons les liens sortants et ceux entrants. Un acteur prestigieux est un

acteur ayant beaucoup de liens entrants. Le prestige d'un acteur est mesuré par le nombre des liens entrants.

L'émergence du Web sémantique amène à appliquer les méthodes d'analyse des réseaux sur de nouvelles traces générées par les usages du Web.

2.2 Le Web sémantique

Tim Berners-Lee (Berners-Lee et al., 2001) décrit le Web de demain comme un vaste espace d'échange de ressources entre êtres humains et machines permettant une exploitation, qualitativement supérieure, de grands volumes d'informations et de services variés. Le Web actuel est essentiellement syntaxique, dans le sens où la structure des documents est bien définie, mais que son contenu reste quasi inaccessible aux traitements machines. Seuls les humains peuvent interpréter leurs contenus. Le Web sémantique a pour ambition de lever cette difficulté. Les ressources du Web seront plus facilement accessibles aussi bien par l'homme que par la machine, grâce à la représentation sémantique de leurs contenus. Le Web sémantique est d'abord une infrastructure pour permettre l'utilisation de connaissances formalisées en plus du contenu informel actuel du Web, même si aucun consensus n'existe sur jusqu'où cette formalisation doit aller. Cette infrastructure doit permettre d'abord de localiser, d'identifier et de transformer des ressources de manière robuste et saine tout en renforçant l'esprit d'ouverture du Web avec sa diversité d'utilisateurs. Elle doit s'appuyer sur un certain niveau de consensus portant, par exemple, sur les langages de représentation ou sur les ontologies utilisées. Elle doit contribuer à assurer, le plus automatiquement possible, l'interopérabilité et les transformations entre les différents formalismes et les différentes ontologies. Ainsi le Web sémantique offre la possibilité aux machines de comprendre et d'exploiter les ressources du web de manière inter opérable. Pour cela le W3C propose des formalismes dotés d'une syntaxe XML permettant de modéliser les concepts du web, de les instancier et de les interroger (W3C).

2.3 Le Web Sémantique peut-il être social ?

Cette question mérite d'être posée car des articles comme (Caussanel et al., 2002), (Zacklad et Barbaud, 2004) et (Zacklad, 2005) ont défendu l'importance de la dimension sociale dans la construction d'un cycle de vie du Web sémantique et proposent une nouvelle approche - le Web socio sémantique - que les auteurs opposent radicalement, à l'approche classique du Web sémantique. Dans leur démarche ils scindent le Web sémantique en deux entités qu'ils opposent : le Web computationnellement sémantique et le Web cognitivement sémantique.

Selon eux le Web computationnellement sémantique « vise essentiellement à automatiser la recherche d'information via des agents logiciels (…) et qu'on représentera les ontologies et les réseaux sémantiques à l'aide de langages formels supportant des inférences et des traitements puissants, comme les langages logiques ou orientés objets » (Caussanel et al., 2002) alors que le Web cognitivement sémantique « vise à soutenir les activités de recherche d'utilisateurs humains dans des corpus complexes et évolutifs » (Zacklad, 2005). Ainsi « tout en prolongeant cette perspective, le Web socio sémantique se positionne vis-à-vis du "Web

social" (…) et vise lui à soutenir des activités de coopération plus structurées dans lesquelles les interactions s'appuient également sur des informations ou des documents partagés par un collectif poursuivant, au moins pour un temps, des objectifs communs » (Zacklad et Barbaud, 2004). Mais comme le montre (Gandon, 2006), il existe une grande différence entre le Web sémantique et les logiques formelles. Selon Gandon, le Web sémantique constitue une famille de langages d'expressivité croissante dont la brique de base n'est pas une logique mais le modèle pivot de RDF (un modèle de triplets pour représenter des graphes descriptifs des ressources) et le Web sémantique ne s'oppose pas aux dimensions sémiotiques, sociales ou pragmatiques du Web. Cependant depuis cet article le camp du Web socio sémantique a changé son approche selon Manuel Zacklad « en considérant qu'il existait bien une forme de complémentarité » entre les deux même s'il revendique « que l'initiative du Web socio sémantique est un courant particulier à l'intérieur du Web sémantique ».

Dans notre contexte nous voyons deux formes de socialité.

La première forme est le réseau social de communautés. Dans ce réseau, les formalismes du W3C nous permettent de modéliser notre réseau social, ce qui est en adéquation avec la position défendue par Fabien Gandon qui soutient que le « Web Sémantique n'est pas antisocial » (Gandon, 2006) car le Web sémantique est une évolution et non une révolution du Web. De plus, comme nous comptons utiliser certains indicateurs du Web social nous aurons besoin d'un langage de requêtes puissant or « le Web cognitivement sémantique ne permet généralement pas de faire des inférences logiquement valides de manière automatique » (Caussanel et al., 2002) .

La deuxième forme de socialité se situe au sein de la communauté. Pour sa prise en compte, il sera certainement nécessaire de recourir au Web socio sémantique pour avoir les différents points de vue des membres d'une communauté ce qui va amener « l'intercompréhension qui englobe l'ensemble des questions liées aux dimensions culturelles et linguistiques pour l'établissement d'un accord entre des participants » selon Manuel Zacklad.

2.4 Représentation et analyse sémantiques des réseaux sociaux

Le projet FOAF (Friend Of A Friend) est un des plus grands projets sur le Web Sémantique. FOAF est devenu un vocabulaire standard largement accepté pour représenter des réseaux sociaux (Finin et al, 2005), (Goldbeck et al, 2008) et (Erétéo et al 2008). Cependant, il s'agit d'un vocabulaire RDF permettant de décrire des personnes et les relations qu'elles entretiennent entre elles alors que dans notre approche nous voulons modéliser les connaissances socioculturelles des différentes localités. L'utilisation de l'ontologie FOAF ne s'avère donc pas adéquate dans notre contexte.

C'est pourquoi nous proposons l'utilisation d'une ontologie OWL (Web Ontology Language) que nous allons construire en s'appuyant sur la théorie de Vygotsky.

Par ailleurs, l'analyse des réseaux sociaux s'intéresse aux nœuds et non aux types de nœuds. Nous pouvons donc l'utiliser dans notre cas. Toutefois les approches actuelles des algorithmes d'analyse des réseaux sociaux sont basées sur des définitions et les caractéristiques des graphes représentant les réseaux sociaux. La sémantique des indicateurs me-

surés n'est pas prise en compte. Les données sociales décrites en RDF forment un graphe typé qui fournit une représentation plus puissante et plus riche des réseaux sociaux du web par rapport aux modèles de graphe classiques d'analyse des réseaux sociaux. La plupart des recherches visent à calculer les métriques des réseaux sociaux en utilisant les relations «knows» et «interest» de l'ontologie FOAF (Erétéo et al 2009) avec le langage de requête SPARQL (SPARQL Protocol And RDF Query Language), une recommandation du W3C, qui permet tout particulièrement, l'interrogation de descriptions RDF. Cependant SPARQL montre certaines limites concernant l'analyse sémantique des réseaux sociaux. Comme le montre (San Martin et al, 2009), RDF et SPARQL présentent toutes les caractéristiques pour l'échange, l'interopérabilité, la transformation et l'interrogation de données sociales sur le web. Toutefois ils montrent aussi que la version standard de SPARQL n'est pas assez expressive pour effectuer des requêtes "globales" sur un réseau social, nécessaires pour calculer la plupart des métriques de l'analyse des réseaux sociaux. Ainsi des extensions à SPARQL sont proposées comme SPARQLeR (Kochut et Janik, 2007) qui permet de manipuler plus de caractéristiques sur les chemins : contraintes sur la longueur des chemins, possibilité d'imposer la présence d'une ressource sur les chemins, prise en compte ou non de l'orientation des chemins qui sont par nature orientés dans les graphes RDF, expressions régulières permettant de filtrer la séquence et type des ressources et propriétés contenues dans les chemins, prise en compte du polymorphisme des ressources.

Nous ne pouvons pas terminer cette partie sans mentionner les travaux qui ont été fait autour du projet Towntology[1] même s'ils n'ont pas été développés dans le cadre du Web sémantique. Nous en faisons état car leur finalité – concevoir une ontologie urbaine – nous semble être une partie de nos travaux parce que vouloir modéliser les aspects socioculturels d'une localité implique forcément une prise en compte de l'aspect urbain. Cependant lors du développement de cette ontologie, ses concepteurs ont senti le besoin de développer leur propre langage basé sur XML, ce qui fait qu'il nous est impossible de la réutiliser dans notre contexte d'autant plus que l'un des buts du projet Towntology est de proposer « un outil de conception d'ontologie pré-consensuelle pour aider les experts du domaine (dans notre cas les urbanistes) à décrire leur connaissance. » (Keita et al., 2006). Néanmoins dans notre modélisation nous comptons utiliser certains concepts (classes) du projet Towntology pour construire notre ontologie socioculturelle.

3 Ontologie socioculturelle

Nous proposons une méthodologie pour identifier les caractéristiques représentant une communauté dans ses aspects sociaux (au sens large). La modélisation de ces caractéristiques formera notre ontologie socioculturelle. L'approche que nous utilisons s'inspire d'un processus appelé « Vygotskian Framework » proposé par Vygotsky. Ce processus examine la relation entre le savoir et le développement d'une société suivant trois axes : a) les humains (subject), b) les objets (bâtiments, parc, etc.), c) les artefacts (outils mentaux, outils matériels, images, dessins visuels, etc.).

[1] http://www.towntology.net/

Nous entendons par méthodologie, les procédures de travail, les étapes, qui décrivent le pourquoi et le comment de la conceptualisation puis de l'artefact construit. L'absence de directives structurées et communes fait qu'il n'existe pas un seul mode ou qu'une seule méthodologie « correcte » pour développer des ontologies (Noy et al, 2001).

3.1 La « Vygotskian Framework »

La théorie de Vygotsky (Ivic, 1994) est un espace métaphorique représentant le lieu de développement cognitif, lieu occupé par des pairs, des experts, et de tout autre dispositif susceptible de contribuer au développement. La théorie offre, au moins en principe, la possibilité de conceptualiser de façon scientifique des processus métacognitifs, qui permet de lier cette dimension du développement cognitif au développement général et de comprendre l'origine de cette capacité du sujet à contrôler ses propres processus intérieurs par le schéma de la figure 1 et qui décrit le passage du contrôle extérieur et interindividuel au contrôle intrapsychique individuel.

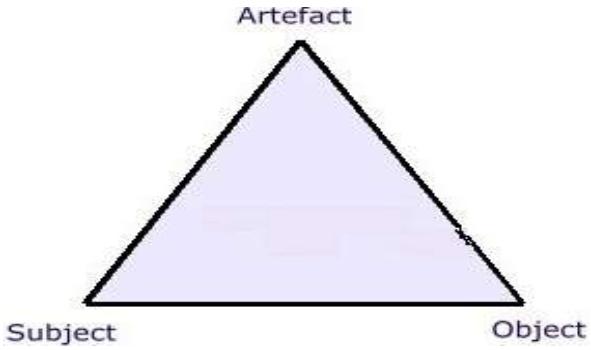


FIG. 1 – *Triangle de médiation de Vygotsky.*

Ainsi nous pourrions dire que la théorie de Vygotsky est une « théorie socio-historico-culturelle du développement». Avec les trois axes de la théorie de Vygotsky on peut modéliser les différents concepts de l'ontologie :

- *Subject* : puisque dans notre « réseau social » on substitue les personnes aux communautés, cet axe représentera les communautés.
- *Object* : cet axe correspond aux différentes infrastructures d'une localité.
- *Artefact* : cet axe représente les activités socioculturelles d'une communauté, les faits historiques et les localités d'un terroir.

3.2 Concepts et Relations

Les trois axes du processus de Vygotsky constituent les classes fondamentales de l'ontologie. La figure 2 montre un extrait de l'ontologie qui regroupe les principales classes et leurs sous-classes. Ainsi nous retrouvons les trois axes de la théorie de Vygotsky avec les différentes sous-classes des trois concepts. Notons que la classe *Site_Historique* est l'union

des classes *Espace_Bati* et *Espace_Non_Bati*. Avec ces trois classes et leurs sous classes nous pouvons capturer l'ensemble des connaissances socioculturelles d'un terroir.

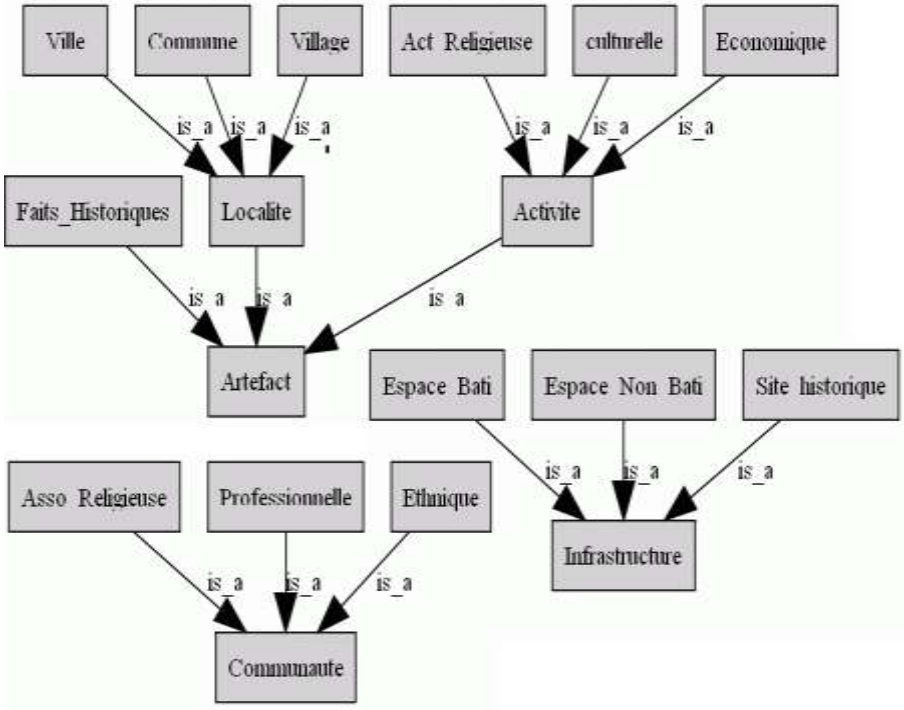


FIG. 2 – *Extrait de notre ontologie Socioculturelle (Concepts de base).*

Les classes seules ne fourniront pas assez d'information, il faudrait leur associer des attributs qui jouent un rôle essentiel dans le développement d'une ontologie. Ils décrivent les propriétés des classes et des instances. Mais malheureusement par manque de place, nous ne pouvons pas détailler les attributs de nos différentes classes. Cependant nous allons présenter deux ontologies que nous avons réutilisées, qui sont des recommandations du W3C: l'ontologie OWL-Time[2] proposée pour la modélisation de phénomènes temporels complexes pour le Web sémantique et l'ontologie GeoRSS-Simple[3] qui constitue un vocabulaire de référence pour la description des propriétés géospatiales des ressources Web.

Nous exploitons l'ontologie OWL-Time en définissant deux relations entre, respectivement, les concepts *Activite* et *Faits_Historiques* de l'ontologie socioculturelle et les concepts *Interval* et *DateTimeDescription* de OWL-Time. Avec la première relation le concept *Activite* dispose des propriétés comme *hasBeginning* et *hasEnd* qui marqueront respectivement l'instant de début et de fin d'une activité. Comme le concept *DateTimeDescription* est utilisé

[2] http://www.w3.org/TR/owl-time/
[3] http://www.georss.org/simple

pour décrire des intervalles implicites, tels que “8 Mai 2007 à 12H 03mn 08s”, qui représentent un intervalle de 24 heures, avec la deuxième relation nous bénéficions de ce type de description pour notre concept *Faits_Historiques*.

Pour ce qui concerne l’ontologie GeoRSS-Simple, nous définissons une relation entre le concept *Infrastructure* de l’ontologie socioculturelle et le concept *gml:_Feature* de GeoRSS. Grâce au constructeur *owl:equivalentClass* du langage OWL-DL, nous obtenons toutes les propriétés du concept *gml:_Feature*. Ainsi plusieurs attributs - *box*, *point*, *line* et *polygon* - peuvent être utilisés pour attacher aux instances du concept *Infrastructure*, des géométries concrètes, spécifiées en utilisant des chaînes de caractères qui respectent un format donné. De même nous bénéficions de la relation *where* qui permet de lier les instances de *Infrastructure* et les différentes géométries du concept *gml:_Geometry*.

Les relations sont, de même que les classes, des éléments importants dans le développement d’une ontologie. Un choix de conception qui doit être fait durant l’élaboration d’une ontologie est de décider si une connaissance doit être modélisée dans une propriété ou à l’aide d’une relation pointant sur un autre concept. Un critère peut être de dire que c’est une propriété dès lors que les valeurs possibles sont d’un type dit primitif (entier, chaîne de caractères), et c’est une relation dès lors que les valeurs possibles sont d’un type dit complexe c’est-à-dire un autre concept de l’ontologie. Ainsi la figure 3 illustre les différentes relations qui peuvent exister entre les classes de notre ontologie socioculturelle.

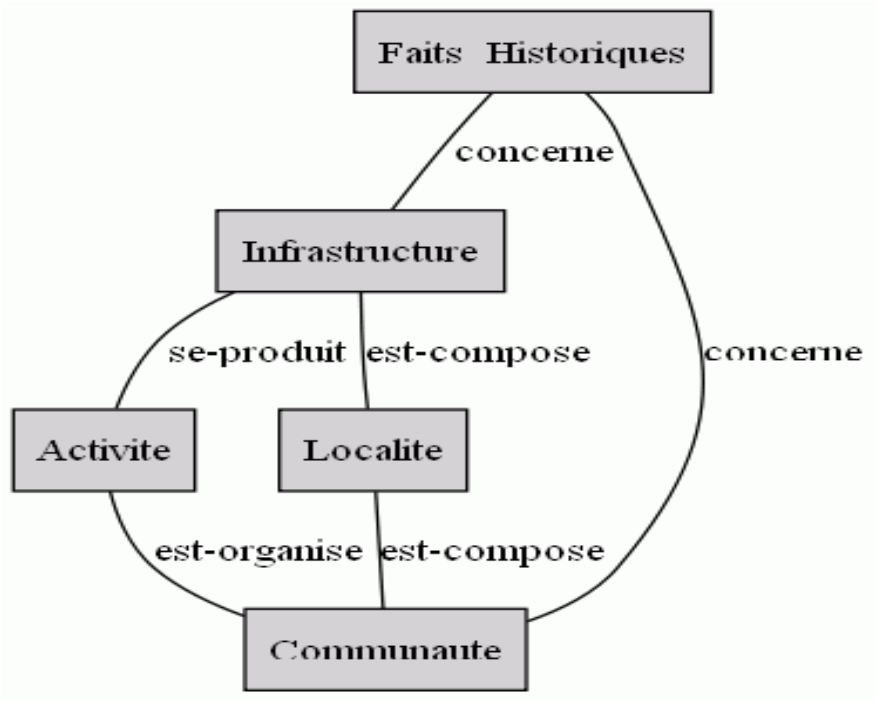


FIG. 3 – *Relations entre les classes.*

Avec ces relations, nous retrouvons le triangle de médiation de Vygotsky à différents niveaux. Elles permettent de représenter différentes connaissances socioculturelles:

- Avec les relations « est-organise », « est-compose » et « se-produit » nous pourrons savoir les différents centres d'intérêts d'une « Communaute » en fonction des « Activite » qu'elle organise. Ainsi de même nous aurons une idée des manifestations qui se produisent dans une «Localite».
- Avec la relation « concerne » nous aurons les différents récits historiques d'une «Infrastructure » ou d'une « Communaute ». Notons que l'image de la relation est défini comme l'union entre les classes « Infrastructure » et « Communaute ».

3.3 Analyse du « réseau social »

L'analyse de notre « réseau social » sera faite en considérant deux niveaux dans la définition des métriques.

Un premier niveau où l'on considère un réseau social intra-communautaire. Dans ce cas les éléments de base considérés sont les différents composants d'une communauté. Par composant on entend les associations qui sont dans une localité par exemple. En considérant les associations on peut calculer la centralité de degré pour voir celles qui sont plus actives en considérant la relation « Organise » c'est-à-dire celles qui organisent le plus d'activités. De même, pour une localité, nous pourrons savoir les différentes infrastructures et associations qui s'y trouvent et leur nombre grâce à la centralité de degré. Ainsi avec la centralité de degré il sera possible de trouver plusieurs informations concernant une localité.

Un second niveau consiste à considérer un réseau intercommunautaire. Dans ce cas les éléments de base considérés sont les localités qui composent notre « réseau social ». Comme dans notre premier niveau nous pouvons connaître l'importance des activités des différents composants d'une communauté. Ici l'idée est de créer un nouvel indicateur pouvant montrer la "similitude d'intérêts" entre communautés. La mesure de similitude est construite par rapport aux diverses activités intra-communautaires. Cet indicateur nous permet de scinder en clusters les communautés en fonction de leur centre d'intérêts. Par exemple en calculant la centralité de degré des différentes activités qui sont organisées dans une localité, si nous nous rendons compte que les associations religieuses sont plus actives nous pouvons dire que celle-ci a un intérêt religieux.

Avec ces différentes métriques nous pourrons avoir les centres d'intérêts de chaque communauté et avec l'indice de "similitude d'intérêts" il sera possible de scinder le réseau.

4 Conclusion

Dans cet article, nous avons présenté une méthode de développement d'ontologie socioculturelle dans le but de vulgariser et de pérenniser la culture d'un pays à travers un partage des coutumes et l'histoire des différentes localités du pays. Cette méthode s'articule autour du processus dénommé « Vygotskian Framework » qui nous a permis de modéliser les principaux concepts de l'ontologie. En s'appuyant sur ce qui se fait au niveau du Web social, nous avons défini un nouveau point de vue de la notion de communauté qui est moins portée

sur les individus (qui sont classiquement les points centraux) que sur les croyances et connaissances que ceux-ci partagent. En modélisant sémantiquement les ressources manipulées dans notre « réseau social », nous exploitons une approche sémantique de l'analyse des réseaux sociaux. Ainsi nous avons divisé notre réseau en deux niveaux. Le premier niveau nous permet d'avoir des informations riches concernant une communauté et le deuxième niveau nous permet de scinder nos communautés en fonction de leur centre d'intérêt grâce à notre indice de "similitude d'intérêts".

Nous venons de terminer une enquête réalisée dans la région de Louga au Sénégal et envisageons, dans un premier temps, utiliser la monographie obtenue pour peupler et valider l'ontologie socioculturelle.

Ensuite, il s'agira de concevoir une plate-forme web sémantique autour de cette ontologie, plate-forme qui doit servir de cadre de partage de connaissances sur les communautés sénégalaises.

Références

Berners-Lee T., Hendler J., Lassila O. (2001). The Semantic Web. Scientific American, Vol 284, n 5, pages 34-43, Mai 2001.

Caussanel J, Cahier J.-P., Zacklad M., Charlet J. (2002). Les Topic Maps sont-ils un bon candidat pour l'ingénierie du Web Sémantique ? Conférence Ingénierie des Connaissances IC'2002.

Erétéo, G., Gandon, F., Buffa, M., Grohan, P., Leitzelman, M., Sander P. (2008). A state of the art on Social Network Analysis and its applications on a semantic web. Proc. SDoW2008 (Social Data on the Web), Workshop held at the 7th International Semantic Web Conference, Karlsruhe, October 2008.

Erétéo, G., Gandon, F., Buffa, M., Grohan, P. (2009). Analyse des réseaux sociaux et web sémantique : un état de l'art. Rapport interne, Edelweiss, INRIA.

Finin, T., Ding, L., Zou, L. (2005) Social networking on the semantic web. The Learning organization journal. Vol. 12, Number 5, pages 418--435

Freeman, L. C. (1979). Centrality in social networks: Conceptual Clarification. Social Networks. 1, 215-239.

Gandon F. (2006). Le Web sémantique n'est pas antisocial », actes des 17ème journées francophones de l'ingénierie des connaissances, IC 2006

Goldbeck, J., Rothstein, M. (2008) Linking social Networks on the web with FOAF. Proceedings of the twenty-third conference on artificial intelligence, AAA08. (2008).

Gruber, T. (1993). A translation Approach to portable ontology specifications. Knowledge Acquisition. Vol. 5. 1993. 199-220.

Gruber T. (2007). Collective Knowledge Systems: Where the Social Web meets the Semantic Web. Web Semantics: Science, Services and Agents on the World Wide Web.

Ivic, I. (1994). Lev. S. Vygotsky. Perspectives : revue trimestrielle d'éducation compare (Paris, UNESCO : Bureau international d'éducation), vol. XXIV, n° 3/4, 1994 (91/92), p. 793-820.

Keita A. K., Roussey, C., Laurini, R. (2006). Un outil d'aide à la construction d'ontologies pré-consensuelle: le projet Towntology. In actes du 24ème congrès d'Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID).

Kochut, K., J., Janik, M. (2007). SPARQLeR: Extended Sparql for Semantic Association Discovery. Proceedings of ESWC '07 Proceedings of the 4th European conference on The Semantic Web: Research and Applications.

Moreno J.L. (1933), Emotions mapped by new geography, New York Times (1933).

Noy N. F., McGuinness D. L. (2001). Ontology Development 101: a guide to creating your first ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05.

San Martin, M., Gutierrez, C., (2009). Representing, Querying and Transforming Social Networks with RDF / SPARQL. ESWC'09.

Scott, J. (2000). Social network analysis, a handbook. Deuxième édition, Edition Sage.

Uschold M. (1998). Knowledge level modeling : Concepts and terminology. Knowledge Engineering Review, 13 : 1, 1998, 5-29. Printed in the United Kingdom. Copyright 1998, Cambridge University Press.

Zacklad, M. (2005). Introduction aux ontologies sémiotiques dans le Web Socio Sémantique. In actes des 16èmes journées francophones d'Ingénierie des Connaissances, Grenoble.

Zacklad, M. et Barbaud, X. (2004). Vers une application du Web Socio Sémantique pour la réalisation d'un système d'information destiné aux réseaux de santé, Second séminaire francophone du Web Sémantique Médical – 9 mars 2004, Rouen.

W3C. Semantic Web Activity, http://www.w3.org/2001/sw/ et http://www.w3.org/2001/sw/Activity

Summary

In this paper we propose a process of ontology development in the sociocultural context to popularize and perpetuate the culture of a country through a shared customs and history of different localities. It can be compared with the construction of a platform that would straddle a « corporate memory » and a « social network », but applied in the context of a country. This process is based on the theory of Russian psychologist Lev Vygotsky called « Vygotskian Framework ».

Index des auteurs