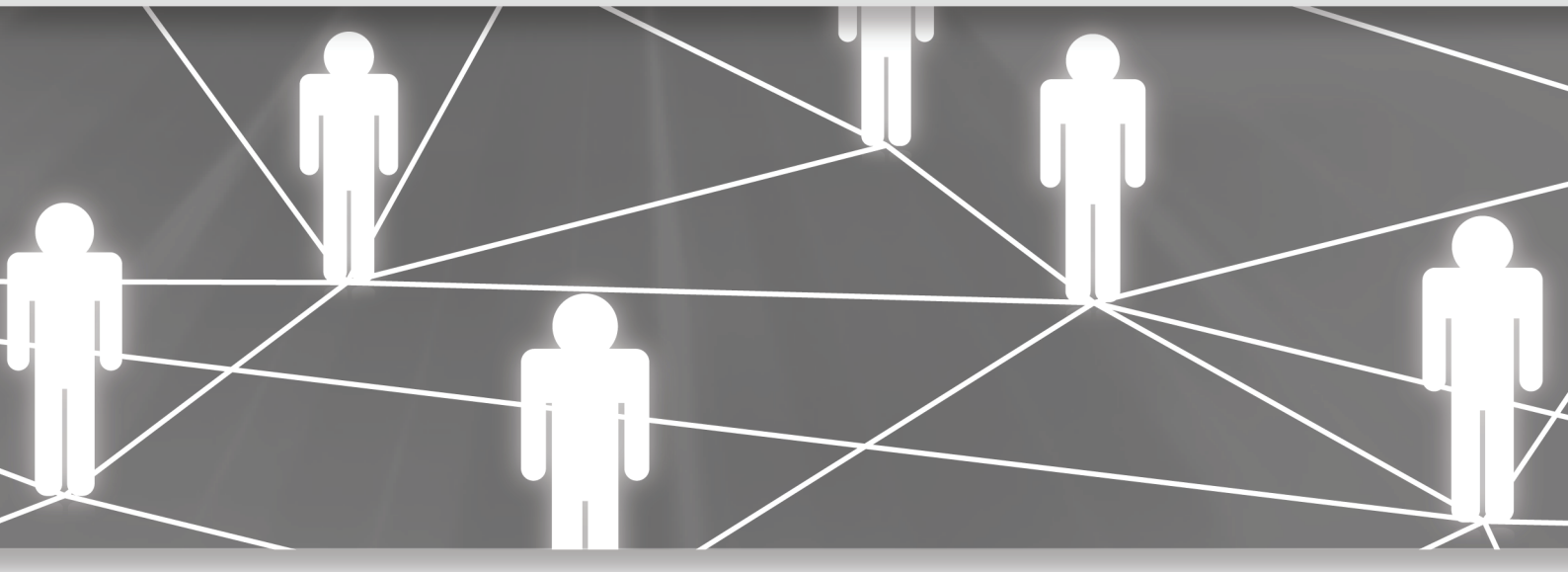

REiSO 2010

Atelier REcherche et REcommandation d'information dans les RESeaux sOciaux

En conjonction avec le 28ème Congrès sur l'INFormatique des ORganisations et des Systèmes d'Information Décisionnels (INFORSID 2010) Marseille.

Marseille, France
25 Mai 2010
http://www.irit.fr/REISO

Organisateurs
Lynda Tamine, IRIT Toulouse
Eric Gaussier, LIG Grenoble
Gregory Grefenstette, Exalead France

Atelier REcherche et REcommandation
d'information dans les RESeaux sOciaux

REiSO 2010

En conjonction avec le 28ème Congrès sur l'INFormatique des ORganisations et des Systèmes d'Information Décisionnels (INFORSID 2010) Marseille.

Organisateurs
Lynda Tamine, IRIT Toulouse
Eric Gaussier, LIG Grenoble
Gregory Grefenstette, Exalead France

Marseille, France
25 Mai 2010
http://www.irit.fr/REISO

Préface

Les applications sociales sont, depuis leur apparition dans les années 1990, en mutation constante. Chaque année apporte de nouvelles directions de partage, de recommandation, et de production interactive d'information, créant de nouvelles connexions entre utilisateurs et de nouvelles communautés dynamiques.

Cet atelier a pour ambition de débattre, à travers la présentation de travaux théoriques, empiriques, ou de démonstrateurs, des réponses possibles à des questions fondamentales : Comment puiser dans ces nouveaux flux d'information pour comprendre les nouveaux facteurs liés à la connectivité et l'interactivité des utilisateurs, leurs nouveaux besoins, leurs avis, leur position sociale? Comment traiter de nouvelles tâches telles que la production et la recherche collaboratives d'information, la recherche d'opinion, la recherche d'expert, etc.

Au-delà de ces aspects, cet atelier vise également à encourager le développement de la thématique "recherche et recommandation d'information dans les réseaux sociaux" en offrant un lieu privilégié d'échanges entre jeunes chercheurs, chercheurs, et industriels, d'où devraient émerger les besoins et les enjeux futurs de la discipline.

La liste non exhaustive des thèmes de l'atelier est :

- Recherche, production, recommandation d'information sociale
- Recherche dans les blogs et tweets
- Recherche, détection et synthèse d'opinion
- Recherche et génération d'annotations sociales (e.g. folksonomie)
- Fraîcheur, autorité, fiabilité dans les réseaux sociaux
- Recherche de personnes, d'experts
- Interactions collaboratives pour la recherche ou recommandation sociale
- Systèmes de Question-Réponse orientés communauté
- Evaluation de systèmes de recherche ou de recommandation sociale
- Découverte, analyse, évolution de communautés
- Apprentissage dans les réseaux sociaux
- Applications de la recherche d'information sociale : réseaux de professionnels, réseaux de marketing, réseaux de divertissement, réseaux bibliographiques
- Transversalité : fondements des réseaux sociaux en mathématiques, psychologie, sociologie...

Lynda Tamine, Eric Gaussier et Gregory Grefenstette

Organisation

La 1ère édition d'atelier "REcherche et REcommandation d'information dans les RESeaux sOciaux (REiSO 2010)" est organisée en conjonction avec le 28ème Congrès sur l'INFormatique des ORganisations et des Systèmes d'Information Décisionnels (INFORSID) Marseille, 25 Mai 2010.

Organisateurs

- Lynda Tamine, IRIT Toulouse
- Eric Gaussier, LIG Grenoble
- Gregory Grefenstette, Exalead France

Comité de Programme

- Frédéric Amblard, IRIT Toulouse
- Gilles Bisson, Université de Grenoble, La Tronche
- Mohand Boughanem, IRIT Toulouse
- Guillaume Cabanac, IRIT Toulouse
- Boris Childovskii, Centre de Recherche Xerox
- Ludovic Denoyer, LIP6 Paris
- Bernard Fallery, Université de Montpellier 2
- Patrick Gallinari, LIP6, Paris
- Christine Largeron,Université Jean Monnet, Saint Etienne

Actes de l'atelier

- Lamjed Ben Jabeur, IRIT Toulouse

Remerciements

Nous remercions les organisateurs de 28ème Congrès sur l'INFormatique des ORganisations et des Systèmes d'Information Décisionnels (INFORSID) pour leur soutien.

Table des matières

Des réseaux de photos aux réseaux sociaux. Création et utilisation d'un réseau social à partir de photos.

Michel Plantié*, Michel Crampes*

*EMA-LGI2P, Parc Scientifique Georges Besse, 30035 Nîmes Cedex, France
{michel.crampes, michel.plantie}@ema.fr, http://www.lgi2p.ema.fr

Résumé. Avec les nouvelles possibilités apparues dans les domaines des communications et du management, les réseaux sociaux et les photos font l'objet d'un intérêt grandissant dans le monde numérique. Dans cet article, nous montrons comment les photos sociales, prises lors d'événements familiaux ou lors de soirées entre amis, représentant des individus ou des groupes d'individus, peuvent être considérées comme génératrice d'un réseau social révélant des attributs sociaux. A partir de ce réseau de photos, nous constituons des sous-groupes de personnes dans l'objectif de diffuser des albums photos personnalisés. L'organisation du réseau de photos utilise des méthodes d'analyse formelle de concepts.

1 Introduction

Avec l'arrivée des photographies digitales et la disponibilité récente des appareils photos dans les téléphones mobiles, il est maintenant possible pour une seule personne de prendre des milliers de photos en une seule année. Dans le passé, une personne prenait en général un petit nombre de photos au cours d'un événement. Ceci était en particulier dû au coût de développement des photos qui générait clairement une limitation du photographe amateur. Aujourd'hui, le problème est très différent avec l'avènement des technologies numériques. Actuellement, chacun fait face à d'autres problèmes soulevés par l'abondance : comment organiser, visualiser, rechercher et partager des milliers de photos.

Dans ce papier, nous considérons uniquement les photos "sociales" indexées manuellement. Nous nous référons avec ce terme aux photos qui sont prises durant des événements familiaux ou lors de soirées entre amis. Ces photos représentent des individus ou des groupes de personnes. En considérant le rôle important que les photos numériques semble prendre pendant des événements "sociaux", il est intéressant d'explorer leur capacité à agir en tant que nouveau langage social reflétant des relations sociales. Il est aussi pertinent d'explorer de nouvelles applications dérivées de ce nouveau langage permettant de renforcer la socialisation. Des applications telles que "Facebook" ou Flick'r manipulent déjà des photos dans des réseaux sociaux. Cependant, dans ces cas les réseaux sont construits en utilisant les personnes. Les photos sont uniquement des objets passifs qui sont volontairement partagés. Elles ne jouent pas un rôle actif dans la construction des réseaux sociaux. Pour considérer les photos comme des acteurs sociaux, il est nécessaire d'interagir avec elles en les considérant comme des objets

actifs. Cette exigence doit être appliquée à différents niveaux dans le cycle de vie du réseau de photo et du réseau social. Chaque niveau est développé dans les sections ci-dessous. La section 2 décrit les techniques pour organiser les photos sociales de manière à révéler les relation sociales entre personnes. Pour remplir cet objectif, nous utilisons les méthodes d'analyse formelle de concepts pour construire les réseaux de photos expressifs, en particulier les hiérarchies de concepts et les diagrammes de Hasse. La section 3 présente plusieurs solutions pour extraire des réseaux sociaux de collections de photos organisées. Les albums photos personnalisés peuvent être déduits à partir du réseau résultant et visuellement lié aux personnes choisies afin de vérifier qui peut recevoir quelles photos. La section 4 présente un application pratique de constitution de réseaux. La section 5 présente la constitution d'albums personnalisés à partir des réseaux construits.

2 Photos sociales et réseaux sociaux

Organiser socialement des photos et élaborer des réseaux sociaux à partir du contenu des photos est depuis peu un domaine très actif de recherche, cependant nous ne connaissons pas d'article qui intègre l'organisation de photos et l'extraction de réseaux sociaux. La méthode la plus proche de nos travaux est présentée en Golder (2008). L'auteur justifie le fait que les photos représentant des personnes peuvent être utilisées pour élaborer des réseaux sociaux. La méthode d'évaluation est bien fondée, et les résultats sont proches des nôtres, bien que notre méthode d'évaluation ne soit pas aussi précise que celle présentée dans le papier cité. Cependant, l'auteur ne présente pas de réseau social déduit du réseau de photos. Par conséquent, il n'y a pas d'aide pour la construction et le partage d'albums personnalisés. Les photos sont également indexées dans un outil séparé qui n'utilise pas leur contenu social disponible. Il n'y a pas de réelle justification pour l'élaboration du réseau social à partir des photos. Crampes et al. (2009) présente une méthode différente incluant ce couplage entre deux outils que nous étendons dans ce papier. L'utilisation des hiérarchies de Galois pour gérer des ensembles de photos, est décrite dans plusieurs applications telles que dans Eklund et al. (2006); Ferré (2007) . L'analyse formelle de Concepts (FCA, Formal Concept Analysis) et les diagrammes de Hasse ont été introduits, il y a longtemps pour représenter les réseaux sociaux Freeman et White (1993) et restent un domaine fructueux de recherche comme dans Roth et Bourgine (2006).

2.1 Le treillis de concepts de photos indexées

Nous introduisons une caractéristique principale de notre système : la capacité à organiser les photos sociales dans un réseau particulier. Un diagramme de Hasse est utilisé pour révéler le contenu social et élaborer le réseau social sous-jacent. Un diagramme de Hasse est une représentation bien-connue d'un treillis de concepts (ou treillis de Galois). Dans l'Analyse Formelle de Concepts Ganter et Wille (1999), un ensemble d'objets possédant des propriétés (ou attributs) peut être organisé dans un treillis de concepts. Ces concepts contiennent des objets selon leur propriétés communes. Dans notre cas, nous considérons les photos comme des objets et les noms des personnages sur les photos comme des propriétés. Le processus d'organisation débute avec le contexte formel : un tableau avec les objets sur les lignes et les propriétés sur les colonnes. Chaque case est cochée (i.e. valeur =1) si l'objet correspondant possède la propriété correspondante, et reste vide (i.e. valeur =0) si l'objet ne possède pas la

propriété correspondante. Formellement, un concept formel est un triplé (G, M, I) où G est un ensemble d'objets, M un ensemble d'attributs, I est une relation binaire entre les objets et les attributs, i.e. $I \subseteq G \times M$.

L'étape suivante dans la construction du treillis de concepts est de définir les concepts. Un concept est défini par une paire de sous-ensemble : un sous-ensemble d'objets (appelé l'extension) et un sous-ensemble de propriétés (l'intention) que les objets partagent. Dans notre cas, un concept est un sous-ensemble de photos qui partagent le même sous-ensemble de personnes. Formellement, pour un ensemble $O \subseteq G$ d'objets et un ensemble $A \subseteq M$ d'attributs, nous définissons un ensemble d'attributs commun aux objets de O par

$f : 2^G \rightarrow 2^M, f(O) = \{a \in A \mid \forall o \in O, (o, a) \in I\}$

et l'ensemble des objets qui a tous ses attributs dans A, par :

$g : 2^M \rightarrow 2^G, g(A) = \{o \in O \mid \forall a \in A, (o, a) \in I\}$

La paire $\{f, g\}$ est une connexion de Galois entre $(2^G, \subseteq)$ et $(2^M, \subseteq)$.

Un concept formel du contexte (G, M, I) est une paire (O, A) avec $O \subseteq G$, $A \subseteq M$, $A = f(O)$ et $O = g(A)$.

L'étape qui suit le processus d'identification des concepts, est la construction d'un treillis dont les éléments sont les concepts. Nous définissons un ordre partiel sur l'ensemble des concepts, où chaque paire de concepts a un supremum unique (la borne supérieure la plus petite d'un concept ; appelée son joint) et un infimum (la borne inférieure la plus grande ; appelé sa réunion). Formellement, soit L l'ensemble des concepts de (G, M, I) ; l'ordre partiel est défini comme suit : $(o1, a1) \leq_L (o2, a2) \Leftrightarrow a1 \subseteq a2 \Leftrightarrow o2 \subseteq o1$. L'ensemble des concepts L est complété si nécessaire par un concept "top" qui contient tous les objets, et un concept "bottom" qui contient tous les attributs. La paire $(L, \leq_L)$ est appelée treillis de concepts de (G, M, I).

2.2 La sous hiérarchie de Galois

Pour plus de simplicité, il est possible de représenter les extensions réduites avec les "intensions" réduites de concepts. Un extension réduite d'un concept (O, A) est un ensemble d'objets qui appartiennent à O et n'appartiennent à aucun concept de rang inférieur, c.a.d. les objets qui n'ont aucun autre attribut que ceux contenus dans A. De façon duale, l'intension réduite d'un concept (O, A) est l'ensemble d'attributs qui appartiennent à A et n'appartiennent à aucun autre concept de rang supérieur. Nous souhaitons positionner les photos dans le graphe avec les seuls concepts qui contiennent les photos, ainsi nous considérons une réduction du graphe où seul les concepts objets sont considérés. Cela signifie que nous conservons les concepts qui contiennent des objets (photos) à l'endroit où ces objets n'apparaissent qu'une seule fois à leur niveau le plus bas dans la hiérarchie (l'extension réduite). Cet acte de nettoyage est celui proposé en Godin et al. (1995) sous le nom de PCL/X. Ceci est illustré dans notre exemple dans la figure 1. C'est maintenant une sous-hiérarchie de Galois comme expliqué dans Godin et Chau (1999). Pour être plus explicite nous l'appelons une sous-hiérarchie Objet de Galois (SHOG). C'est une visualisation bien plus légère des données où l'on se focalise sur les objets (dans notre cas les photos).

Des réseaux de photos aux réseaux sociaux.

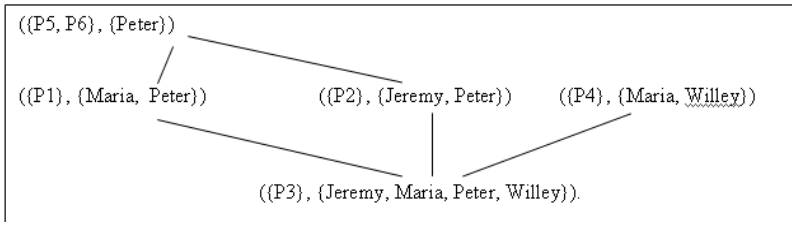

FIG. 1 – *Sous hiérarchie objet de Galois*

3 Des réseaux de photos aux réseaux sociaux

Le diagramme de la figure 1 révèle les différentes co-occurrences entre les personnes dans les photos. Chaque concept montre un groupe particulier de personnes (l'intension du concept) et contient une collection de photos (l'extension du concept) avec ce même groupe de personnes. A partir de ce diagramme, nous pouvons déduire des réseaux sociaux par le calcul de distance, l'inverse de la proximité entre toute paire de personne. Différentes formules de proximité sont possibles et nous introduisons quelques unes d'entre elles. Nous définissons les variables suivantes :

x_i : un individu parmi l'ensemble des personnes trouvées dans la collection de photos

N : est le nombre de personnes de la collection de photos

c : est un concept dans le diagramme de Hasse

C : est le nombre de concepts dans le diagramme de Hasse

(Nota : pour la construction du réseau à partir des photos nous considèrerons uniquement les concepts dont l'intention contient au moins deux personnes, C est donc le nombre de concepts dont l'intention est de cardinalité supérieure ou égale à 2)

n_c : est le nombre de personnes dans un concept c

$[c/x_i]$: est l'ensemble des concepts qui contiennent x_i

$[c/[x_i, x_j]]$: est l'ensemble des concepts qui contiennent uniquement x_i et x_j

$[c/x_i, x_j]$: est l'ensemble des concepts qui contiennent au moins x_i et x_j

$[c/x_i, x_j, x_k]$: est l'ensemble des concepts qui contiennent x_i et x_j et x_k

$[c/x_i \vee x_j]$: est l'ensemble des concepts qui contiennent au moins x_i ou x_j

$Car[.]$: est la cardinalité d'un ensemble particulier

$[c/x_i \oplus x_j]$: est l'ensemble des concepts qui contiennent x_i ou x_j mais non les deux

$[c/\oplus(x_i, x_j)]$: est l'ensemble des concepts qui ne contiennent ni x_i ni x_j

3.1 Un réseau social est un hypergraphe

La plupart des auteurs considèrent qu'un réseau social est un graphe dont les sommets sont les personnes et les arêtes les liens entre ces personnes. Cependant si l'on analyse plus finement la situation, un réseau social est un hypergraphe comme le suggère également Mika (2005). Les hypergraphes sont des objets mathématiques généralisant la notion des graphes, dans le sens où les arêtes ne relient plus un ou deux sommets, mais un nombre quelconque de sommets.

Un hypergraphe H est un couple (V, E) où $V = v_1, v_2, ..., v_n$ est un ensemble non vide (généralement fini) et $E = e_1, e_2, ..., e_m$ est une famille de parties non vides de V. Les éléments de V sont les sommets de H. Les éléments de E sont les arêtes de H. Les hypergraphes correspondent précisément aux matrices de dimension $n \times m$ à coefficients 0 ou 1 (dont chaque colonne a au moins un 1). Tout hypergraphe H correspond de manière univoque à la matrice telle que : $\forall\, a_{i,j} \in A, a_{i,j} = 1$ si $v_i \in e_j$ et $= 0$ sinon. Un hypergraphe est donc décrit par une matrice que l'on nomme matrice d'incidence.

3.2 Le concept de tribu

L'objectif de notre travail est de créer des albums personnalisés selon les liens entre les personnes présentes lors d'un événement et permettre le partage de photos entre les personnes concernées. Plus formellement de trouver les individus x_i (les individus dans les photos) qui sont concernés par l'attribution d'un ensemble de photos. Nous devons pour cela introduire une nouvelle notion. Nous appelons Tribu un sous-ensemble de personnes x_i concernées par un album personnalisé. Tous les individus qui appartiennent à la même tribu possèdent un réseau social commun qui est mis en évidence par l'apparition de ces individus sur différentes photos sous des formes différentes. Nous résonnerons ici sur un couple d'individus x, y, plus précisément sur le lien entre x et y. Notre ensemble de photos et de personnes est un hypergraphe dont les sommets sont des personnes x_i et les arêtes représentent les tribus. Un hypergraphe est la généralisation d'un graphe, et les liens entre les personnes x_i seront toujours représentés par les photos dans l'hypergraphe.

3.3 Règles de cohésion des tribus ou modèle de force

Une tribu est définie par un sous ensemble d'individus obéissant à des règles. Il est possible de définir plusieurs règles d'appartenance. Ci-dessous nous définissons différentes lois qui constitueront des tribus différentes et nous comparerons ensuite ces différents découpages. Dans ce modèle nous considérons les occurrences de personnes en relation dans les concepts. Un concept est représenté par une photo qui caractérise l'intension d'un concept, c.a.d. le groupe de personnes.

3.3.1 La force simple d'un couple

Ici nous définissons la "fréquence" d'un couple (deux individus dans une certaine forme de relation) comme le nombre d'occurrences de ce couple parmi les concepts divisé par le nombre de concepts :

$$FORCESIMPLE(x_i, x_j) \;=\; \frac{Car[c/x_i, x_j]}{C} \tag{1}$$

Cette métrique est intéressante car elle représente la fréquence d'apparition d'un couple capturé par le photographe dans différentes situations sociales. Si un couple est vu dans de nombreuses situations alors la relation entre ces deux personnes est plutôt stable. Par contre cette distance n'exprime pas la force de la liaison entre deux personnes.

3.3.2 La proximité ou force pondérée d'un couple

En conformité avec Golder (2008) nous considérons que plus il y a de personnes dans une photo, moins les liens sont forts entre ces personnes. En conséquence pour exprimer la force d'un couple, le nombre de personnes qui sont présentes dans les photos avec ce couple doit être pris en compte. Avec ce point en tête, l'auteur du Golder (2008) défini la force d'une paire d'individus par la formule suivante :

$$STR(P_a, P_b) = \sum_{i=1}^{m} \frac{1}{\sqrt{(n_i - 1)}} \tag{2}$$

Où P_k est une photo montrant k, $STR(P_a, P_b)$ est la force du lien entre a et b, n_i est le nombre de personnes dans la photo i et m est le nombre de photos (montrant a et b). Nous voyons trois écueils dans cette définition. Tout d'abord elle est définie par le sens commun mais sans justification formelle. Ensuite, la valeur n'est pas bornée (par exemple à 1). Enfin, notre principal intérêt est d'appliquer cette loi à toutes les photos. En conséquence, si un nombre important de photos des même personnes ont été prises à la suite, elles vont renforcer les liens entre ces personnes de façon incorrecte. Nous préférons considérer les concepts et donc toutes les photos qui représentent le même groupe d'individus ne sont comptées qu'une seule fois. Nous estimons que ce qui différencie un couple dans un groupe de personnes est le fait qu'ils sont plus ou moins socialisés, c.a.d. qu'ils apparaissent plus ou moins dans différents groupes de personnes dans différents concepts comme cela est défini dans l'équation 1. S'ils apparaissent une seule fois et seul, leur force devrait être égale à 1 car le couple peut être identifié clairement. Rigoureusement, la force de toute paire d'individus doit être observée dans les concepts et dans un concept particulier, elle doit être proportionnelle à l'existence de ce couple dans un groupe de personnes qui sont présentes dans le concept. Une formule plus évoluée permet de dire qu'un couple qui apparait souvent dans des petits groupes d'individus, il semble logique qu'il y ait plus de chance pour que ce couple soit plus proche que s'il apparaissait dans de grands groupes d'individus. La proximité d'une tribu (respectivement d'un couple) est la somme des inverses des nombres d'individus présents dans les concepts qui contiennent la tribu, divisée par le nombre total de concepts. Donc, plus le couple est "dilué" dans un grand nombre d'individus pour un concept, moins la contribution de ce concept à la proximité est forte. Quand une tribu (un couple) n'apparaît dans aucun concept, la force est nulle. Deux personnes apparaissant dans un seul concept, donne la proximité maximum de $2/2/1 = 1$.

$$Proximite(x_i, x_j) = \frac{(\sum_k \frac{2}{Car[c_k/x_i, x_j]})}{C} \tag{3}$$

$\sum_k$ avec k tel que : $x_i, x_j \in c_k$.

3.3.3 Cohésion d'un couple

Intuitivement, on peut comprendre que le modèle social ci-dessus ne capture pas certaines observations qui peuvent être extraites automatiquement à partir du diagramme de Hasse. Parmi elles, il en existe une qui est particulièrement intéressante. Plus deux personnes sont vues séparément, plus elles sont indépendantes l'une de l'autre. Inversement quand deux personnes sont observées toujours ensemble, cela renforce l'hypothèse de liens particuliers entre

eux. Nous l'appelons "cohésion". Le couple (x_i, x_j) a une forte cohésion lorsque x_i et x_j sont souvent représentés dans le même concept par rapport au nombre de fois où ils apparaissent dans un concept. On a donc :

$$Cohesion(x_i, x_j) = \frac{Car[c/x_i, x_j]}{Car[c/x_i \vee x_j]} \quad (4)$$

3.4 Formation des tribus

Dans cette étape, nous construisons une matrice $B_{n,n}$ qui dépend des règles de formation des tribus. Les règles ci-dessus constituent autant de fonctions tribales. Une fonction tribale révèle le comportement d'un couple d'individus (x_i, x_j) avec $i, j \in [1, n]$. Nous fixons ensuite un seuil ϵ tel que : $\forall b_{i,j} \in B,\ b_{i,j} = 1$ si la fonction tribale est $> \epsilon$ et 0 sinon. On note que $\forall i \in [1, n]\ b_{i,i} = 1$.

3.4.1 Co-occurrence et force de tribus

La notion de tribu et les notions de forces définies ci-avant sont plus générales que la simple co-occurrence entre personnes. Ces notions prennent en compte des mesures sur l'ensemble des concepts et non pas uniquement la présence conjointe de deux personnes.

3.5 Matrice d'incidence de l'hypergraphe

L'étape précédente nous permet d'obtenir, après avoir choisi un seuil ϵ, une matrice B associée à l'hypergraphe qui met en évidence les relations entre chaque individu présent à un événement. A partir de cette matrice, nous devons donc trouver les tribus de ces individus. Ainsi, nous cherchons la matrice d'incidence de l'hypergraphe H tel que $H = T(G)$, avec $T(G)$ matrice triangulaire supérieure de la matrice associée. La i-ème ligne de H représente une tribu. La j-ème colonne de H représente un individu. Pour un i fixé, si $h_{i,j} = 1$ alors l'individu j est dans la tribu i.

4 Application pratique

Nous supposons un événement où cinq personnes Alain (a), Bernard (b), Céline (c), Daniel (d), Emilie (e) ont été pris en photo. Ces photos sont au nombre de 9. Le tableau suivant relate quels sont les individus présents sur chaque photo.

concept	1	2	3	4	5	6	7	8	9
personnes	a	a,c	d,c,e	d,c	a,c,d,e	b,c,d	b	a,d,e	d,e

TAB. 1 – *photos, personnes et concepts*

4.1 Matrices associées à chaque règle de tribus

Nous présentons dans les tableaux 2, les matrices générées par les différentes règles de constitution de tribus, du paragraphe 3.3.

FORCESIMPLE	A.	B.	C.	D.	E.
Alain	1	0	0,22	0,22	0,22
Bernard		1	0,11	0,11	0
Céline			1	0,44	0,22
Daniel				1	0,44
Emilie					1

PROXIMITÉ	A.	B.	C.	D.	E.
Alain	1	0	0,17	0,13	0,13
Bernard		1	0,07	0,07	0
Céline			1	0,31	0,13
Daniel				1	0,31
Emilie					1

COHESION	A.	B.	C.	D.	E.
Alain	1	0	0,29	0,29	0,33
Bernard		1	0,17	0,14	0
Céline			1	0,57	0,29
Daniel				1	0,67
Emilie					1

TAB. 2 – *forces et couples*

4.2 Forces et tribus

le nombre de tribus possibles est en théorie égal à 2^N mais dans la pratique il ne peut pas dépasser le nombre de photos au carré. Les tribus à considérer sont celles obtenues lorsque les forces sont non nulles. Nous en déduisons donc 14 tribus sans compter les individus uniques qui constituent des tribus particulières dont les scores sont 1 pour toutes les forces. Les forces calculées précédemment (voir leur valeurs dans le tableau 3) permettent d'apporter un coefficient de pondération que nous nommerons "CONNIVENCE" d'une tribu. La connivence peut avoir plusieurs significations, si nous adoptons :

1. La valeur "max" des forces pour chaque tribu, exprime le fait que les individus voient leur empathie déterminée par l'évènement le plus heureux qui les rassemble. Dans le cas des photos, un individu x_i se souviendra d'un individu x_j plus par le fait qu'ils ont discuté en tête à tête (témoigné par une photo) que par le fait qu'ils ont été à un moment donné dans un groupe plus large.

2. La valeur "moyenne" des forces pour chaque tribu, à l'inverse, indique que la connivence est une impression générale sur toute la soirée.

3. La valeur "minimum" des forces pour chaque tribu, indique que les personnes d'une tribu se souviendraient des moments de la soirée où elles sont avec d'autres personnes avec lesquelles elles ont moins d'affinités.

Nous voyons ainsi que le choix de la mesure "CONNIVENCE" peut varier selon l'intention que l'on veut donner à une tribu. Prendre une valeur élevée de connivence c'est favoriser les tribus aux liens sociaux étroits. Prendre une connivence faible correspond à l'observation de tribus dont les membres sont faiblement liés et peu soucieux de partager leurs photos (selon le point de vue socio-photographique que nous adoptons dans cet article). Nous adopterons la valeur max du tableau 3 pour pondérer la distance d'une photo à une tribu. dans le tableau 3

force	FORCE SIMPLE	PROXIMITÉ	COHÉSION	max
tribu (a,c)	0,22	0,17	0,29	0,29
tribu (a,d)	0,22	0,13	0,25	0,25
tribu (a,e)	0,22	0,13	0,33	0,33
tribu (b,c)	0,11	0,07	0,17	0,17
tribu (b,d)	0,11	0,07	0,14	0,14
tribu (c,d)	0,44	0,31	0,57	0,57
tribu (c,e)	0,22	0,13	0,29	0,29
tribu (d,e)	0,44	0,31	0,66	0,66
tribu (a,c,d)	0,11	0,06	0,13	0,13
tribu (a,c,e)	0,11	0,06	0,13	0,13
tribu (a,d,e)	0,22	0,13	0,25	0,25
tribu (b,c,d)	0,11	0,07	0,13	0,13
tribu (c,d,e)	0,22	0,13	0,33	0,33
tribu (a,c,d,e)	0,11	0,06	0,13	0,13

TAB. 3 – *forces et tribus*

5 Politique de diffusion de photos, constitution d'albums personnalisés

La constitution des albums personnalisés découle d'une règle simple : toute photo contenant un membre d'une tribu est diffusée à toute la tribu. Nous devons donc déterminer la règle la plus intéressante pour choisir les tribus.

5.1 Distance brute et pondérée d'une photo à une tribu

Afin de déterminer les albums à construire, nous calculons la distance de chaque photo à une tribu. Nous utilisons la distance de jaccard (voir KAUFMAN et al. (1990)) qui mesure la dissimilarité entre deux ensembles. Elle est obtenue en divisant la différence de taille de l'union et de l'intersection de deux ensembles : $J_\delta(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$. Ici nos deux ensembles sont les individus présents dans une photo et les individus présents dans une tribu. Le tableau 4 montre les distances (plus exactement : le complément à 1 des distances) des concepts aux tribus en fonction des différentes tribus. Plus la mesure indiquée est grande plus la distance entre les éléments est faible et leur proximité importante. Ce tableau indique également que plus une tribu est grande plus la somme de ses distances est grandes. On remarque également que certaines photos obtiennent un score somme important qui n'est pas lié uniquement au nombre de personnes présentes dans les photos. Nous utilisons ensuite la valeur "max" du tableau 3 des forces pour pondérer la distance d'une photo à une tribu. Le tableau des distances des concepts aux tribus est donc modifié sur le tableau 5.

5.2 Diffusion de photos

Le critère de choix entre différentes alternatives peut être fondé sur un critère économique : le nombre total d'exemplaires de photos diffusés. Pour diffuser les photos nous pouvons choisir entre cinq alternatives qui vont de la plus restrictive à la plus sociale :

1. Diffusion des photos aux seules personnes présentes sur les photos

concepts	1	2	3	4	5	6	7	8	9	Total
personnes	a	a,c	d,c,e	d,c	a,c,d,e	b,c,d	b	a,d,e	d,e	
tribu (a,c)	0,5	1	0,25	0,25	0,5	0,25	0	0,25	0	3
tribu (a,d)	0,5	0,33	0,25	0,33	0,5	0,33	0	0,66	0,66	3,58
tribu (a,e)	0,5	0,66	0,5	0	0,5	0	0	0,66	0,66	3,5
tribu (b,c)	0,33	0,33	0,25	0,33	0,20	0,66	0,5	0	0	2,61
tribu (c,d)	0	0,5	0,66	1	0,5	0,66	0	0,25	0,66	4,25
tribu (c,e)	0	0,33	0,66	0,33	0,5	0,25	0	0,25	0,66	3
tribu (d,e)	0	0	0,66	0,66	0,5	0,25	0	0,66	1	3,75
tribu (a,c,d)	0,33	0,66	0,5	0,66	0,75	0,40	0	0,5	0,25	4,07
tribu (a,c,e)	0,33	0,66	0,5	0,25	0,75	0,20	0	0,5	0,25	3,45
tribu (a,d,e)	0,33	0,25	0,5	0,25	0,75	0,20	0	1	0,66	3,95
tribu (b,c,d)	0	0,66	0,5	0,66	0,4	1	0,33	0,2	0,25	4,02
tribu (c,d,e)	0	0,33	1	0,66	0,75	0,75	0	0,75	0,66	4,92
tribu (a,c,d,e)	0,25	0,5	0,75	0,5	1	0,6	0	0,75	0,5	4,85

TAB. 4 – *distances brutes : concepts - tribus*

concepts	1	2	3	4	5	6	7	8	9	Total
personnes	a	a,c	d,c,e	d,c	a,c,d,e	b,c,d	b	a,d,e	d,e	
tribu (a,c)	0,15	0,29	0,07	0,07	0,15	0,07	0	0,07	0	0,87
tribu (a,d)	0,13	0,08	0,06	0,08	0,13	0,08	0	0,17	0,17	0,9
tribu (a,e)	0,17	0,22	0,17	0	0,17	0	0	0,22	0,22	1,16
tribu (b,c)	0,06	0,06	0,04	0,06	0,03	0,11	0,09	0	0	0,44
tribu (c,d)	0	0,07	0,09	0,14	0,07	0,09	0	0,04	0,09	0,6
tribu (c,e)	0	0,19	0,38	0,19	0,29	0,14	0	0,14	0,38	1,71
tribu (d,e)	0	0	0,19	0,19	0,15	0,07	0	0,19	0,29	1,09
tribu (a,c,d)	0,22	0,44	0,33	0,44	0,5	0,27	0	0,33	0,17	2,71
tribu (a,c,e)	0,04	0,09	0,07	0,03	0,1	0,03	0	0,07	0,03	0,45
tribu (a,d,e)	0,04	0,03	0,07	0,03	0,10	0,03	0	0,13	0,09	0,51
tribu (b,c,d)	0	0,17	0,13	0,17	0,1	0,25	0,03	0,05	0,06	1
tribu (c,d,e)	0	0,04	0,13	0,09	0,10	0,40	0	0,10	0,09	0,64
tribu (a,c,d,e)	0,08	0,17	0,25	0,17	0,33	0,2	0	0,25	0,17	1,62

TAB. 5 – *distances pondérées : concepts - tribus*

2. Diffusion des photos aux tribus dont le seuil dans le tableau 4 dépasse une valeur donnée pour la tribu. Fixons comme règle pour diffuser une photo à une tribu (c'est à dire à tous ses membres) le fait que la distance d'une photo à cette tribu soit supérieure à $0,5$ au sens de la distance de Jaccard, qui indique qu'une photo montre au moins une majorité des individus de la tribu.

3. Diffusion des photos aux tribus dont le seuil dans le tableau 5 dépasse une valeur donnée pour la tribu. Ce seuil est déterminé par un critère économique c'est à dire le nombre de photos à diffuser. Prenons par exemple le nombre de 35 photos maximum, le seuil est alors de $0,15$.

4. Nous Classons les couples "photo-tribu" par ordre décroissant selon la mesure du tableau 5 et nous diffusons les photos en déterminant un seuil à partir duquel on ne diffuse plus les photos aux tribus désignées. Prenons comme seuil les 10 photos les plus importantes

au sens des couples “photo-tribu”. Le classement pour les 10 premiers couples photo-tribu est montré dans le tableau 6.

5. Diffusion de toutes les photos à toutes les personnes.

Les alternatives 1 et 5 sont déjà pratiquées. Nous allons donc comparer toutes ces alternatives et plus particulièrement les solutions 2, 3 et 4. Prenons la solution 2 nous constatons que toutes

priorité	1	2	3	4	5	6	7	8	9	10
score	0,5	0,44	0,44	0,38	0,38	0,33	0,33	0,33	0,29	0,29
photo	(a,c,d,e)	(a,c)	(d,c)	(c,d,e)	(d,e)	(a,d,e)	(c,d,e)	(c,d,e)	(a,c,d,e)	(d,e)
tribu	(a,c,d,e)	(a,c,d)	(a,c,d)	(c,e)	(c,e)	(a,c,d)	(a,c,d)	(a,c,d)	(c,e)	(d,e)

TAB. 6 – *Classement des 10 premiers couples photos-tribus par ordre décroissant de distance*

les tribus sont sélectionnées pour au moins une photo. Prenons maintenant la solution 3, le tableau 7 montre les photos diffusées. Plusieurs tribus ne sont sélectionnées dans aucune photo : (b,c),(c,d),(a,c,e),(a,d,e),(c,d,e). Selon la solution 4, plusieurs tribus ne sont sélectionnées pour aucune photo : (a,c),(a,d),(b,c),(c,d),(a,c,e),(a,d,e),(b,c,d),(c,d,e). La solution 1 représente 21

concepts	1	2	3	4	5	6	7	8	9
	a	a,c	d,c,e	d,c	a,c,d,e	b,c,d	b	a,d,e	d,e
Alain	X	X	X	X	X	X		X	X
Bernard		X		X		X			
Céline	X	X	X	X	X	X		X	X
Daniel	X	X	X	X	X	X		X	X
Émilie	X	X	X	X	X	X		X	X

TAB. 7 – *diffusion des photos pour chaque personne selon la distance brute de Jaccard*

photos, la solution 2 représente 38 photos, la solution 3 représente 35 photos, la solution 4 représente 27 photos et la solution complète 5 représente 45 photos. Chacune des solutions présente une stratégie de sélection des photos différentes : soit un critère “anti-social”, soit un critère “semi-social”, soit un critère économique, soit un critère sémantique, soit un critère totalement social. Nous disposons maintenant d’un outil de diffusion d’albums personnalisés qui peut procéder soit de manière automatique, soit sous forme d’assistance à la diffusion manuelle. Les différentes stratégies présentées sont modifiables et montre la richesse du modèle de sélection présenté. De nombreuses voies sont encore à explorer. Nous constatons dans notre exemple, que la “proximité” et la cohésion évoluent dans le même sens. Cependant elles expriment des réalités différentes et des exemples plus grands montreront leur différence.

6 Conclusion

De nombreuses améliorations sont possibles, et les modèles peuvent être étendus en particulier la prise en compte de la sémantique pour l’extraction du réseau social. C’est à dire la prise en compte de certains biais comme celui du photographe, qui n’est pas forcément objectif ou de la personne qui produit les photos ou celle qui diffuse les photos. Nos travaux futurs vont

explorer ces modèles, avec des tests plus approfondis. Le passage à l'échelle sera également étudié par l'utilisation de centaines de photos pour révéler l'évolution des réseaux sociaux dans le temps avec la fusion de plusieurs événements sociaux. Nous exploiterons également les principes énoncés ici pour explorer une autre voie : la construction de réseaux sociaux à partir de l'analyse sémantique de documents partagés entre personnes.

Références

Crampes, M., J. de Oliveira-Kumar, S. Ranwez, et J. Villerd (2009). Indexation de photos sociales par propagation sur une hiérarchie de concepts. *Actes de la conférence IC 2009, Hammamet Tunisie*, 13–24.

Eklund, P., J. Ducrou, et T. Wilson (2006). An intelligent user interface for browsing and search mpeg-7 images using concept lattices. *Proc of the 4th International Conference on Concept Lattices and Their Applications, LNAI, Springer-Verlag*.

Ferré, S. (2007). Camelis : Organizing and browsing a personal photo collection with a logical information system. *Proc. of the 5th International. Conference on. Concept Lattices and Their Applications*, 112–123.

Freeman, L. et D. White (1993). Using galois lattices to represent network data. *Sociological Methodology 23*, 127Ű146.

Ganter, B. et R. Wille (1999). Formal concept analysis. *Mathematical Foundations, Springer*.

Godin, R. et T. Chau (1999). Comparaison d'algorithmes de construction de hiérarchies de classes. *L objet 5(3/4)*.

Godin, R., G. Mineau, et R. Missaoui (1995). Incremental structuring of knowledge bases. *Proceedings of the International Knowledge Retrieval, Use, and Storage for Efficiency Symposium (KRUSE'95), Santa Cruz 179-198*.

Golder, S. A. (2008). Measuring social networks with digital photograph collections. *ACM Conference on Hypertext and Hypermedia. June 19-21. Pittsburgh, Pennsylvania*.

KAUFMAN, L., P. ROUSSEEUW, et J. B. P. (1990). *Finding groups in data : An introduction to cluster analysis*. WILEYInterscience.

Mika, P. (2005). Ontologies are us : A unified model of social networks and semantics. *In International Semantic Web Conference*, 522–536.

Roth, C. et P. Bourgine (2006). Lattice-based dynamic and overlapping taxonomies : The case of epistemic communities. *Scientometrics (impact factor : 1.74) 69(2)*, 429Ű447.

Summary

With the new possibilities in communication and information management, social networks and photos have received plenty of attention in the digital age. In this paper, we show how social photos, captured during family events, representing individuals or groups, can be visualized as a network that reveals social attributes. From this photo network, a social network is extracted that can help to build personalized albums. The photo network organization makes use of Formal Concept Analysis methods and a Hasse Diagram representation.

Classification à partir d'une collection de matrices

Clément Grimal*, Gilles Bisson**

*Laboratoire d'Informatique de Grenoble, UMR 5217
Domaine Universitaire de Saint-Martin-d'Hères
38400 Saint Martin d'Hères, France
Clement.Grimal@imag.fr
**Laboratoire TIMC, CNRS / UJF 5525
Université de Grenoble - Domaine de la Merci
38710 La Tronche, France
Gilles.Bisson@imag.fr

Résumé. La classification simultanée de deux types d'objets – par exemple les mots et les documents dans les applications de recherche d'information – encore appelée *co-classification*, a été largement étudiée ces dernières années, et permet de découvrir la structure de ces objets, qu'elle soit explicite (classes) ou latente. Cependant, ce type d'approches ne permet de traiter qu'une seule relation entre deux types d'objets. Or, il existe de nombreux cas où des données multi-relationnelles sont disponibles, chaque paire d'objets pouvant être liée par une relation décrite par des données. Dans le cas des réseaux sociaux, on possède des données de co-occurrence décrivant les relations entre les acteurs et les différents objets. Dans cet article, nous décrivons un travail en cours ayant pour objectif d'exploiter ces données multi-relationnelles afin d'améliorer la qualité des classes obtenues. Ces méthodes sont une extension de l'algorithme de mesure de co-similarité χ-Sim développé par Bisson et Hussain (2008).

1 Introduction

L'objectif de la classification est d'organiser un ensemble d'*individus* suivant des critères de similarité, afin de retrouver, ou de découvrir la structure selon laquelle ils s'organisent. La classification a pour objectif de regrouper les objets dans des classes homogènes et contrastées : ainsi, la similarité d'un couple d'objets appartenant à la même classe doit être maximisée, tandis que celle d'une paire d'objets appartenant à deux classes différentes doit être minimisée. Les données utilisées, issues d'une collection d'observations, sont classiquement « non-structurées », mais représentables sous la forme d'une matrice dans laquelle les lignes correspondent aux *individus* et les colonnes aux *variables* les décrivant. Or, dans de nombreux

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-08-CORD-009, dans le cadre du projet FRAGRANCES.

problèmes, les *variables* sont suffisamment homogènes pour que l'on puisse envisager de les catégoriser au même titre que les *individus*.

Récemment, le domaine de recherche visant à co-classifier deux types d'objets distincts a été intensivement exploré à la fois dans le domaine de la recherche documentaire, où le but est de co-classifier des documents selon la catégorie de sujets qu'ils traitent, et des mots qui les composent ; mais aussi dans le domaine de la bioinformatique afin, notamment, d'analyser les données d'expression génique.

Cependant, dans de nombreuses applications, les données de co-occurrence impliquant plus de deux types d'objets sont courantes, et les méthodes classiques de co-classification ne permettent pas de tirer partie de la richesse de cet aspect multi-dimensionnel. Nous nous proposons donc d'exploiter ces informations supplémentaires pour améliorer la classification des objets qu'elles décrivent. C'est dans ce contexte que nous travaillerons avec une collection de matrices de co-occurrence partageant certaines dimensions. Par exemple, ainsi que nous le verrons dans la partie 4, si le problème est de classifier des *films* selon leur genre, il peut être intéressant de prendre en compte simultanément plusieurs sources d'informations comme les matrices *films - acteurs* et *films - mots-clés*, qui partagent donc la dimension *films*, comme détaillé dans la partie 4.1.

L'objectif de notre méthode est alors de calculer les mesures de similarités entre les objets du même type, en utilisant toutes les données disponibles, le travail se base sur l'algorithme de calcul de co-similarité χ-Sim (Bisson et Hussain, 2008). En effet, l'un des intérêts de cette approche, outre les bons résultats obtenus sur la classification de données matricielles est qu'elle permet à l'utilisateur d'utiliser les matrices de similarité obtenues avec l'algorithme de classification qui lui convient le mieux (K-means, classification ascendante, classification par densité...) pour élaborer les classes d'objets.

L'article est organisé ainsi : dans la partie 2, nous décrivons les différentes classes de problèmes de co-classification, puis dans la partie 3, nous présentons nos extensions de la mesure de co-similarité χ-Sim au cas des données multi-relationnelles ; finalement, dans la partie 4, nous comparons nos mesures avec des méthodes courantes de (co-)classification.

2 Les différentes classes de problèmes

2.1 Bi-classification

Comme décrit dans l'introduction, lorsque les *variables* d'un problème de classification sont suffisamment homogènes, on peut choisir de les catégoriser au même titre que les *individus*. Dans le domaine de la recherche d'information, la co-classification apporte le double bénéfice d'améliorer la qualité des classes lorsque l'on travaille avec des matrices creuses et de grandes dimensions Long et al. (2005) ; mais également de mettre en évidence des ressemblance entre *individus* ayant des *variables*, a priori, très différentes. Ces ressemblances reposent sur l'hypothèse, dans le cas de la classification de textes, que deux documents appartenant à une même thématique, et donc appartenant à la même catégorie, peuvent certes contenir des termes différents, mais que ces termes doivent être a priori fréquents dans la collection des documents relatifs à cette thématique.

Il existe de nombreuses approches permettant de traiter cette classe de problèmes. Dans le domaine de la recherche documentaire, on peut citer entre autres les travaux de Dhillon (2001);

Liu et al. (2002); Dhillon et al. (2003); Long et al. (2005); Nadif et Govaert (2005); Rege et al. (2008); Bisson et Hussain (2008), ainsi que, dans le domaine de la bioinformatique, les travaux de Cheng et Church (2000); Madeira et Oliveira (2004); Speer et al. (2004).

En particulier, l'algorithme développé par Bisson et Hussain (2008) introduisant une nouvelle mesure de co-similarité χ-Sim, permet, en utilisant un algorithme de classification simple, de traiter cette classe de problèmes. C'est donc sur cet algorithme que nous baserons nos propositions dans la partie 3.

2.2 Classification de données multi-relationnelles

Les données issues d'applications réelles ont souvent des structures plus riches que celles traitées par les algorithmes de bi-classification, impliquant différents types d'objets, liés par des relations de co-occurrence, deux à deux. Par exemple, dans le cas de la classification de films selon leur genre, les *films* sont à la fois décrits par une liste d'*acteurs*, et par une liste de *mots-clés*. On peut donc mettre ces données sous la forme de deux matrices de co-occurrence, ayant une dimension commune, les *films*. Dans ce cas, l'objectif est donc de classifier simultanément les *films*, les *acteurs* et les *mots-clés*, en tenant compte simultanément de l'ensemble des relations de co-occurrence qui lient ces données. Il faut noter qu'il n'est pas trivial d'évaluer la classification de tous les types d'objets concernés (ici, les *acteurs* et les *mots-clés*), et l'utilisateur pourra ne pas être intéressé par le résultat de la classification de tous les types d'objets. Cependant, à l'instar de la bi-classification, la qualité de la classification de l'un de ces types d'objets (ici, les *films)*, pourra être grandement améliorée par cette approche.

Remarque 1 : *la représentation sous la forme d'une collection de matrices n'est pas la seule représentation valable. En effet, il est également possible de représenter ces données sous la forme de graphe k-partis*[1]*, chaque couche du graphe représentant un type d'objets différent Long et al. (2006).*

Bien que cette classe de problèmes soit plus récente que la précédente, et ait donc été moins fréquemment abordée, nous pouvons citer, entre autres, les travaux de Bekkerman et al. (2005); Wang et al. (2006); Long et al. (2006); Banerjee et al. (2007); Tang et al. (2009).

C'est cette classe de problèmes que nous traiterons dans la partie 3, en basant nos approches sur l'algorithme de calcul de co-similarité χ-Sim.

3 Algorithmes

3.1 Notations utilisées

De façon classique, les matrices sont notées en caractères majuscules gras, les vecteurs en minuscules gras et les autres variables en minuscules italiques.

Matrices de données : soit $\mathbf{M}_p$ l'ensemble des matrices contenant les informations de la base de données. Chaque matrice comporte r_p lignes (films) et c_p colonnes (acteurs ou mots-clés) : chaque valeur $m_{p,ij}$ caractérise *la nature de la relation* entre le $j^{\text{ème}}$ acteur (ou mot-clé) et le $i^{\text{ème}}$ film ; $\mathbf{m}_{p,i} = \left[m_{p,i1} \cdots m_{p,ic_p}\right]$ est le vecteur ligne correspondant au film i et $\mathbf{m}_p^j = \left[m_{p,1j} \cdots m_{p,r_pj}\right]$ est le vecteur colonne correspondant à l'acteur (ou mot-clé) j.

1. Un graphe est dit k-partis s'il existe une partition de son ensemble de sommets en k sous-ensembles, telle que deux sommets d'un même sous-ensemble ne puissent pas être adjacents.

Matrices de similarité : $\mathbf{SR}_p$ (de taille $r_p \times r_p$) et $\mathbf{SC}_p$ (de taille $c_p \times c_p$) représentent les matrices carrées et symétriques contenant respectivement les valeurs de similarité entre films et entre acteurs (ou mots-clés) avec $sr_{p,ij} \in [0,1]$, $1 \leqslant i,j \leqslant r_p$ et $sc_{p,ij} \in [0,1]$, $1 \leqslant i,j \leqslant c_p$. Le vecteur $\mathbf{sr}_{p,i} = \left[sr_{p,i1} \cdots sr_{p,ir_p}\right]$ (respectivement $\mathbf{sc}_{p,j} = \left[sc_{p,j1} \cdots sc_{p,jc_p}\right]$ indique la similarité entre le film i et l'ensemble des autres films (respectivement entre l'acteur, ou mot-clé j et l'ensemble des autres acteurs ou mots-clés).

Fonction de similarité : la fonction générique $\mathrm{F_s}(\cdot,\cdot)$ prend en entrée deux valeurs $m_{p,ij}$ et $m_{p,kl}$ de $\mathbf{M}_p$ et retourne la similarité entre ces valeurs avec $\mathrm{F_s}(m_{p,ij}, m_{p,kl}) \in [0,1]$; en outre, on pose l'hypothèse que $\mathrm{F_s}(m_{p,ij}, m_{p,kl}) = 0$ lorsque $m_{p,ij} = 0$ ou $m_{p,kl} = 0$.

3.2 L'algorithme de co-similarité χ-Sim

L'algorithme χ-Sim ne travaillant que sur une seule matrice de données, l'ensemble $\mathbf{M}_p$ se réduit à un seul élément que nous noterons $\mathbf{M}$, nous pouvons dans ce paragraphe simplifier les notations précédentes pour supprimer les indices p. Pour rappeler simplement le principe de cet algorithme de calcul de co-similarité χ-Sim (Bisson et Hussain, 2008), nous considérons que la matrice $\mathbf{M}$ est une matrice de co-occurrence *films-acteurs*, et où nous cherchons à calculer la matrice de similarité entre *films* $\mathbf{SR}$.

Dans le calcul de similarité entre les *films* $\mathbf{m}_i$ et $\mathbf{m}_j$, le principe fondamental de l'algorithme χ-Sim est de ne pas seulement considérer les *acteurs* jouant à la fois dans $\mathbf{m}_i$ et dans $\mathbf{m}_j$, mais bien l'ensemble de toutes les paires d'*acteurs* qui apparaissent dans ces *films*. De cette manière, l'algorithme capture non seulement les similarités des *acteurs* partagés entre $\mathbf{m}_i$ et $\mathbf{m}_j$, mais également les similarités des acteurs qui ne sont pas simultanément présents dans les deux *films*. La similarité entre deux *films* $\mathbf{m}_i$ et $\mathbf{m}_j$ est alors calculée par une fonction que nous noterons ici $\mathrm{Sim}(\mathbf{m}_i, \mathbf{m}_j)$ tel que :

$$\begin{aligned}
&\mathrm{Sim}(\mathbf{m}_i, \mathbf{m}_j) = \\
&\quad \mathrm{F_s}(m_{i1}, m_{j1}) \times sc_{11} + \mathrm{F_s}(m_{i1}, m_{j2}) \times sc_{12} + \ldots + \mathrm{F_s}(m_{i1}, m_{jc}) \times sc_{1c} + \\
&\quad \mathrm{F_s}(m_{i2}, m_{j1}) \times sc_{21} + \mathrm{F_s}(m_{i2}, m_{j2}) \times sc_{22} + \ldots + \mathrm{F_s}(m_{i2}, m_{jc}) \times sc_{2c} + \\
&\qquad \ldots \\
&\quad \mathrm{F_s}(m_{ic}, m_{j1}) \times sc_{c1} + \mathrm{F_s}(m_{ic}, m_{j2}) \times sc_{c2} + \ldots + \mathrm{F_s}(m_{ic}, m_{jc}) \times sc_{cc}
\end{aligned} \tag{1a}$$

Symétriquement, la similarité entre deux *acteurs* $\mathbf{m^i}$ et $\mathbf{m^j}$ est donnée par :

$$\begin{aligned}
&\mathrm{Sim}(\mathbf{m}^i, \mathbf{m}^j) = \\
&\quad \mathrm{F_s}(m_{i1}, m_{j1}) \times sr_{11} + \mathrm{F_s}(m_{i1}, m_{j2}) \times sr_{12} + \ldots + \mathrm{F_s}(m_{i1}, m_{jr}) \times sr_{1r} + \\
&\quad \mathrm{F_s}(m_{i2}, m_{j1}) \times sr_{21} + \mathrm{F_s}(m_{i2}, m_{j2}) \times sr_{22} + \ldots + \mathrm{F_s}(m_{i2}, m_{jr}) \times sr_{2r} + \\
&\qquad \ldots \\
&\quad \mathrm{F_s}(m_{ir}, m_{j1}) \times sr_{r1} + \mathrm{F_s}(m_{ir}, m_{j2}) \times sr_{r2} + \ldots + \mathrm{F_s}(m_{ir}, m_{jr}) \times sr_{rr}
\end{aligned} \tag{1b}$$

Cependant, les valeurs des éléments sr_{ij} de $\mathbf{SR}$ et sc_{ij} de $\mathbf{SC}$ doivent appartenir à l'intervalle $[0,1]$, or les valeurs calculées à partir des équations (1a) et (1b) ne vérifient pas cette propriété, ce qui implique une étape de normalisation. Par définition dans la partie 3.1, la fonction de similarité $\mathrm{F_s}$ renvoie des valeurs comprises entre 0 et 1, ainsi, la valeur maximale de

$\mathrm{Sim}(\mathbf{m}_i, \mathbf{m}_j)$ correspond au produit du nombre d'éléments non nuls de $\mathbf{m}_i$ et de $\mathbf{m}_j$. En notant ce produit $|\mathbf{m}_i| \times |\mathbf{m}_j|$, et en suivant le même raisonnement pour les éléments de $\mathbf{SC}$, les valeurs sr_{ij} de $\mathbf{SR}$ et sc_{ij} de $\mathbf{SC}$ peuvent se calculer ainsi :

$$\forall i,j \in 1..r, sr_{ij} = \frac{\mathrm{Sim}(\mathbf{m}_i, \mathbf{m}_j)}{|\mathbf{m}_i| \times |\mathbf{m}_j|} \qquad \forall i,j \in 1..c, sc_{ij} = \frac{\mathrm{Sim}(\mathbf{m}^i, \mathbf{m}^j)}{|\mathbf{m}^i| \times |\mathbf{m}^j|} \tag{2}$$

Lorsque la fonction de similarité $\mathrm{F_s}(m_{ij}, m_{kl})$ est définie comme le produit des deux valeurs m_{ij} et m_{kl}, le système d'équations permettant de calculer $\mathbf{SR}$ et $\mathbf{SC}$ peut être reformulé comme le produit de trois matrices. Ainsi, $\mathbf{SR}$ se calcule de cette manière [2] :

$$\mathbf{SR} = \left(\mathbf{M} \times \mathbf{SC} \times \mathbf{M}^T\right) \circ \mathbf{NR} \quad \text{avec } nr_{ij} = \frac{1}{|\mathbf{m}_i| \times |\mathbf{m}_j|} \tag{3a}$$

Symétriquement, $\mathbf{SC}$ se calcule ainsi :

$$\mathbf{SC} = \left(\mathbf{M}^T \times \mathbf{SR} \times \mathbf{M}\right) \circ \mathbf{NC} \quad \text{avec } nc_{ij} = \frac{1}{|\mathbf{m}^i| \times |\mathbf{m}^j|} \tag{3b}$$

Comme les équations décrites dans (2) ne sont pas indépendantes, il faut utiliser une méthode itérative pour calculer $\mathbf{SR}$ et $\mathbf{SC}$. Les deux matrices de similarité sont initialisées avec la matrice identité $\mathbf{I}$, soit $\mathbf{SR}^{(0)} = \mathbf{I}_r$ et $\mathbf{SC}^{(0)} = \mathbf{I}_c$. En effet, on considère qu'au départ, sans aucune connaissance a priori sur les données, la similarité entre un objet et lui-même vaut 1, et la similarité entre deux objets distincts est nulle. Puis, on calcule tour-à-tour $\mathbf{SR}$ en utilisant la matrice de similarité $\mathbf{SC}$ de l'itération précédente, et $\mathbf{SC}$ en utilisant la matrice de similarité $\mathbf{SR}$ de l'itération précédente. À chaque étape, les diagonales de $\mathbf{SR}$ et $\mathbf{SC}$ sont forcées à 1 afin de contraindre la similarité entre un objet et lui-même au maximum.

Finalement, il est intéressant de s'interroger sur le nombre d'itérations nécessaires pour calculer les valeurs finales de similarité. Dans cet algorithme, la notion d'itération a une forte signification car itérer n fois correspond à prendre en compte le $n^{\text{ème}}$ degré de co-occurrence entre les *films* et les *acteurs*.

3.3 Extensions aux données multi-relationnelles

On cherche maintenant à utiliser l'algorithme précédent permettant de calculer les co-similarités de deux types d'objets, pour calculer des co-similarités de différents types d'objets liés par des relations liant un type d'objets à un autre. Pour plus de simplicité, nous considérons par la suite le cas où nous cherchons à classifier des films selon leur genre (action, comédie...), connaissant pour chaque film, la liste des acteurs y apparaissant et la liste des mots-clés affectés par des utilisateurs. Nous possédons donc deux matrices de co-occurrence, une décrivant la relation *films - acteurs* notée $\mathbf{M}_1$ et une décrivant la relation *films - mots-clés* notée $\mathbf{M}_2$. L'objectif est donc d'examiner comment les informations contenues dans les matrices $\mathbf{M}_1$ et $\mathbf{M}_2$ peuvent être utilisées de manière conjointe et de vérifier que l'on obtient ainsi une classification des films, qui permette de retrouver les genres.

Dans les différentes approches, les « instances » de l'algorithme χ-Sim travailleront chacune avec une matrice de données distincte. On peut ainsi considérer que chaque « instance » de

2. L'opérateur « $\circ$ » désigne le produit de Hadamard, défini tel que, pour $\mathbf{C} = \mathbf{A} \circ \mathbf{B}$, les éléments de $\mathbf{C}$ sont calculés ainsi : $c_{ij} = a_{ij} \times b_{ij}$.

χ-Sim, travaillant avec sa propre matrice de données, est une *vue* différente du problème, et que l'objectif commun de toutes ces approches et de faire communiquer ces *vues* afin de tirer partie de toutes les informations disponibles.

3.3.1 Structure en cascade

Le principe de cette méthode est de calculer une première matrice de similarité des films à partir d'une des deux matrices de co-occurrence, puis de l'utiliser pour initialiser un second algorithme utilisant l'autre matrice de co-occurrence. On effectue n_1 itérations avec la première « instance » de χ-Sim, puis on initialise $\mathbf{SR}_2^{(0)}$ avec $\mathbf{SR}_1^{(n_1)}$, et on effectue n_2 itérations avec la seconde « instance ». Finalement, on utilise $\mathbf{SR}_2^{(n_2)}$ pour classifier les films.

Le schéma présenté sur la figure 1 présente un exemple de structure en cascade utilisant 2 « instances » fe χ-Sim.

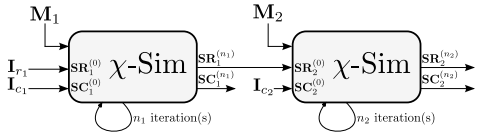

FIG. 1 – *Schéma de la méthode de structure en cascade. La matrice notée* $\mathbf{I}_{r_1}$ *représente la matrice identité de taille* r_1

Ainsi, lorsque la seconde «instance» de χ-Sim est utilisée, elle bénéficie des connaissances sur la similarité entre les objets partagés (ici les films), provenant de la première *vue*, qui doivent permettre d'améliorer les similarités calculées. On peut aisément généralisé cette structure à plus de deux « instances » de l'algorithme χ-Sim.

3.3.2 Structure en anneau

Le principe de cette méthode, utilisant une structure en anneau est de réaliser une seule itération avec chacune des matrices de données puis d'utiliser la matrice de similarité sur les lignes obtenue pour initialiser l'algorithme χ-Sim utilisant une autre matrice de données.

On alterne ainsi entre deux «vues» sur les données, ce qui peut permettre de faire circuler les informations de similarité de l'une à l'autre.

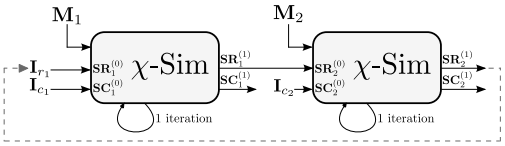

FIG. 2 – *Schéma de la méthode de structure en anneau. La matrice notée* $\mathbf{I}_{r_1}$ *représente la matrice identité de taille* r_1

Le schéma présenté sur la figure 2 représente le principe de cette méthode. À l'instar de la figure 1, on y voit deux «instances» de l'algotithme χ-Sim, la première utilisant la matrice de données $\mathbf{M}_1$, et la seconde utilisant $\mathbf{M}_2$. Chaque calcul de co-similarité à partir de l'une des matrices de données est donc utilisé par l'autre instance de χ-Sim par la suite.

3.3.3 Combinaison

Le principe de cette méthode est de calculer séparément deux matrices de similarité du même type d'objet, ici les films, puis de «combiner» ces deux matrices avant d'effectuer la classification des films. Différentes stratégies pour combiner ces matrices de similarité sont envisagées, parmi lesquelles, la moyenne, le minimum et le maximum des valeurs qu'elles contiennent.

Ainsi, en combinant les matrices de similarité obtenues, on unifie les différentes *vues* suivant un critère déterminé par la stratégie de combinaison (moyenne, minimum, maximum, etc.) choisie. Par exemple, en choisissant la stratégie du minimum, on considère que pour que deux films soient similaires, il faut qu'ils soient similaires dans les deux *vues* ; alors que pour la stratégie du maximum, il suffit qu'ils soient similaires dans au moins une des deux *vues*.

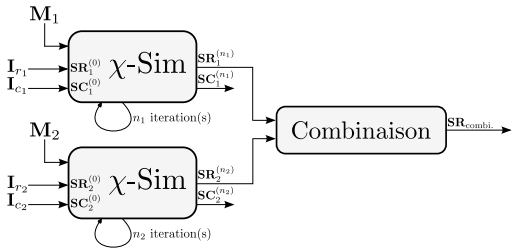

FIG. 3 – *Schéma de la méthode de combinaison. La matrice notée* $\mathbf{I}_{r_1}$ *représente la matrice identité de taille* r_1

Le schéma présenté sur la figure 3 représente le principe de cette méthode. On y voit deux «instances» de l'algorithme χ-Sim, la première utilisant ma matrice de données $\mathbf{M}_1$, et la seconde utilisant $\mathbf{M}_2$. De la stratégie de combinaison, représentée ici par le bloc $\mathbf{combi}$., dépendent donc les résultats de cette méthode.

4 Expérimentations et résultats

L'évaluation de nos algorithmes a été réalisée à partir d'une série d'expérimentations sur un jeu de données où les *films* sont déjà catégorisés. Le critère d'évaluation consiste alors à mesurer l'adéquation entre les classes de films engendrées par un algorithme de classification utilisant la matrice de similarité $\mathbf{SR}$ produite par notre algorithme, et les catégories existantes. Nous évaluons donc ici l'apport de la co-similarité, et des données multi-relationnelles dans la classification des films par genre.

4.1 Base expérimentale

Les travaux dans le domaine de la classification de données multi-relationnelles étant encore peu nombreux, des jeux de données formés d'une collection de matrices sont peu fréquents. C'est pourquoi nous avons choisi de créer nos propres données, décrivant des films.

Nous avons utilisé *IMDb* [3] afin de recueillir le genre principal de chaque film [4] ainsi que la liste des acteurs y apparaissant et la liste des mots-clés affectés par des utilisateurs. Ainsi, nous avons collecté trois types d'objets différents : films, mots-clés et acteurs ; avec les deux relations suivantes : films - mots-clés et films - acteurs. Ceci nous donne donc deux matrices de co-occurrence, ayant la dimension commune *films*. De plus, il est important de noter que seule la qualité de la classification des films est évaluable, l'objectif étant de classer ensemble les films du même genre.

Les films choisis parmi la base de données de IMDb correspondent à un sous-ensemble de ceux utilisés dans le jeux de données *MovieLens* [5]. Nous avons cependant veillé à ce que les films retenus se répartissent de façon équilibrée dans les différents genres. Ainsi, les films de notre jeu de données ne représentent pas tous les genres d'IMDb mais les genres représentés possèdent tous un nombre similaires de films.

Par ailleurs, nous avons effectué un pré-traitement en supprimant les mots-clés utilisés pour moins de 3 films, les acteurs apparaissant dans moins de 2 films, et les films étant étiquetés par moins de 2 mots-clés ou ayant moins de 2 acteurs.

Finalement, notre jeux de données possèdent :

– 617 films de 17 genres différents (environ 36 films par genre)
– 1878 mots-clés
– 1398 acteurs

Chaque film est étiqueté par 33 mots-clés en moyenne, et une moyenne de 9 acteurs est à son casting. Un mot-clé est en moyenne utilisé pour étiqueter 11 films et un acteur apparait dans 4 films en moyenne.

4.2 Algorithmes utilisés

Nous avons comparé nos algorithmes aux mesures de similarité suivantes :

– *La similarité de cosinus.* Bien que cette mesure ne soit pas liée à la notion de co-classification, nous l'avons retenue car elle est intensivement utilisée en recherche d'information. Elle permet donc de définir une « ligne de base » pour les résultats et de mieux estimer l'apport d'une analyse conjointe des films et des acteurs/mots-clés, même si, comme dans cette expérimentation, on ne s'intéresse qu'à la seule classification des films.
– *La mesure de similarité induite par LSA (Latent Semantic Analysis).* Développée par Deerwester et al. (1990), elle est très souvent utilisée en recherche d'information entre autres. LSA permet d'établir des ressemblances entre objets ne partageant pas explicitement des occurrences avec d'autres objets du même type Kontostathis et Pottenger (2004). Il est donc intéressant de comparer une classification basée sur cette approche avec nos méthodes d'autant plus que les mécanismes sous-jacents sont différents : LSA est basée sur la décomposition en valeurs singulières de la matrice de donnée $\mathbf{M}$.
– *L'algorithme de co-classification ITCC (Information Theoretic Co-Clustering)* de Dhillon et al. (2003).

3. http://www.imdb.com/interfaces/
4. Dans IMDb, chaque film se voit attribuer plusieurs genres, mais celles-ci sont cependant hiérarchisées. Nous avons donc décidé de garder le premier comme genre.
5. http://www.grouplens.org/node/73

– *La co-similarité χ-Sim.* Cette mesure est quand à elle liée à la notion de co-classification et elle va nous permettre de démontrer l'apport des matrices de co-occurrence supplémentaires.

Il est évidemment nécessaire pour construire des classes de documents d'utiliser conjointement à ces mesure un algorithme de classification. Dans tous les cas, nous avons retenu l'algorithme de Classification Ascendante Hiérarchique (CAH) avec l'indice de Ward (1963), car, comme l'ont montré Bisson et Hussain (2008), cette méthode est celle qui permet d'obtenir les meilleurs résultats.

4.3 Implémentation et critère d'évaluation

Pour effectuer les tests, χ-Sim ainsi que nos algorithmes ont été implémentés en Java, en utilisant la bibliothèque JAMA [6] pour faciliter la manipulation des matrices. De même, les classifications ascendantes hiérarchiques ont été implémentées en Java.

De nombreuses mesures ont été proposées pour quantifier la ressemblance entre les classes initiales et les classes produites par un algorithme de classification. Elles sont basées sur l'analyse de la matrice de confusion $\mathbf{C}$ dans laquelle chaque élément c_{ij} indique le nombre de films de la classe réelle i qui ont été affectés à la classe apprise j (idéalement on devrait obtenir une matrice diagonale). Pour évaluer les résultats et permettre la comparaison avec les autres travaux nous avons utilisé l'indice classique suivant : la « précision micro-moyennée » ou micro-averaged Precision (Pr) introduite par Dhillon et al. (2003), pour laquelle une valeur de 1 correspond à une classification idéale.

4.4 Résultats expérimentaux

Pour tous les algorithmes disponibles, le test a porté sur l'intégralité de la base, il n'y a donc pas de valeur moyenne, ni d'écart-type disponible.

Nous avons utilisé deux matrices de données pour tester nos méthodes de calcul de mesure de co-similarité :

– $\mathbf{M_1}$: matrice de co-occurrence *films - mots-clés*
– $\mathbf{M_2}$: matrice de co-occurrence *films - acteurs*

De plus, dans ce cas particulier où les deux matrices ont la dimension *films* (les lignes) en commun, il est possible de juxtaposer ces deux matrices. On peut ensuite y appliquer les méthodes de bi-classification classiques. Bien que non générale, nous expérimentons cette stratégie et présentons les résultats dans la dernière ligne du tableau 1.

Pour LSA, nous avons fait varier le nombre k de valeurs singulières conservées dans l'intervalle $[10 \dots 200]$ avec un pas de 10 afin de trouver le nombre optimal de dimensions. La valeur rapportée pour chaque matrice de données est celle correspondant à la représentation des données et la valeur de k qui maximise la précision micro-moyennée.

Pour ITCC, nous avons fait varier pour chaque ensemble de tests le nombre de classes de mots dans la plage de valeur suggérée par Dhillon et al. (2003). Par ailleurs, comme ITCC repose sur une initialisation aléatoire des classes, chaque problème a été soumis 3 fois. Comme pour LSA les résultats rapportés correspondent aux paramètres qui maximisent la valeur moyenne de Pr.

6. http://math.nist.gov/javanumerics/jama/

Les tableaux 1, 2 et 3 présentent respectivement les résultats obtenus avec les méthodes classiques de mesures de co-similarité, les méthodes utilisant la structure en cascade et en anneau. Les résultats sur la méthode de combinaison sont omis pour des raisons de place.

Méthode	Cosinus	LSA	ITCC	χ-Sim	
				2 itération(s)	3 itération(s)
$\mathbf{M}_1$	0,225	0,277	0,280	0,256	**0,282**
$\mathbf{M}_2$	0,212	0,216	0,160	0,214	**0,217**
$\mathbf{M}_1\mathbf{M}_2$	0,284	**0,295**	0,266	0,261	0,280

TAB. 1 – *Résultats des expérimentations en terme de précision micro-moyennée (Pr) pour les tests issus du jeux de données décrit en 4.1 utilisant les méthodes classiques.*

	$\mathbf{M}_1 \rightarrow \mathbf{M}_2$	$\mathbf{M}_2 \rightarrow \mathbf{M}_1$
1 itération(s) $\rightarrow$ 1 itération(s)	0,207	0,254
1 itération(s) $\rightarrow$ 2 itération(s)	**0,227**	0,241
1 itération(s) $\rightarrow$ 3 itération(s)	0,222	0,285
2 itération(s) $\rightarrow$ 1 itération(s)	0,216	**0,292**
2 itération(s) $\rightarrow$ 2 itération(s)	0,227	0,246
2 itération(s) $\rightarrow$ 3 itération(s)	0,225	0,285

TAB. 2 – *Résultats des expérimentations en terme de précision micro-moyennée (Pr) pour les tests issus du jeux de données décrit en 4.1 utilisant la structure en cascade. Dans la première colonne,* 1 *itération(s)* $\rightarrow$ 2 *itération(s) signifie que l'on effectue 1 itération avec la première matrice, puis 2 itérations avec la seconde.*

Le tableau 2 regroupe l'ensemble des résultats permettant de mettre en évidence l'intérêt de la structure de cascade, il apparaît que les résultats obtenus en utilisant l'algorithme χ-Sim sur une seconde matrice de données sont en général meilleurs en utilisant la matrice de similarité obtenue à partir d'une première matrice de données, qu'avec la matrice identité. L'amélioration n'est cependant pas systématique et il serait intéressant de pouvoir déterminer à l'avance, l'apport d'une matrice de similarité donnée.

Notons également que le meilleur résultat que nous ayons obtenu à l'aide de cette structure en cascade donne une précision micro-moyennée de **0,292** avec deux itérations pour la première instance de χ-Sim et une itération pour la seconde instance de χ-Sim.

Le tableau 3 regroupe l'ensemble des résultats obtenus à l'aide de la structure en anneau.

	$\mathbf{M}_1 \leftrightarrow \mathbf{M}_2$	$\mathbf{M}_2 \leftrightarrow \mathbf{M}_1$
3 itération(s)	**0,292**	0,224
4 itération(s)	0,219	**0,266**

TAB. 3 – *Résultats des expérimentations en terme de précision micro-moyennée (Pr) pour les tests issus du jeux de données décrit en 4.1 utilisant la structure en anneau.* $\mathbf{M}_1 \leftrightarrow \mathbf{M}_2$ *signifie que l'on commence par faire une itération avec* $\mathbf{M}_1$*, puis une avec* $\mathbf{M}_2$*, jusqu'a ce que le nombre total d'itérations figurés dans la première colonne du tableau soit atteint.*

Il apparaît que la structure en anneau permet d'obtenir de très bons résultats, mais, à l'instar de la structure en cascade, l'amélioration de ces résultats n'est pas systématique. Notons le meilleur résultat de **0,292** – obtenu avec la structure en cascade et la structure en anneau – qui reste cependant inférieur à celui obtenu par LSA sur la juxtaposition des deux matrices avec **0,295**.

5 Conclusion

Dans cet article, nous avons proposé différentes méthodes permettant de tirer partie des informations fournies par des jeux de données multi-relationnels, impliquant chacune deux types d'objets différents. Nos méthodes s'appuient toutes sur l'algorithme de calcul de co-similarité χ-Sim développée par Bisson et Hussain (2008) très efficace dans des applications de recherche d'information (bi-classification de documents et des termes qui les composent), et dans le domaine de la bio-informatique (bi-classification de gènes et de leurs expressions).

Les différentes « instances » de l'algorithme χ-Sim que nos méthodes utilisent, correspondent à plusieurs vues sur les données dont nous disposons, l'objectif est alors d'agréger ces vues afin de tirer profit au mieux, des données qu'elles apportent. Bien entendu, si les différentes matrices de données apportent des informations contradictoires sur les types d'objets que l'on cherche à classifier, nos propositions ne permettront pas d'obtenir de meilleurs résultats qu'en utilisant une seule matrice de données.

Nous souhaitons continuer à expérimenter nos propositions sur des données représentées par plus de deux matrices. Nous aimerions également tester nos algorithmes sur des jeux de données concernant d'autres domaines, tels que la biologie ou la recherche documentaire.

Références

Banerjee, A., S. Basu, et S. Merugu (2007). Multi-way clustering on relation graphs. In *SDM*. SIAM.

Bekkerman, R., R. El-Yaniv, et A. McCallum (2005). Multi-way distributional clustering via pairwise interactions. In *ICML '05 : Proceedings of the 22nd international conference on Machine learning*, New York, NY, USA, pp. 41–48. ACM.

Bisson, G. et F. Hussain (2008). Chi-sim : A new similarity measure for the co-clustering task. In *Seventh International Conference on Machine Learning and Applications (ICMLA)*, pp. 211–217. IEEE Computer Society.

Cheng, Y. et G. M. Church (2000). Biclustering of expression data. In *Proceedings of the International Conference on Intelligent System for Molecular Biology*, Boston, pp. 93–103.

Deerwester, S., S. T. Dumais, G. W. Furnas, Thomas, et R. Harshman (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science 41*, 391–407.

Dhillon, I. S. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. In *KDD '01 : Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, pp. 269–274. ACM.

Dhillon, I. S., S. Mallela, et D. S. Modha (2003). Information-theoretic co-clustering. In *Proceedings of The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2003)*, pp. 89–98.

Kontostathis, A. et W. M. Pottenger (2004). A framework for understanding latent semantic indexing (lsi) performance.

Liu, X., Y. Gong, W. Xu, et S. Zhu (2002). Document clustering with cluster refinement and model. In *In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 191–198. ACM Press.

Long, B., X. Wu, Z. M. Zhang, et P. S. Yu (2006). Unsupervised learning on k-partite graphs. In *KDD '06 : Proceedings of the 12th ACM SIGKDD*, New York, NY, USA, pp. 317–326. ACM.

Long, B., Z. M. Zhang, et P. S. Yu (2005). Co-clustering by block value decomposition. In *KDD '05 : Proceedings of the eleventh ACM SIGKDD*, New York, NY, USA, pp. 635–640. ACM.

Madeira, S. C. et A. L. Oliveira (2004). Biclustering algorithms for biological data analysis : A survey.

Nadif, M. et G. Govaert (2005). Block clustering of contingency table and mixture model. *Lecture notes in computer science 3646*, 249.

Rege, M., M. Dong, et F. Fotouhi (2008). Bipartite isoperimetric graph partitioning for data co-clustering. *Data Min. Knowl. Discov. 16*(3), 276–312.

Speer, N., C. Spieth, et A. Zell (2004). A memetic clustering algorithm for the functional partition of genes based on the gene ontology.

Tang, W., Z. Lu, et I. Dhillon (2009). Clustering with Multiple Graphs.

Wang, X., J.-T. Sun, Z. Chen, et C. Zhai (2006). Latent semantic analysis for multiple-type interrelated data objects. In *SIGIR '06 : Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, pp. 236–243. ACM.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association 58*(301), 236–244.

Summary

Simultaneous clustering of two types of objects – words and documents for instance – also known as *co-clustering*, has been thoroughly studied in the past years, and allows discovering the structure, whether explicit or latent, of objects. However, many examples of real-world data involve objects of multiple types that are related to each other, and this kind of approaches is unable to take advantage of such data, available in various contexts, such as social networks. This article presents an ongoing work on extending the (co-)similarity measure χ-Sim developped by Bisson et Hussain (2008), in order co-cluster multi-type interrelated data objects.

Recommendation of Key Messages Extracted from Forums

Anna Stavrianou, Julien Velcin, Jean-Hugues Chauchat

ERIC Laboratoire - Université Lumière Lyon 2
Université de Lyon
5 avenue Pierre Mendès-France 69676 Bron Cedex, France

Abstract. The abundance and popularity of online discussions in the Web require their modeling for mining purposes. Most of the existing techniques model a forum as a social network of users. In this paper we consider the structure and the opinion flow of a discussion and we use Post-Reply Opinion Graphs for its representation. We propose using the information extracted from such graphs in order to automatically classify the discussion postings according to how key they are for a specific discussion. We experiment with various criteria and we distinguish between key and non-key messages. Such a distinction enables recommending to a user the most interesting discussion messages. In this way, a user can quickly get an idea of the content of the discussion and identify how to participate or to whom to talk to in the first place. The experiments carried out with forums found on the Web reveal interesting observations.

1 Introduction

The analysis of behaviors and social relations developed between the participants of online discussions has lately received great attention. Currently, most online discussions are modeled by a social network in the form of user-based graphs where the vertices represent the users of the network. User-based graphs capture the relations between the discussion participants and enable the study of user interactions and network dynamics.

One difference between online discussions that usually come in the form of web forums, and standard social network interactions is the relationship between the participants. In such discussions, the participants do not necessarily know each other and they do not necessarily intend to develop a relationship between them. Moreover, the importance of forums lies in their content, since the posted messages may include opinions and criticism on certain ideas, product reviews, general knowledge and belief exchange. The current social network models do not allow focusing on the opinionated content of a forum. This shows that another type of modeling is required that not only takes advantage of the social network models but it also exploits the content and offers additional knowledge regarding the discussion.

In this paper, we represent an online discussion from the point of view of postings rather than just the users who participate in the discussion. We use a Post-Reply Opinion Graph which allows us to define criteria so as to determine the most interesting messages that appear inside an online discussion. The recommendation to an end-user of a list of discussion messages would help the user navigate quicker and more efficiently inside the discussion.

Our main objective is to distinguish between the key and the non-key messages that are posted in an online discussion. Most end-users assume a message to be interesting if it follows at least one of the following assumptions which were collected after having interviewed 10 end-users who visit forums in an almost daily basis. The assumptions are presented here in descending order of popularity: a)Opinion: A message is interesting when it contains opinion, b)Size: A message is interesting when it is present inside a long discussion thread, c)Reactions: An interesting message has caused many reactions, d)Initial: The initial message of a discussion thread is interesting, e)Time: The most recent message is interesting.

The extraction of key messages permits a kind of summarization of the discussion and it allows the user to get an idea of the content and the main ideas that have been expressed. Considering the fact that most important discussions can be very long, with hundreds and thousands of messages, our model helps the user to identify quickly how to participate or to whom to start talking to, since there is no need to spend time reading the whole discussion.

The contributions of our work are summarized in the following:

– We use Post-Reply Opinion Graphs to extract key messages from online discussions.
– A number of criteria are studied and they are correlated with user preferences.
– Extensive experiments are carried out which allow us to see the conditions under which a recommendation set is acceptable by a user.

This paper begins by discussing related work in Section 2. It continues with the presentation of the Post-Reply Opinion Graph in Section 3. In Section 4 we present our criteria that are based on the theory of graphs and in Section 5 we carry out experiments with web forums. Section 6 concludes by highlighting future perspectives.

2 Related Work

The recommender systems are systems whose main goal is to recommend items to users. These items can range from movies and news articles to proposals for holiday destinations or for meeting people.

One example of a recommender system is the Syskill & Webert system presented in (Pazzani and Billsus, 1997). The purpose of this system is to recommend interesting web pages on a particular topic. The system learns user profiles that differ per topic. It identifies the most informative words of each web page by calculating the expected information gain that the presence or absence of a word gives towards the classification of elements of a set of pages. The user can rate the recommended pages as positive or negative examples and these examples are used when a profile is learnt.

Among the first collaborative recommender systems is Tapestry (Goldberg et al., 1992) that recommends interesting e-mails and news to the user. The system relies on direct as well as implicit user feedback. For example, a mail that is replied to by an interesting user is an interesting mail since it has been given attention to by this user. Tapestry is filtering items into two steps; the first step includes separating the items in «good» or «bad» and the second step is prioritizing the «good» items.

Another work which is outside the recommender systems domain but it is highly related with ours, is a quality document identification system (Elkan, 2007). This work classifies mainly text documents with an application of ranking messages in discussion groups according to quality. The classifier is based on various criteria such as the vocabulary used, the length

of words/sentences and the usage of grammar. The training data are collected with the help of humans who determine which types of documents/messages are of high quality. This work, though very interesting, differs from ours in that we do not use a learning system. We are based on the structure of the forum and its graph representation rather than on the low-level features of a text (e.g. average length of words, usage of punctuation, orthography etc.).

3 Post-Reply Opinion Graph

The web discussions in the form of forums are characterized by a number of postings sent by users who have registered to the forum with a username. The postings are either a reply to an already existing message or they are sent as a new, independent message that signals the beginning of a new discussion thread. We consider the relations «reply-to» between the messages of a forum to be known.

The model we use for the representation of web forums is based on graphs. Most graph-based existing approaches consider users to be the vertices of the graph. In our model, we propose to use message objects as the vertices. More information about the model can be found in (Stavrianou et al., 2010).

Definition. A **Post-Reply Opinion Graph** is a directed graph $G = (V, E)$ with a vertex set V and an edge set E. Each vertex represents a *posting* and each edge $e_{v'v} = (v', v)$ points out a reply direction from the vertex v' to the vertex v. A vertex $v \in V$ is defined as:

$$v = (m_v, op_v, u_v, tm_v),$$

where m_v is the actual content of the posting, op_v the opinion polarity included in the message, u_v the user that has written it, and tm_v the timestamp that shows when the posting was posted.

The opinion $op_v \in \{n, o, p\}$ in the definition of a Post-Reply Opinion Graph captures the opinion expressed in the message m_v. It may be negative (n) or positive (p) or we may have no opinion included (o). The opinion can be calculated by Opinion Mining techniques (e.g. (Hu and Liu, 2004), (Turney, 2002)).

The author of the message u_v is encapsulated in the message object. As a result, the social network of users can be extracted from the proposed model. This is an important property of the Post-Reply Opinion Graphs, since the information provided by the social networks can still be exploited.

Representing a forum as a graph of related message objects allows us to identify discussion threads. The set of the **discussion threads** in a Post-Reply Opinion Graph G is the union of all the maximal connected components of G. In Figure 1, we can see an example of a Post-Reply Opinion Graph which is composed of three discussion threads.

The proposed model differs from the existing representations that focus on user-based graphs in many ways. It allows having a content and opinion-oriented view of the discussion. In this way, we can identify easier the messages that have initiated a thread and the messages that have caused many reactions. We can also observe the overall opinion flow of a discussion i.e. if it is consisted mainly of positive or negative statements. Moreover, we can extract information regarding the sentiment behavior of users and towards certain users. This information is not straightforward from the existing discussion models.

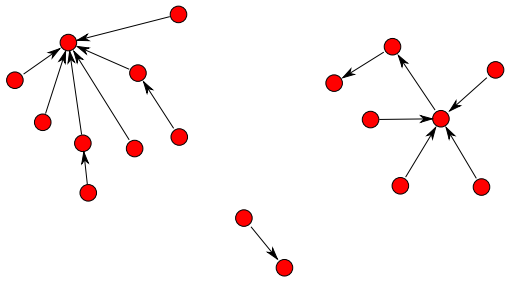

FIG. 1 – *A Post-Reply Opinion Graph of a forum.*

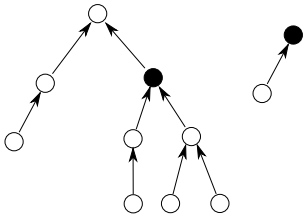

FIG. 2 – *The importance of a vertex differs according to the order of the discussion thread it belongs to. The black vertex on the left hand side has played a more significant role in the discussion than the black vertex on the right hand side.*

In this paper we will use the Post-Reply Opinion Graph in order to extract key messages from an online discussion.

4 Extracting Key Messages

Choosing the criteria for providing key messages to users, can be considered as the «cold-start» case of a recommender system. This is the case when a new user enters the system and the system knows nothing about this user. It does not know any of the user's likes/dislikes and as a result, it cannot match the user with already known user profiles. In such cases, a system has to deal with very little data.

Based on the aforementioned assumptions of which type of message an average user considers to be key, we present the criteria used in order to extract key messages from a forum that is represented by a Post-Reply Opinion Graph G.

4.1 Order of a discussion thread

A posting that belongs to a discussion thread with many vertices does not have the same importance as a posting that belongs to a thread with one or two vertices. In Figure 2, for example, the black vertex on the left-hand side of the Figure may have influenced a part of the discussion, while the black vertex on the right-hand side of the Figure does not have the same weight.

Each discussion thread consists of a number of vertices. The order of the discussion thread G_{thr} where the vertex v belongs to is defined as:

$$orderThr(G_v) =\mid V_{thr} \mid,$$

where $G_v = (V_{thr}, E_{thr})$ is the subgraph of G that represents the discussion thread G_{thr} which contains the vertex v.

We define a criterion that retrieves a set of messages which belong to a thread with a minimum order, defined by the threshold d. A vertex v of a graph $G = (V, E)$ will be part of the set of these messages if it is retrieved by the following criterion:

$$order(G, d) = \{v \in V : orderThr(G_v) > d\}, \quad (1)$$

where G_v is the discussion thread of the graph G that contains the vertex v, and d is a threshold that can be either calculated automatically or defined by the user. In the rest of the paper we refer to this criterion as «order».

4.2 Root Vertices

A discussion is often divided into sub-topics, since discussion participants tend to discuss about a particular topic or elaborate on a specific argument. A user that wants to speak about an argument/topic that has not been referred to until then, may initialize a new discussion thread by sending a new message that is not a reply to an existing posting. This would be a «root» message.

We define a criterion that retrieves a set of messages which are root and they also belong to a thread with a minimum order, defined by the threshold d:

$$root(G, d) = \{v \in V : outDegree(v) = 0, orderThr(G_v) > d\}, \quad (2)$$

where G_v is the discussion thread that contains the vertex v and d is a user-defined or automatically calculated threshold that points out the minimum desired number of vertices of the discussion thread. In the rest of the paper we refer to this criterion as «root».

4.3 Vertex Popularity

According to the theory of graphs, the $inDegree$ of a vertex of the graph $G = (V, E)$ denotes its predecessors:

$$inDegree(v) =\mid \{v' \in V : (v', v) \in E\} \mid.$$

In the case of a Post-Reply Opinion Graph, the predecessors of a vertex point out the number of reply-postings it has received. The more the replies, the more attention the message represented by the vertex has been paid to.

A vertex v of a graph $G = (V, E)$ can be considered to be popular if it satisfies the following criterion:

$$popular(G, d) = \{v \in V : inDegree(v) > d\}, \quad (3)$$

where $d > 0$ is a threshold that can be either calculated automatically or defined by the user. This criterion will be referred to as «popularity».

4.4 Opinion content

One advantage of a Post-Reply Opinion Graph is that it encapsulates the opinion information of each posting. A user that scans a forum with many postings is mainly looking for messages that contain opinion or messages where there is a presence of arguments from the participants. As a result, a message that contains opinion is more probable to be a key message than a message that is just informative (subjective message).

The opinion messages are represented by vertices v = (m_v, op_v, u_v, tm_v), where $op_v \neq o$. This criterion will be referred to as simply «op. content».

4.5 Opinion Reactions

Another criterion we define that uses the opinion information of a Post-Reply Opinion Graph is the number of reactions which contain opinion.

The $\sum_{r\in\{n,p\}}(reply(v,r))$ is the number of replies that hold an opinion. The $reply(v,p)$ and $reply(v,n)$ denote the number of replies expressing a positive and a negative opinion respectively.

This criterion is an indication of whether a posting has caused reactions that contain opinion or just information and it is 0 only if all reactions are opinion-free.

A vertex v of a graph $G = (V, E)$ can be considered to satisfy the criterion of opinion reactions if it belongs to the set:

$$reply(G,d) = \{v \in V : \sum_{r\in\{n,p\}} (reply(v,r)) > d\}, \tag{4}$$

where $d > 0$ is a threshold that can be either calculated automatically or defined by the user. The criterion will be referred to as «Op. Reactions».

4.6 Entropy

This criterion is similar to the previous one that counts the reactions with opinion, but it is not a counter. The entropy H of a node v measures the variety in the opinion of the replies a message has received and it is defined as:

$$H(v) = -\sum_{r\in\{n,o,p\}} \left(\frac{reply(v,r)}{inDegree(v)} \log \frac{reply(v,r)}{inDegree(v)} \right).$$

According to this definition, if a vertex has received a number of replies that are all of the same opinion orientation, then the entropy will be 0. This criterion captures phenomena where a particular posting has caused disagreement or replies with various opinion degrees of arguments. Such a posting can be interesting for a discussion analyst in order to investigate the reasons why people argue.

A vertex v of a graph $G = (V, E)$ can be considered to satisfy the criterion of entropy if it belons to the set:

$$entropy(G,d) = \{v \in V : H(v) > d\}, \tag{5}$$

where $d > 0$ is a threshold that can be either calculated automatically or defined by the user. This criterion will be referred to as «entropy».

5 Experimental Evaluation

The evaluation of a system that extracts key messages from discussions is not straightforward (Herlocker et al., 2004; Shardanand and Maes, 1995). The evaluation we perform here is done in two parts: a)We initially asked the experts to label the messages that they consider to be key and then we compared the labeled messages with the messages the system found as key and b)we showed the messages identified as key by our system to the experts and the experts told us whether they agree or not.

The first part of the evaluation helps us to analyze how the different criteria or a combination of criteria can be applied to real online discussions in order to help us distinguish and extract interesting messages. In addition, it helps us identifying whether these criteria correlate with how the humans proceed in classifying the messages as key ones or not. The second part of the evaluation confirms that the criteria used are actually approved by the users when it comes to selecting key messages.

The initial experiments that have been carried out involve French forums from the http://www.liberation.fr/forums/ web site. In these discussions the users identify themselves by user names. The «reply-to» links between the messages are known. We have analyzed a total of 1,147 messages that appear in eight forums. The messages were manually annotated with opinion polarities, since the automatic Opinion Mining is considered out of scope for the particular evaluation. We applied each of the pre-mentioned criteria to the set of the messages of each forum, and we identified the key messages per forum and per criterion.

In order to evaluate the results given by each criterion, we asked human raters to classify the messages as being key or not for the flow of the discussion. For each forum, two experts identified the key messages. The experts were free to choose an unlimited number of key messages per discussion.

5.1 Choosing the right threshold

One of the decisions that has to be taken when evaluating each criterion is the threshold d to be used. The threshold makes sense for the criteria where there is some granularity. For criteria such as the «op. content», where the answer is binary, all thresholds give the same results.

Let us consider a particular discussion represented by the graph $G = (V, E)$, for which the maximum value of a criterion C is M, and the value for the respective criterion for a vertex v is $criterion_C(v)$. Let also the threshold be t.

Definition. The posting represented by the vertex $v \in V$ can be added in the list of recommended postings if, for the criterion C and a predefined threshold t, we have $\frac{criterion_C(v)}{M} \geq t$.

We experimented with various thresholds $t \in (0, 1]$, in order to study the difference in our results. For each criterion we tried to choose the threshold that minimizes the number of messages retrieved while giving good performance results.

We gave importance to the values of the F1-measure so as to decide which threshold is good for each criterion. For the criteria «Order» and «Popularity» the maximum value for the F1-measure was achieved for a 0.2 threshold. For the criterion «Root» we chose not to apply

a threshold so that even the nodes which result to discussion threads with an order equal to 1 are considered. Independent of the threshold was also the criterion «Op. Content».

For the criterion «Op. Reactions», the maximum value for the F1-measure is achieved for a 0.1 threshold. Although this threshold results in the extraction of many messages, it is leveraged by the fact that this criterion is hard to be satisfied and, thus, there are not many messages that comply with it. Finally, for the criterion «Entropy» the maximum value for the F1-measure is achieved for a 0.6 threshold.

5.2 Evaluation of each criterion separately

We looked at the standard recall, precision and F1 measures in order to evaluate the correlation between the results of each criterion and those of the human raters. For each criterion we used the thresholds mentioned in the previous Section. The average results of the Recall, Precision and F1-measure per criterion are presented in Table 1.

	Rec.	Prec.	F1
Order	**0.71**	0.14	0.22
Root	0.64	0.20	0.26
Popularity	0.44	0.27	0.29
Op. Content	0.59	0.22	0.28
Op. Reactions	0.31	0.36	**0.30**
Entropy	0.20	**0.53**	0.27

TAB. 1 – *Average Recall, Precision and F1-measure results per criterion when the optimum threshold value per criterion is used.*

From Table 1, the results indicate that the criterion «Order» achieves the highest average recall and the lowest average precision of all the criteria. This can be attributed to the fact that important messages tend to be involved in long discussion threads, since they cause reactions and sub-discussions. The low precision, though, is because not all messages of long threads are considered to be key. Similarly, the results show that when the root and the popularity criterion is applied, the recall is much higher than the precision. On the other hand, when we choose the messages that have had reactions which contain opinions or those whose reactions include a variety of opinion polarities, we achieve higher precision than recall. This is due to the fact that the forum messages that satisfy these criteria are usually very few.

From Table 1, we notice that the precision is on average low. The low precision shows that whatever criteria we use we will retrieve non-interesting messages among the interesting ones. This is inevitable because a message that is considered key for one user may be skipped by another one and vice versa. Also, when an expert classifies the forum messages, the fact of reading the message A before the message B, affects his/her decision on which message is a key message. Even if, under certain circumstances, they would be both regarded as key, the expert tends to select the one that appeared first. This concept is mentioned in (Robertson, 1977) regarding the Probability Ranking Principle. According to this, the fact that a document A is retrieved before a document B may affect the usefulness of A.

The low F1-measure results of each criterion separately show the need of aggregating the different criteria in order to achieve a more satisfactory coverage of the users' needs.

5.3 Aggregation of Criteria

There are many ways in which multiple criteria can be aggregated (Adomavicius and Tuzhilin, 2005). For this paper, we choose to do a simple linear aggregation, considering that each weight equals to 1. The linear aggregation gives ranked results. Therefore, the precision will have more sense if we calculate the precision at a cut-off value n of the ranking (Salton and Lesk, 1968). In our case, n will be the number of the user's answers.

We give the results of $precision@n$ and $recall$ per forum and per expert in Table 2. We consider n to be the number of messages the user has said to be the interesting ones and we consider a message to be «correctly assigned» by our aggregated measure if it is ranked lower than the 50% of the total messages of the forum. If, for example, a forum has 100 messages, then we consider as «correctly assigned» only the first 50 ranked.

Forum	Expert	Recall	Precision@n	F1
1	1	0.88	0.53	0.66
	2	0.86	0.43	0.57
2	1	0.72	0.28	0.40
	2	0.88	0.25	0.39
3	1	1.00	0.33	0.50
	2	0.73	0.4	0.52
4	1	0.8	0.4	0.53
	2	0.86	0.29	0.43
5	1	0.64	0.07	0.13
	2	0.68	0.25	0.37
6	1	1.00	0.67	0.80
	2	0.76	0.59	0.66
7	1	0.86	0.45	0.59
	2	0.7	0.35	0.47
8	1	1.00	0	0
	2	0.78	0.56	0.65
Average		**0.82**	**0.37**	**0.48**

TAB. 2 – *Recall, Precision@n and F1-measure results for the linear aggregation of criteria.*

From the Table 2 we see that the average F1-measure increases quite a lot when compared to the F1-measure of the criteria applied separately. This increase shows that by having a simple linear aggregation of the proposed criteria we have much better results than having one criterion every time. In the future, experimenting with different weights (other than 1) may optimize the aggregation results.

5.4 Additional Experiments

In this Section, we present the second part of the experiments we have carried out in order to see whether the messages that we extract are useful when they are recommended to the users. We performed the experiment with 6 users, 8 French forums and 7 English forums. We evaluated a total of 35 answers (almost 6 forums per user). For each evaluation we started by

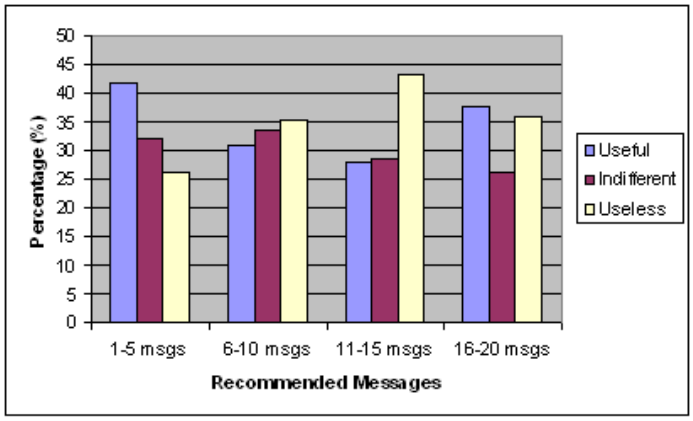

FIG. 3 – *Results of the evaluation of a recommended set of messages.*

presenting the forum to the user by giving its title and a list of the 20 first extracted messages given in random order. The 20 messages are the ones we extract when we aggregate the presented criteria. We asked the users to read each of the recommended messages separately and rate them according to how useful they are in helping the user to start navigating inside it. The rating varied between «useful», «indifferent» and «useless».

The outcome of this experiment is shown in Figure 3 where we see the percentage of «useful», «indifferent» and «useless» to the users messages in the set of the first five recommended messages, the set of the second five recommended messages and so on. This experiment showed that within the first 5 extracted messages we find the maximum of «useful» and the minimum of «useless» postings. At the same time when we recommend more than 10 messages to the user, we include a lot of useless ones.

During our experiment, we noticed that some short messages may have been popular but they made no sense to the user when they were presented on their own. As a result we decided to carry out another experiment after having removed these short messages from the set. This improved the results of the evaluation as it is shown in Figure 4, by increasing the rate of useful messages by an average of 6% and reducing the rate of useless messages by an average of 7%. When the short messages are excluded, the number of useless messages lowers even when we go up to 20 recommended messages, while the number of useful messages remains high.

We also noticed that there are differences in how users rate each posting. For example, a «useful» posting for one user was sometimes rated as «useless» by another one. This shows that there is a need for personalization techniques which will improve even more our results and will lead to a more appropriate recommendation per user. This agrees with the comments we have had from the users during the experiment which are summarized in the following:

– If a posting is in favor of an idea for which the user is totally against, then the posting is more likely to be rated as «useless».
– A posting that contains information which the user already knows about may be rated as «useless» or «indifferent».

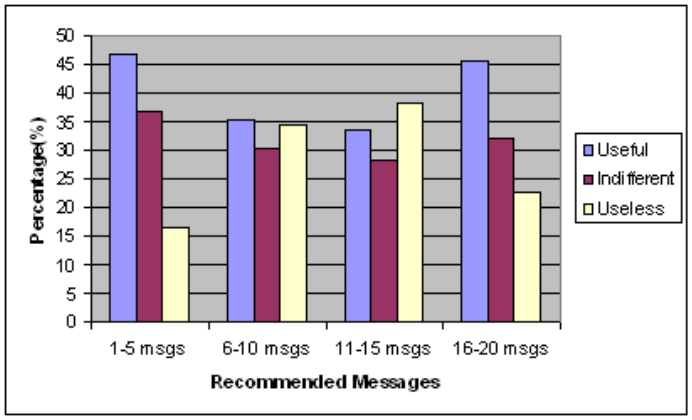

FIG. 4 – *Results of the evaluation of a recommended set of messages when the short messages are excluded.*

– If the user's mother language is not that of the language of the posting, the user may rate the posting as «useless» even if it is not.
– For some users the style of a posting plays a role in what they consider as «useful». For example, postings that contain slang are attractive to some users but not to others.

The results presented in the tables and figures of this Section are the total results of all the 15 forums together. The results can differ significantly according to the forum. This is a general problem in the evaluation of recommender systems. Even if the system does its best, if none of the available postings are interesting to the user anyway, the performance will be low.

6 Conclusion and Future Work

We have used a Post-Reply Opinion Graph in order to classify the discussion postings according to how key they are. These key messages can be recommended to users in order to help them browse through an online discussion without having to read all the existing postings.

We have experimented with real users and forums in both English and French. Various criteria have been applied and an aggregation of these criteria seems to improve the results of the recommendation task. Recommending messages to a user is quite a subjective issue and as a result personalization techniques (Eirinaki and Vazirgiannis, 2005; Mobasher, 2007) should be applied in the future in order to make more appropriate per user recommendations.

Apart from adding personalization techniques to our approach, we could consider a PageRank-style criterion in order to choose the key messages. Two messages with the same values for all criteria can be differentiated by the importance of the messages they reply to or they are replied to. This is an interesting issue that needs further research.

References

Adomavicius, G. and A. Tuzhilin (2005). Towards the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering 17*(6), 734–749.

Eirinaki, M. and M. Vazirgiannis (2005). Usage-based pagerank for web personalization. In *Proceedings of 5th IEEE International Conference on Data Mining (ICDM)*.

Elkan, C. (2007). Method and system for selecting documents by measuring document quality. *US Patent 7200606*.

Goldberg, D., D. Nichols, B. Oki, and D. Terry (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM 35*(12), 61–70.

Herlocker, J., J. Konstan, L.G.Terveen, and J.T.Riedl (2004). Evaluating collaborative filtering recommender systems. *ACM TOIS*.

Hu, M. and B. Liu (2004). Mining and summarizing customer reviews. In *KDD '04*, pp. 168–177. ACM.

Mobasher, B. (2007). Data mining for personalization. *Adaptive Web: Methods and Strategies of Web Personalization, LNCS 4321*, 90–135.

Pazzani, M. and D. Billsus (1997). Learning and revising user profiles: The identification of interesting web sites. *Machine Learning 27*, 313–331.

Robertson, S. (1977). The probability ranking principle in ir. *Journal of Documentation 33*.

Salton, G. and M. Lesk (1968). Computer evaluation of indexing and text processing. *Journal of the ACM 15*(1), 8–36.

Shardanand, U. and P. Maes (1995). Social information filtering: Algorithms for automating "word of mouth". In *CHI-95*.

Stavrianou, A., J. Velcin, and J.-H. Chauchat (2010). *PROG: A Complementary Model to the Social Networks for Mining Forums*. Social Networks Analysis and Mining, Springer.

Turney, P. (2002). Thumbs up or down? semantic orientation applied to unsupervised classification of reviews. In *ACL-2002*, pp. 417–424.

Résumé

La plupart des techniques existantes modélisent un forum comme un réseau social d'utilisateurs. Dans cet article, nous considérons la structure et le déroulement de l'opinion dans la discussion et nous utilisons le Post-Reply Opinion Graphs pour sa représentation. Nous proposons d'utiliser l'information extraite de ce type de graphes afin de classifier automatiquement les messages des discussions selon qu'ils soient ou non clés pour une discussion spécifique. Nous expérimentons selon plusieurs critères et nous distinguons les messages clés de ceux qui ne le sont pas. Cette distinction permet de recommander à un utilisateur les messages les plus intéressants de la discussion. De cette manière, l'utilisateur peut avoir rapidement une idée du contenu de la discussion, identifier comment participer ainsi que trouver l'interlocuteur privilégié. Les expériences faites sur des forums du Web révèlent des observations intéressantes.

Un modèle de Recherche d'Information Sociale pour l'Accès aux Ressources Bibliographiques : Vers un réseau social pondéré

Lamjed Ben Jabeur*, Lynda Tamine* et Mohand Boughanem*

*IRIT, Université Paul Sabatier
118 Route de Narbonne
F-31062 Toulouse CEDEX 9 , FRANCE
{jabeur, tamine, bougha}@irit.fr

Résumé. Cet article propose une nouvelle approche, basée sur les réseaux sociaux, pour l'accès aux ressources bibliographiques. Nous introduisons un modèle d'information sociale dont les auteurs sont les principales entités et les relations sont extraites à partir des liens de coauteur et de citation. En effet, ces relations sont pondérées en tenant compte des interactions entre les auteurs et des annotations sociales produites par les utilisateurs. Dans ce modèle, la pertinence d'un document est estimée par combinaison de la pertinence thématique et de la pertinence sociale, qui est à son tour dérivée de l'importance sociale des auteurs associés. Nous évaluons la viabilité de notre modèle sur une collection d'articles scientifiques dont les annotation sociales sont extraites depuis le réseau social académique *CiteULike.org*. Les résultats obtenus montrent la supériorité des performances de notre modèle par rapport à la recherche d'information traditionnelle.

1 Introduction

Les moteurs de recherche académiques, tels que GOOGLE SCHOLAR[1] et CITESEERX[2], ont donné aux chercheurs la possibilité d'accéder à diverses sources d'information scientifique et de gérer leurs références bibliographiques. Depuis leur apparition, les systèmes de recherche d'information scientifique ont été confrontés à une problématique majeure qui consiste à évaluer l'importance des publications scientifiques. Les premiers travaux y ont apporté une solution en utilisant les indicateurs bibliométriques. Certains travaux successeurs ont modélisé les ressources bibliographiques avec des structures hypertextes dont les hyperliens représentent les citations. Dans une telle approche, l'importance des documents est calculée en appliquant l'algorithme de *PageRank* sur le graphe de citation.

[1] http ://scholar.google.fr/
[2] http ://citeseerx.ist.psu.edu/

Avec l'apparition des réseaux sociaux académiques tels que CITEULIKE[3] et ACADEMIA[4], cette vision a été élargie en considérant que les articles scientifiques sont produits et consommés par des entités sociales et leur importance peut être estimée à partir du contexte de production et d'utilisation. Cette approche a été portée par le courant de la recherche d'information sociale où les acteurs sont représentés au moyen d'un réseau social et la pertinence du document est calculée en appliquant les mesures de centralité sociale introduites par l'analyse des réseaux sociaux Wasserman et al. (1994). En effet, la recherche d'information sociale suppose qu'un document pertinent est produit par un auteur important d'où l'importance scientifique des ressources bibliographiques peut être dérivée à partir de l'importance sociale des auteurs associés.

Dans ce contexte, inspirés des travaux de Mutschke (2001) et Kirsch et al. (2006) qui s'intéressent à l'accès aux ressources bibliographiques et à la représentation des auteurs par un réseau social, nous proposons un modèle générique de la recherche d'information sociale qui est déployé particulièrement pour l'accès aux ressources bibliographiques. Comparativement aux approches précédentes, ce modèle comprend des nouvelles entités sociales représentées par les annotateurs et les annotations sociales, des nouvelles relations sociales telles que la citation et l'annotation et des mesures de pondération attribuées aux différentes relations du réseau.

Le reste de cet article est organisé comme suit. Dans la section 2, nous présentons une synthèse des approches proposées pour l'accès aux ressources bibliographiques. La section 3 donne un aperçu de notre modèle générique de recherche d'information sociale. La section 4 décrit l'instanciation du modèle générique dans le cadre précis de l'accès aux ressources bibliographiques. La section 5 décrit les expérimentations menées pour valider notre modèle. Enfin, la section 6 conclut le papier et annonce les perspectives.

2 Recherche d'information dans les ressources bibliographiques

Les travaux précurseurs dans le domaine de la recherche d'information dans les ressources bibliographique ont considéré les liens de citation comme étant un indicateur de qualité et d'autorité des publications scientifiques. Nous citons dans ce contexte les indicateurs bibliométriques basés sur le nombre de citations tel que le facteur d'impact mesurant l'importance d'une publication ou d'une revue scientifique Garfield (2006). Cependant, l'indice de citation est insuffisant pour estimer la pertinence d'un document. Des nouvelles mesures qui tiennent compte de la croissance de citations reçues sont alors proposées pour ordonner les documents selon leurs ages et le nombre de citation prévues Hauff et Azzopardi (2005) Meij et de Rijke (2007). D'autre part, certaines approches représentent les citations par des hyperliens et modélisent les ressources bibliographiques sous forme d'un graphe. Dans ce cas, les documents sont classés selon leurs autorités calculées par les algorithmes de la recherche d'information hypertexte tels que *PageRank* et *HITS* Page et al. (1998) Langville et Meyer (2005).

Certains travaux récents réutilisent les mesures de centralité introduites par le domaine d'analyses des réseaux sociaux telles que les mesures de *Betweeness* et de *Closeness* afin

[3]http ://www.citeulike.org/
[4]http ://www.academia.edu/

d'identifier les ressources centrales dans le graphe des documents Bollen et al. (2005) ou pour déduire la pertinence d'un document à travers la centralité de ses auteurs appliquant ainsi ces mesures sur le graphe des auteurs Mutschke (2001) Kirchhoff et al. (2008). Pour représenter le réseau social, la plupart des travaux se limitent à des simples liens de coauteur entre le nœuds du graphe Yan et Ding (2009) Mutschke (2001). Toutefois, les modèles présentés par Newman (2000) et Liu et al. (2005) proposent de pondérer les liens entre les auteurs selon la fréquence et l'exclusivité de leurs associations de coauteur.

D'autre modèles de recherche d'information sociale intègrent les documents comme des nœuds du réseau social Korfiatis et al. (2006). Dans ce cas, les relations entre les documents et les auteurs sont extraites à partir des interactions de collaboration, de publication et de citation. Des telles approches interprètent la pertinence par le degré de confiance, d'autorité et de popularité des documents dans le réseau social Kazai et Milic-Frayling (2008). Ces facteurs de pertinence sociale sont modélisés soit par des probabilités de transition sur le graphe social (approche intégrée) Amer-Yahia et al. (2007), soit par un score combiné (approche modulaire) Kirchhoff et al. (2008) Kirsch et al. (2006).

Inscrit dans ce cadre, nous proposons un modèle de recherche d'information sociale qui associe la pertinence des ressources bibliographiques à l'importance sociale de leurs auteurs. Comparativement aux travaux proches du domaine et de notre précédente contribution Tamine et al. (2009), notre proposition présentée dans ce papier s'en distingue par les points clé suivants :

1. nous proposons une formalisation d'un modèle générique de recherche d'information sociale,
2. en plus des liens de coauteur, nous exploitons deux autres types de relations sociales : la citation et l'annotation sociale,
3. nous attribuons à ces relations des poids qui tiennent compte de la position des acteurs dans le réseau social et de leurs mutuelles collaborations.

3 Le modèle générique de recherche d'information sociale

Un modèle de recherche d'information offre un support théorique pour représenter des documents et des requêtes et mesure leur degré de similitude assimilée à la pertinence. Formellement, et en se basant sur la représentation proposée par Baeza-Yates et Ribeiro-Neto (1999), nous décrivons le modèle générique de recherche d'information sociale par un quintuplet $[D, Q, G, F, R(q_i, d_j, G)]$ où D est l'ensemble des documents, Q est l'ensemble des requêtes, G est le réseau d'information sociale, F est la fonction d'appariement des documents et des requêtes et $R(q_i, d_j, G)$ est une fonction de classement qui intègre divers facteurs de la pertinence sociale et qui tient compte de la topologie du réseau. Cette fonction peut être définie par la combinaison de sous-ensemble des facteurs de la pertinence sociale suivants : la pertinence thématique, l'importance sociale des acteurs, la distance sociale, la popularité, la fraîcheur de l'information et le nombre de marque-pages reçus Amer-Yahia et al. (2007).

En ce qui concerne le réseau d'information sociale G, il représente les entités sociales qui interagissent au voisinage du document. Nous proposons donc d'y inclure tous les acteurs et les données qui permettent d'évaluer sa pertinence sociale comme illustré dans la figure 1. Les acteurs y représentent les producteurs et les consommateurs d'information (respectivement les

auteurs et les utilisateurs) tandis que les données comprennent les documents et les annotations sociales (les *tags*, les votes, et les avis). Dans le cadre de leurs collaborations et interactions sociales, les acteurs participent à produire de l'information et à enrichir les documents par les annotations sociales.

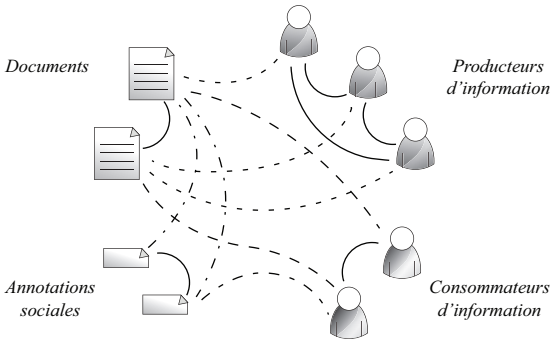

FIG. 1 – *Le réseau d'information sociale*

Le réseau d'information sociale peut être représenté par un graphe $G = (V, E)$ où l'ensemble des nœuds $V = A \cup U \cup D \cup T$ représente les entités sociales avec A, U, D et T correspondent respectivement aux auteurs, aux utilisateurs, aux documents et aux annotations sociales. L'ensemble des arcs $E \subseteq (V \times V)$ représente les relations sociales reliant les différents types des nœuds (publier, co-auteur, amitié, citation, annotation...etc.).

Le réseau d'information sociale est appréhendé différemment. Du point de vue du producteur d'information, le réseau social regroupe les nœuds documents et auteurs et met en évidence le contexte social de la production des ressources. De même, la vue consommateurs d'information représente le contexte d'utilisation sociale des documents et l'interaction entre les utilisateurs. Cette vue intègre 3 types des nœuds à savoir : les documents, les annotations sociales et les utilisateurs.

Dans la suite, nous instancions ce modèle générique au cadre de la production et de la consommation de ressources bibliographiques.

4 Le réseau d'information sociale des ressources bibliographiques

Les travaux du domaine modélisent essentiellement le réseau social des ressources bibliographiques en se basant uniquement sur le point de vue producteurs d'information. Cependant, l'introduction des réseaux sociaux académiques sur le web (par exemple CITEULIKE et ACADMEICA) a permis aux utilisateurs de participer et de fournir par la suite des descripteurs sociaux aux ressources bibliographiques. Contrairement aux réseaux des amis tels que FACEBOOK[5] et MYSPACE[6] et dont les relations entre les individus expriment principalement un lien

[5]http ://www.facebook.com/
[6]http ://www.myspace.com/

d'amitié, les réseaux sociaux académiques incluent des relations sociales spécifiques entre les entités d'information.

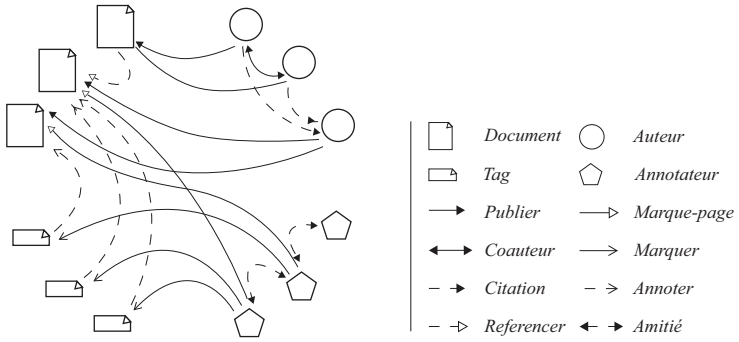

FIG. 2 – *Le réseau d'information sociale pour les ressources bibliographiques*

D'un point de vue des producteurs d'information, nous identifions les relations sociales suivantes qui impliquent les documents et les auteurs :

Publier : une relation dirigée relie chaque auteur $a_i \in A$ avec sa publication $d_j \in D$.

Référencer : une relation dirigée qui associe un document $d_i \in D$ avec ses références bibliographiques,

Coauteur : une relation entre deux auteurs $a_i, a_j \in D$ ayant collaboré pour produire un document en commun.

Citation : une relation sociale dirigée entre un auteur $a_i \in A$ et un second auteur $a_j \in A$ avec a_j est cité dans l'un des documents écrit par a_i.

Du point de vue des consommateurs d'information, nous identifions les relations sociales suivantes qui intègrent les documents, les *tags* et les annotateurs :

Marque-page : en attribuant un *tag* à un document, l'annotateur $u_i \in U$ et le document $d_j \in D$ sont associés avec une relation sociale de marque-page.

Annoter : relie un *tag* $t_j \in T$ avec un document $d_i \in D$ assigné au moins une fois pour décrire son contenu.

Marquer : relie un annotateur $u_i \in U$ et un tag $t_j \in T$ utilisé au moins une fois pour marquer un document.

Amitié : relie deux annotateurs $u_i, u_j \in U$. Cette relation peut être décrite explicitement par un lien d'amitié ou implicitement par le fait d'appartenir à un même groupe.

Le réseau d'information sociale pour les ressources bibliographiques peut être représenté en utilisant une notation graphique illustrée dans la figure 2.

4.1 Pondération des relations sociales

Les arcs reliant les nœuds sociaux expriment divers types de relations sociales et permettent d'optimiser d'une façon significative le processus d'exploration du réseau social. Si nous ex-

plorons le voisinage social d'un nœud[7], les poids sur les arcs nous permettent de sélectionner le nœud suivant à chaque saut. Dans cet article, nous nous intéressons essentiellement au réseau social des publications scientifiques, pour cela nous définissons un modèle de pondération pour les associations auteur-auteur $e(a_i, a_j) \in (A \times A)$ et les associations auteur-document $e(a_i, d_j) \in (A \times D)$.

a. La relation de coauteur : représentée par un arc dirigé, cette relation connecte deux coauteurs ayant collaboré pour produire un document. Les coauteurs ont souvent des contacts personnels directs néanmoins la multiplicité de leurs collaborations exprime la similarité et le partage d'intérêt entre eux. En effet, les auteurs des publications scientifiques ont tendance à s'échanger les connaissances et diversifier leurs collaborations. Pour cette raison et afin de quantifier la similarité entre les coauteurs, nous proposons de tenir compte de la totalité des collaborations. Nous proposons d'assigner des poids asymétriques aux relations de coauteur comme suit :

$$Co(i,j) = \frac{A(i,j)}{A(i)} \tag{1}$$

Avec $A(i,j)$ est le nombre de documents co-écrits par les auteurs a_i et a_j. $A(i)$ représente le nombre des documents publiés par l'auteur a_i.

b. La relation de citation : représentée par un arc dirigé, les liens de citation expriment le transfert de connaissances entres les auteurs des publications scientifiques. Par conséquent, plus un auteur cite les publications d'un second auteur, plus il est influencé par ses idées d'une manière que tout les deux partagent des sujets similaires. Cette relation est asymétrique et son importance est souvent proportionnelle au nombre des publications. Afin de mesurer l'importance des cette association, nous tenons compte du nombre des citations entre les auteurs ainsi que le nombre total des citations énoncées par l'auteur source de la relation. Les relations de citation sont alors pondérées comme suit :

$$Ci(i,j) = \frac{C(i,j)}{C(j)} \tag{2}$$

Avec $C(i,j)$ est le nombre de fois que l'auteur a_i a cité l'auteur a_j et $C(i)$ représente le nombre de citations énoncées par l'auteur a_i.

c. La relation de publication : un auteur sera plus affilié à un sujet S s'il l'a fréquemment abordé dans ses publications. Ainsi, un coauteur sera davantage associé à son document d que ses coauteurs s'il a publié plusieurs documents sur le même sujet de d. Pour estimer les connaissances d'un coauteur a_k sur le sujet de son document d, nous proposons de comparer la quantité d'information importée via ses autres publications. Du point de vue des consommateurs, cela peut être estimé par la distribution des *tags* affectés au sous-ensemble des publications de chaque coauteur représenté par $\mathcal{A}_k$ avec $a_k \in A$ est le coauteur que nous souhaitons mesurer l'affiliation au sujet de son document d. Nous calculons ainsi une distribution

[7] Voisinage social : l'ensemble des nœuds de proximité accessibles directement ou indirectement à partir d'un nœud du réseau.

de probabilité de l'ensemble des *tags* T assignés au document d dans la sous-collection des documents publiés par les coauteurs du document d.

$$w(a_k, d) = \sum_{t_i \in T} \frac{tf(t_i, \mathcal{A}_k)}{tf(t_i, \mathcal{A})} \tag{3}$$

Avec T l'ensemble des tags assignés au document d. $\mathcal{A} = \bigcup_{k=1}^{m} \mathcal{A}_k$ représente la sous-collection des documents publiée par les m coauteurs du document d. $tf(t_i, \mathcal{A}_k)$ est la fréquence de *tag* t_i dans le sous-ensemble des documents $\mathcal{A}_k$ publiés par l'auteur a_k. $tf(t_i, \mathcal{A})$ représente la fréquence de *tag* dans la sous-collection des documents publiés par les coauteurs du document d.

Certains algorithmes de centralité sociale ne supportent pas la multiplicité des arcs de même sens entre deux nœuds. Nous proposons donc de combiner les poids des relations de coauteur et de citation comme suit :

$$w(a_i, a_j) = \frac{1}{4} * (1 + Co(i, j)) * (1 + Ci(i, j)) \tag{4}$$

4.2 Estimation de la pertinence sociale des documents

La sémantique de l'importance sociale des documents dépend de la nature de l'application. Par exemple, l'importance sociale d'un article de blog est estimé à travers la popularité et le nombre des *tags* reçus. La mesure de *Degree* est alors la plus appropriée pour calculer son importance. Pour notre part, l'objectif est de sélectionner la mesure d'importance sociale qui met en évidence les ressources bibliographiques de qualité. Ainsi, nous calculons pour chaque auteur a_i un score d'importance sociale $C_G(a_i)$ en utilisant une des mesures d'importance suivantes : *Betweeness*, *Closeness*, *PageRank*, le score *"Authority"* de *HITS* et le score *"Hub"* de *HITS*. Ces mesures sont appliquées seulement sur le sous-graphe d'auteurs $G_a = (A, E_a)$ avec $E_a \subseteq (A \times A)$. Les arcs sont pondérés comme décrit précédemment et désignent soit une relation de coauteur ou un lien de citation. Ensuite, un score d'importance sociale est propagé aux documents par une somme pondérée des scores sociaux des ses auteurs :

$$Imp_G(d) = \sum_{i=1}^{m} w(a_i, d) C_G(a_i) \tag{5}$$

Nous combinons par la suite le score $Imp_G(d)$ avec une métrique de la recherche d'information traditionnelle. L'idée consiste à estimer la pertinence du document dans le graphe social et de présenter une réponse plus précise à l'utilisateur, en combinant la pertinence thématique et l'importance sociale. Intuitivement, un utilisateur est susceptible d'évaluer un document comme pertinent s'il couvre le sujet de la requête et si les auteurs correspondants sont socialement importants. Sur cette base, nous définissons la fonction de classement R par la combinaison linéaire de deux scores normalisés de la pertinence comme suit :

$$R(q, d, G) = \alpha\, RSV(q, d) + (1 - \alpha) Imp_G(d) \tag{6}$$

Avec $\alpha \in [0,1]$ est un paramètre de pondération, $RSV(q,d)$ est une mesure de similarité thématique entre la requête q et un document d et $Imp_G(d)$ est l'importance sociale du document d dans le réseau social G.

5 Évaluation expérimentale

Dans le but d'évaluer l'impact de notre modèle étendu sur l'efficacité de la recherche, nous avons mené une série d'expérimentations sur une collection d'articles scientifiques. Les objectifs de cette évaluation sont de :

- mesurer l'impact de la pondération des relations sociales sur l'estimation de la pertinence sociale des documents comme proposé précédemment,
- comparer les différentes mesures d'importance sociale afin de déterminer la mesure permettant de mieux exprimer l'importance des ressources bibliographiques,
- mener une évaluation comparative de notre modèle relativement à un modèle recherche d'information classique.

5.1 Cadre d'évaluation

Les compagnes d'évaluations tel que TREC proposent un cadre standard pour évaluer et comparer les systèmes de recherche d'information. Cependant, les collections disponibles ne sont pas adaptées pour évaluer les modèles de recherche d'informations sociale en l'absence des données indispensables à la construction du réseau social. Afin de valider notre proposition, nous avons construit un corpus des documents scientifiques issus du la conférence ACM SIGIR de 1978 à 2008. Nous décrivons dans la suite les caractéristiques de la collection des documents ainsi que les requêtes et les mesures d'évaluation utilisées.

- *Corpus des ressources bibliographiques*
  La collection SIGIR comprend 2871 auteurs avec une moyenne de 2 relations de coauteur et 16 liens de citation par auteur. Comme indiqué au tableau 1, les relations de citation dominent le réseau social avec 9 fois plus des liens que les associations de coauteur. En intégrant les relations de citation dans le réseau social, les communautés dispersées et de petite taille se restructure en plus larges composantes connexes. Par conséquent, la composante "géante" reliant la majorité des nœuds est élargie pour inclure 84% des auteurs comme le montre la figure 3.

Nombre de documents	2053
Nombre d'auteurs	2871
Relations de coauteur	5047
Relations de citation	45880
Relations de coauteur et/ou citation	52516
Composant le plus large	2430 (84%)

Tableau. 1 – *Caractéristiques statistiques de corpus SIGIR*

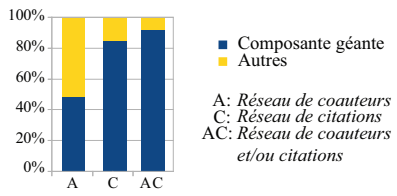

Figure. 3 – *Composante géante du réseau social SIGIR*

En plus du contenu textuel des publications scientifiques, le corpus SIGIR inclut d'autres informations sur les auteurs et les citations. Un auteur fait partie du réseau social s'il a publié au moins un article dans le cadre de la conférence ACM SIGIR. Deux auteurs sont en relation sociale à condition :

- Qu'ils aient publié un article en commun dans la conférence ACM SIGIR.
- Que dans son article SIGIR, le premier auteur cite une publication du deuxième auteur. L'article cité ne doit pas forcément appartenir au corpus SIGIR .

Nous avons enrichi ce corpus avec les données collectées depuis le réseau social académique CITEULIKE. Nous regroupons ainsi tous les *tags* utilisés pour annoter les documents SIGIR ainsi que les identifiants des utilisateurs correspondants. Pour indexer cette collection, nous avons utilisé le moteur de recherche TERRIER[8].

- *Collection de requêtes test et jugements de pertinence associés*
  Étant donné que les *tags* sont des mots clés générés par les utilisateurs dans le but d'annoter les documents et que les requêtes sont une représentation les utilisateurs d'un besoin en information, les *tags* peuvent alors constituer les requêtes dans notre cadre d'évaluation. Nous considérons que les *tags* le plus populaires sont de haute importance sociale et nous sélectionnons comme des requêtes les 25 *tags* le plus utilisés pour annoter les documents SIGIR.
  Pour constituer la collection des documents pertinents, nous supposons qu'un document est pertinent s'il est annoté au moins une seul fois par le *tag* (requête) et que ce dernier est parmi le 3 *tags* les plus affectés au documents. Le collection finale contient 25 requêtes et 223 documents pertinents avec une moyenne de 8.9 documents par requête.
- *Mesures d'évaluation*
  Afin d'étudier les performances de notre modèle et de comparer les mesures d'importance sociale nous étudions les précisions aux $5^{ème}$ et $10^{ème}$ documents résultats que nous notons respectivement $p@5$ et $p@10$. Ces mesureurs évaluent la capacité du systèmes à retourner des résultats pertinents parmi les premiers documents retournés.

5.2 Comparaison des mesures d'importance sociale

Les différentes mesures d'importance sociale mettent en évidence les entités clés d'un réseau social. Ces mesures ont une sémantique qui varie d'une application sociale à une autre. Dans le contexte des publications scientifiques, la mesure de *Betweeness* est considérée comme un indicateur d'interdisciplinarité et met en évidence les auteurs connectant plusieurs partitions dispersées de la communauté scientifique. La mesure de *Closeness*, basée sur les chemins le plus courts, reflète la proximité et l'indépendance d'un auteur à son voisinage social. Les mesures de *PageRank* et le score d' *Authority* de *HITS* distinguent les sources d'autorité dans le réseau social. En revanche, le score de *Hub* de *HITS* identifie les auteurs ayant une importante activité sociale tout en se basant sur des sources d'autorité, appelés les auteurs *"Centraux"*.

Nous avons appliqué les mesures d'importance sociale, à savoir : *Betwenness*, *Closeness*, *PageRank*, *Authority* (*HITS*), et *Hub* (*HITS*) sur un modèle binaire et pondéré du réseau social. Nous notons l'application des ces mesures sur le modèle pondéré du réseau respectivement

[8] http ://www.terrier.org

par *W-Betwenness*, *W-Closeness*, *W-PageRank*, *W-Authority*, et *W-Hub*. Les performances de recherche sont présentées dans le tableau 2.

	p@5	p@10		p@5	p@10
Closness	0,0211	0,0526	W-Closness	0,0316	0,0579
Betweeness	**0,0421**	0,0526	W-Betweeness	0,0316	0,0316
PageRank	0,0211	0,0421	W-PageRank	0,0316	0,0421
Authority	0,0316	0,0368	W-Authority	0,0316	0,0368
Hub	0,0316	**0,0579**	W-Hub	**0,0421**	**0,0632**

TAB. 2 – *Comparaison des mesures d'importance sociale*

En comparant les précisions $p@5$ et $p@10$, nous constatons que la mesure de *Hub* permet de mieux classer les ressources bibliographiques retournées initialement par un modèle de recherche d'information classique. Nous concluons donc que l'importance des publications scientifiques peut être estimée par la *Centralité* de leurs auteurs. En effet, la pertinence sociale dans le contexte de ressources bibliographiques est interprétée par l'intense activité de l'auteur, proportionnelle au nombre de publications et de collaborations, tout en s'appuyant sur des ressources d'autorité.

Pour la plupart des mesures d'importance, le modèle pondéré permet d'améliorer l'efficacité de la recherche. Cela est constaté avec les valeurs des précisions obtenues par les mesures de *W-Closeness*, de *W-PageRank* et de *W-Hub* dépassant leurs analogues appliquées sur un modèle binaire du réseau social. Nous concluons donc que les propriétés exprimées via la pondération des relations sociales, à savoir le partage des centres d'intérêt, l'influence et le transfert de connaissances, permettent de mieux identifier les auteurs *Centraux* et par la suite estimer la pertinence des ressources bibliographiques.

Pour évaluer l'efficacité de notre modèle, nous retenons la mesure de *W-Hub* comme étant la mesure qui permet de mieux exprimer l'importance sociale des ressources bibliographiques.

5.3 Evaluation de l'efficacité de notre modèle

Dans l'étape précédente, les résultats sont ordonnés uniquement selon les scores sociaux des documents. Dans ce cas, les mesures de précision $p@5$ et $p@10$ ne dépassent pas le seuil de 46% comparé à celles du système de recherche d'information classique basé sur le modèle *Okapi BM25* Jones et al. (2000) dont $p@5 = 0,158$ et $p@10 = 0,137$. Les mesures d'importance sociale ne sont donc pas capables de trier les résultats sans prendre en considération la similarité entre le document et la requête. Nous nous intéressons à présent à une combinaison linéaire de deux scores pour estimer la pertinence d'un document à la requête, comme indiqué dans la formule 6.

Nous avons étudié l'impact du paramètre α sur le processus de recherche d'information, et ceci pour la mesure de *W-Hub*. Lorsque $\alpha = 0$, seule la pertinence sociale est prise en considération. D'autre part, $\alpha = 1$ correspond au *baseline* BM25 puisque seule la pertinence thématique est prise en considération pour classer les documents. L'analyse des mesures $p@5$ et $p@10$ en fonction du paramètre α montre que les courbes, présentées sur la figure 4, présentent des pics dont les valeurs dépassent la valeur obtenue pour $\alpha = 1$, et cela lorsque la pertinence thématique est uniquement prise en considération. Donc, la combinaison des deux scores per-

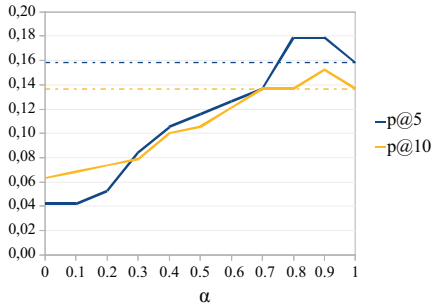

FIG. 4 – *Ajustement du paramètre* α

met effectivement d'améliorer l'ordonnancement final des documents. Les meilleures valeurs du paramètre α sont obtenues entre 0.8 et 0.9.

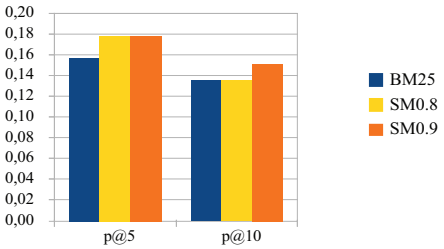

FIG. 5 – *Évaluation l'efficacité de notre modèle*

Nous avons comparé notre modèle avec un système de recherche d'information classique basé sur le modèle BM25 et utilisant l'algorithme de lemmatisation *"SnowBall Stemmer"*. Nous utilisons le même modèle pour sélectionner les documents pertinent et pour calculer le score pertinence thématique $RSV(q, d)$. Comme décrit dans la figure 5, les meilleures valeurs du paramètre α permettent d'aboutir à une amélioration jusqu'à 13% par rapport la *baseline* BM25. Nous concluons donc que la combinaison de la pertinence thématique et l'importance sociale des documents permet d'accéder au ressources bibliographiques de qualité et d'améliorer l'efficacité de la recherche.

En fait, les tags utilisés comme des requêtes dans notre évaluation expérimentale sont des termes utilisateur et ne figurent pas forcément dans le contenu du document. Par conséquent, seuls quelques documents pertinents sont sélectionnés ce qui explique ainsi la faible précision de la *baseline* qui affecte directement performances du modèle proposé.

En outre, les résultats sont comparables au modèle de recherche d'information classique basé sur la pertinence thématique. L'objectif principal des notre proposition est d'améliorer l'efficacité de la recherche en combinant l'importance sociale du document et sa pertinence thématique. Nous avons atteint cet objectif avec un taux d'amélioration de 13% par rapport à la *baseline* BM25.

6 Conclusion

Nous avons proposé un modèle de recherche d'information sociale générique puis nous l'avons instancié pour l'accès aux ressources bibliographiques. Ce modèle a la spécificité d'intégrer les relations sociales de citation, de production et d'annotation ainsi que la pondération des différentes relations sociales. Notre évaluation expérimentale sur la collection des documents scientifiques SIGIR montre que la mesure de *Hub* est la mesure qui permet de mieux évaluer l'importance sociale des documents scientifiques et prouve la supériorité de notre modèle comparativement à un modèle de recherche d'information classique.

En perspective, nous envisageons de comparer les performances de notre approche aux autres modèles qui intègrent la composante sociale. De plus, nous mènerons les expérimentations sur une collection de ressources bibliographiques de plus grande taille.

Références

Amer-Yahia, S., M. Benedikt, et P. Bohannon (2007). Challenges in searching online communities. *IEEE Data Eng. Bull. 30*(2), 23–31.

Baeza-Yates, R. et B. Ribeiro-Neto (1999). *Modern Information Retrieval*. New York : ACM Press.

Bollen, J., H. V. de Sompel, J. A. Smith, et R. Luce (2005). Toward alternative metrics of journal impact : A comparison of download and citation data. *Information Processing & Management 41*(6), 419 – 1440. Special Issue on Infometrics.

Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA 295*(1), 90–93.

Hauff, C. et L. Azzopardi (2005). Age dependent document priors in link structure analysis. In *ECIR*, pp. 552–554.

Jones, K. S., S. Walker, et S. E. Robertson (2000). A probabilistic model of information retrieval : development and comparative experiments. *Inf. Process. Manage. 36*(6), 779–808.

Kazai, G. et N. Milic-Frayling (2008). Trust, authority and popularity in social information retrieval. In *CIKM '08*, New York, NY, USA, pp. 1503–1504. ACM.

Kirchhoff, L., K. Stanoevska-Slabeva, T. Nicolai, et M. Fleck (2008). Using social network analysis to enhance information retrieval systems. In *in Applications of Social Network Analysis (ASNA) (Zurich) , 12-9-2008*.

Kirsch, S. M., M. Gnasa, et A. B. Cremers (2006). Beyond the web : Retrieval in social information spaces. In *In Proceedings of the 28 th European Conference on Information Retrieval (ECIR 2006*. Springer.

Korfiatis, N. T., M. Poulos, et G. Bokos (2006). Evaluating authoritative sources using social networks : an insight from wikipedia. *Online Information Review 30*(3), 252–262.

Langville, A. N. et C. D. Meyer (2005). A survey of eigenvector methods for web information retrieval. *SIAM Rev. 47*(1), 135–161.

Liu, X., J. Bollen, M. L. Nelson, et H. Van de Sompel (2005). Co-authorship networks in the digital library research community. *Inf. Process. Manage. 41*(6), 1462–1480.

Meij, E. et M. de Rijke (2007). Using prior information derived from citations in literature search. In *RIAO*.

Mutschke, P. (2001). Enhancing information retrieval in federated bibliographic data sources using author network based stratagems. In *Reserach and Advanced Technology for Digital Libraries : 5th European Conference, ECDL 2001, Darmstadt, Germany, September 4-9, 2001 ; Proceedings*.

Newman, M. E. J. (2000). Who is the best connected scientist ? a study of scientific coauthorship networks. Working Papers 00-12-064, Santa Fe Institute.

Page, L., S. Brin, R. Motwani, et T. Winograd (1998). The pagerank citation ranking : Bringing order to the web. Technical report, Stanford Digital Library Technologies Project.

Tamine, L., A. B. Jabeur, et W. Bahsoun (2009). An exploratory study on using social information networks for flexible literature access. In *FQAS*, pp. 88–98.

Wasserman, S., K. Faust, et D. Iacobucci (1994). *Social Network Analysis : Methods and Applications (Structural Analysis in the Social Sciences)*. Cambridge University Press.

Yan, E. et Y. Ding (2009). Applying centrality measures to impact analysis : A coauthorship network analysis. *J. Am. Soc. Inf. Sci. Technol. 60*(10), 2107–2118.

Prédiction de Motifs Relationnels par Décomposition Tensorielle dans les Réseaux Sociaux

Sheng Gao Ludovic Denoyer Patrick Gallinari

Université Pierre et Marie Curie - LIP6 - Paris, France
{sheng.gao, ludovic.denoyer, patrick.gallinari}@lip6.fr

Résumé. Beaucoup de données réelles peuvent être représentées sous la forme de collections d'objets liés par différents types de relations. Les réseaux sociaux par exemple correspondent à un ensemble d'éléments - des utilisateurs, des données textuelles, des images ... - connectés par des relations de différentes natures comme les relations d'amitiés, de similarités, ou bien des relations géographiques. Cependant, si certaines de ces relations sont explicites dans le réseau social, comme la relation d'amitié par exemple, certaines peuvent être implicites ou incomplètes. Nous nous intéressons ici à la prédiction automatique de relations entre deux individus d'un réseau social. Nous proposons une formalisation de cette problématique sous forme d'un problème de décomposition de tenseur de dimension 3 et proposons une méthode à base de gradient conjugué pour le résoudre. Nous expérimentons cette méthode sur un corpus issu du jeux de données Enron et montrons l'intéret de la modélisation tensorielle des réseaux sociaux pour la prédiction de relations.

1 Introduction

Beaucoup de données réelles peuvent être représentées sous la forme de collections d'objets liés par différents types de relations. Les réseaux sociaux par exemple correspondent à un ensemble d'éléments - des utilisateurs, des données textuelles, des images ... - connectés par des relations de différentes natures comme les relations d'amitiés, de similarités, ou bien des relations géographiques. Cependant, si certaines de ces relations sont explicites dans le réseau social, comme la relation d'amitié par exemple, certaines sont plus implicites voire incomplètes. Considérons par exemple une relation du type " deux utilisateurs partagent le même centre d'intérêt". Cette relation va être explicitée entre certains utilisateurs, mais non renseignée pour de nombreux autres.

Nous nous intéressons ici à la tâche de la prédiction de relations entre les éléments d'un réseau social. Cette tâche consiste à inférer, à partir des données du réseau et de relations connues, un ensemble de nouvelles relations entre les individus. Les relations peuvent être de plusieurs types. Plus particulièrement, nous considérons le cas d'un réseau multi-relationnel dont les relations entre individus sont incomplètes et cherchons à prédire les relations manquantes entre deux individus. L'ensemble des relations entre deux individus sera dénommé *motif relationnel* dans la suite de l'article. Notre problématique correspond donc à de l'inférence automatique de motifs relationnels dans un réseau social. Cette tâche est illustrée dans la

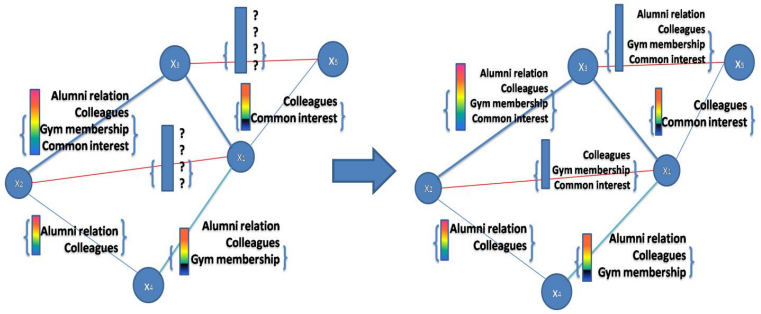

FIG. 1 – *Prédiction de motifs relationnels : la partie gauche représente un réseau social multi-relationnel avec des relations manquantes. En partie droite, les relations manquantes on été prédites par le modèle.*

figure 1. Cette figure (partie gauche) illustre un réseau social composé d'un ensemble d'individus et d'un ensemble de relations. Les relations considérées ici sont au nombre de 4. Cependant, certaines relations entre individus manquent (les " ?" dans la figure). La tâche d'inférence automatique de motifs relationnels (partie droite) consiste à inférer automatiquement de nouvelles relations à partir de l'état connu du réseau.

Cette tâche trouve de nombreuses applications concrètes, aussi bien en tant qu'étape intermédiaire pour des problèmes d'analyse de réseaux sociaux - notamment la détection de communautés ou bien la détection automatique de "rôles" - que pour de la prédiction, permettant alors de déduire les différents types de relations entre individus - par exemple "qui partage les mêmes centre d'intérêts que moi ?".

Nous proposons un modèle de prédiction de motifs relationnels basé sur un algorithme de décomposition d'un tenseur décrivant le réseau social. Si cette problématique a déjà été étudiée dans le cadre de la prédiction mono-relationnelle (voir partie 4), nous nous intéressons ici au problème multi-relationnel. Cet algorithme est basé sur une méthode originale d'optimisation à base de descente de gradient conjugué. La suite de cet article est organisée comme suit : dans la section 2, nous présentons brièvement le formalisme des tenseurs, ainsi que la formulation du problème de prédiction de motifs relationnels. Notre modèle est présenté en partie 2.3. La partie 3 présente des résultats expérimentaux obtenus sur des corpus réels et la section 4 présente les travaux de l'état de l'art connexes à notre approche.

2 Le modèle de prédiction de motif de liens

2.1 Tenseurs

Les tenseurs sont une extension multi-dimensionnelle de la notion de matrice. Un tenseur $\mathcal{X}$ à N dimensions est une fonction à valeurs dans $\mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ telle que $\mathcal{X}_{i_1,..,i_N}$ est la valeur de la cellule de coordonnées $i_1, .., i_N$ où $i_j \in I_j$. Ainsi, un vecteur correspond à un tenseur à 1 dimension, une matrice est un tenseur à 2 dimensions tandis qu'un tenseur à 3 dimensions correspond comme montré en figure 2 à un empilement de matrices. Dans la suite de l'article, nous nous intéressons uniquement à des tenseurs à trois dimensions $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$.

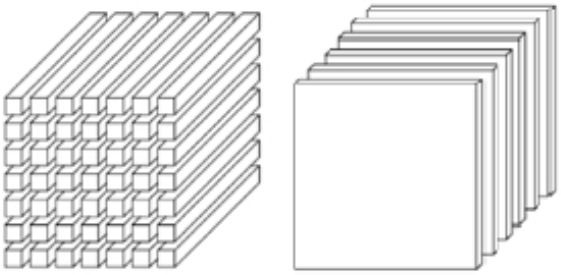

FIG. 2 – *Fibres et tranches d'un tenseur.*

Nous notons $\mathcal{X}_{i,j,k}$ la valeur de la cellule de coordonnées i, j, k. Nous noterons $\mathcal{X}_{i,j,:} \in \mathbb{R}^K$ la "fibre" du tenseur de coordonnées i, j qui correspond au vecteur $(\mathcal{X}_{i,j,1}, \cdots, \mathcal{X}_{i,j,K})$. De même, nous noterons $\mathcal{X}_{:,:,k} \in \mathbb{R}^{I \times J}$ la tranche du tenseur de coordonnées k, c'est à dire la matrice de niveau k dans le tenseur. Ces notions sont illustrées en figure 2.

2.1.1 Décomposition tensorielle

La décomposition tensorielle est une méthode de factorisation matricielle avec des applications importantes en traitement du signal, en statistique ou en analyse de données. Elle permet en effet de faire de la réduction du bruit, de la compression des données, de faire de l'apprentissage non-supervisé, ... Plusieurs méthodes de décomposition existent. Nous utilisarons ici une méthode classique qui est la décomposition CP (CANDECOMP/PARAFAC) (Harshman, 1970) (Carroll et Chang, 1970). Ce modèle considère qu'un tenseur tridimensionnel $\mathcal{X}$ peut s'écrire à l'aide de trois facteurs matriciels A, B et C sous la forme :

$$\mathcal{X}_{i,j,r} = \sum_{k=1}^{K} A_{i,k} B_{j,k} C_{r,k} \tag{1}$$

A, B et C sont des matrices et K correspond au rang du tenseur $\mathcal{X}$, $A_{i,k}$ est la valeur de la matrice A à la coordonnée i, k. Différentes méthodes existent pour trouver cette décomposition. Le calcul du rang K étant problème NP-complet, différentes valeurs de K sont habituellement testées expérimentalement.

2.2 Modélisation de réseaux sociaux par tenseurs

Nous considérons qu'un réseau social est un graphe multi-relationnel $(\mathcal{N}, \mathcal{R})$ définit pour un ensemble de noeuds N et un ensemble de relations multiples $\mathcal{R}$. $\mathcal{N} = (n_1, ...n_N)$ l'ensemble des noeuds ou individus du réseau social. L'ensemble des relations mutiples entre individus sera représenté par un tenseur de dimension 3 noté $\mathcal{R}$ qui correspond à l'empilement des matrices d'adjacences pour chaque type de relations. Soit R le nombre de relations possibles,

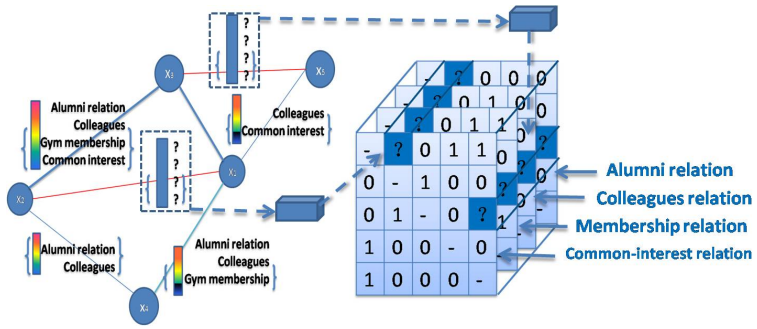

FIG. 3 – *Un réseau social multi-relationnel et sa représentation tensorielle associée.*

$\mathcal{R}$ est un tenseur définit sur l'espaces $N \times N \times R$ à valeur dans $0; 1$ tel que :

$$\forall i, j \in [1..N]^2, \forall r \in [1..R], \mathcal{R}_{i,j,r} = \begin{cases} 1 \text{ si la relation de type r existe entre i et j} \\ 0 \text{ sinon} \end{cases} \tag{2}$$

La fibre $\mathcal{R}_{i,j,:}$ représente l'ensemble des relations entre deux individus i et j. Cette fibre est appelée dans la suite *motif relationnel*. Ces différentes notions sont illustrées en figure 3 (droite). Dans la partie suivante, nous nous intéressons à la tâche de prédiction de motifs relationnels entre individus.

2.3 Prédiction de motifs relationnels

Dans le réseau social $(\mathcal{N}, \mathcal{R})$, nous allons considérer que certains des motifs relationnels entre individus $\mathcal{R}$ sont inconnus. Ces motifs relationnels inconnus sont identifiés à l'aide d'une matrice de poids $\mathcal{W}$ définie dans l'espace $\mathcal{N} \times \mathcal{N}$ tel que :

$$\forall i, j \in [1..N]^2, \mathcal{W}_{i,j} = \begin{cases} 1 \text{ si le motif relationnel entre } i \text{ et } j \text{ est connu} \\ 0 \text{ sinon} \end{cases} \tag{3}$$

La tâche de prédiction de motifs relationnels va consister à tenter de retrouver l'ensemble des motifs relationnels pour toutes les valeurs nulles de la matrice $\mathcal{W}$. Pour cela, nous allons chercher une décomposition CP aussi proche que possible du tenseur original $\mathcal{R}$ pour les relations connues. Ce problème s'écrit comme celui consistant à trouver un tenseur $\mathcal{X}^*$ de même taille que $\mathcal{R}$ qui minimise un coût défini ainsi :

$$L(\mathcal{X}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \mathcal{W}_{i,j} ||\mathcal{R}_{i,j,:} - \mathcal{X}_{i,j,:}||^2 \tag{4}$$

Le problème de prédiction de motifs relationnels peut donc s'écrire comme :

$$\begin{gathered} \text{Trouver } \mathcal{X}^* \text{ tel que } \mathcal{X}^* = argmin_{\mathcal{X}} L(\mathcal{X}) \\ \text{avec : } \mathcal{X}_{i,j,r} = \sum_{k=1}^{K} A_{i,k} B_{j,k} C_{r,k} \end{gathered} \tag{5}$$

où $\mathcal{X}_{i,j,r} = \sum_{k=1}^{K} A_{i,k}B_{j,k}C_{r,k}$ correspond au fait que $\mathcal{X}$ est une décomposition tensorielle d'ordre K et A, B et C sont les facteurs matriciels issus de cette décomposition. Nous appelons le modèle proposé *Link Pattern Prediction Tensor* (LPPT).

2.4 Algorithme de prédiction

Nous proposons ici une méthode de résolution du problème de prédiction de motifs relationnels par l'algorithme du gradient conjugué non-linéaire. L'équation 4 peut être réécrite ainsi :

$$\begin{aligned} L(\mathcal{X}) &= \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{r=1}^{R} \mathcal{W}_{i,j}(\mathcal{R}_{i,j,r} - \mathcal{X}_{i,j,r})^2 \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{r=1}^{R} \mathcal{W}_{i,j}(\mathcal{R}_{i,j,r} - \sum_{k=1}^{K} A_{i,k}B_{j,k}C_{r,k})^2 \end{aligned} \tag{6}$$

et le problème de prédiction revient donc à trouver les valeurs de A, B et C qui minimisent $L(\mathcal{X})$ noté dorénavant $L(A, B, C)$. Pour cela, nous allons calculer le gradient de L et utiliser un algorithme de gradient conjugué non-linéaire pour résoudre le problème. On a :

$$\frac{\partial L}{\partial A_{i,k}} = \sum_{j=1}^{N}\sum_{r=1}^{R} \mathcal{W}_{i,j}(\sum_{k=1}^{K} A_{i,k}B_{j,k}C_{r,k})B_{j,k}C_{r,k} - \sum_{j=1}^{N}\sum_{r=1}^{R} \mathcal{W}_{i,j}\mathcal{R}_{i,j,r}B_{j,k}C_{r,k} \tag{7}$$

Les dérivées partielles des facteurs matriciels B et C peuvent être calculées de la même façon. Notez que la complexité de ce calcul est en $O(N^2 \times R)$ et donc que ce modèle peut difficilement être appliqué pour des données de très grande taille.

3 Résultats expérimentaux

Nous montrons ici la validité de notre modèle LPPT dans le cadre de la tâche de prédiction de motifs de liens. Pour l'évaluation, nous avons construit nos expériences à partir de la collection *Enron email* constitué de 6 977 emails envoyés entre employés de la société Enron. Nous avons extrait les expéditeurs des emails, les destinataires, les destinataires en CC et nous avons obtenu un jeu de données constitué de $N = 107$ individus distincts. Nous avons extrait quatre types relations ($R = 4$) à partir des ces données :

- Le premier type de relation relie l'expéditeur d'un email à son destinataire.
- Le second type correspond à une relation de voisinage : deux individus sont reliés si ils ont au moins un destinataire en commun
- Le troisième type est de même nature que la relation 1, mais prend en compte les destinataires CC.
- Enfin, le quatrième type correspond à une similarité sur le contenu des emails envoyés par deux individus. Cette similarité est calculée à l'aide de la fréquence des mots dans les emails et transformée en relation binaire à l'aide d'un seuil fixé arbitrairement.

Noeuds	107
Nombre de relations	4
Nombre de relations de type 1	1207
Nombre de relations de type 2	4250
Nombre de relations de type 3	624
Nombre de relations de type 4	1070

TAB. 1 – *Statistiques pour le corpus Enron*

	20%		40%	
	$K = 20$	$K = 10$	$K = 20$	$K = 10$
SVD (mono-relationnel)	0.7017	0.7046	0.6971	0.6614
LPPT (multi-relationnel)	0.8843	0.8734	0.8442	0.8403

TAB. 2 – *Performances moyennes de l'AUC sur 10 runs pour le modèle mono-relationnel (SVD) et le modèle multi-relationnel proposé (LPPT).*

Quelques statistiques sur ce corpus sont données en table 1.

Nous avons testé la capacité de notre méthode à retrouver des motifs relationnels manquants entre individus. Pour cela, la matrice $\mathcal{W}$ a été remplie aléatoirement par des valeurs à 0 ou 1, les 1 représentant les motifs manquants i.e les couples d'individus pour lesquels nous souhaiterions retrouver les relations. Nous avons calculé les performances obtenues pour différentes proportions de 0 dans la matrice $\mathcal{W}$. La mesure d'évaluation proposée est la mesure AUC qui est une mesure robuste permettant d'éviter d'avoir à fixer un seuil de classification arbitraire. Nous avons effectué une moyenne des performances sur 10 runs. Nous avons comparé les résultats obtenus avec une méthode de décomposition en valeur singulière de matrices (SVD) appliquée séparément à chaque tranche $\mathcal{R}_{i,j,:}$ du tenseur d'origine. Cette méthode correspond en fait à la version mono-relationnelle de notre modèle et permet de comparer à quelle point l'utilisation de la décomposition de tenseur qui prend en compte simultanément tous les types de relations permet d'améliorer la version classique relation par relation. Les résultats pour deux tailles de test sont présentés en table 2. Ces résultats sont des résultats préliminaires et de plus amples campagnes d'expérimentation sont en cours de lancement. Ces résultats montrent clairement que pour 20% et 40% de valeurs manquantes, le modèle multi-relationnel donne de biens meilleurs résultats que le modèle mono-relationnel calculé indépendamment pour chaque relation. Ces expériences montrent l'intérêt d'une telle méthode et doivent maintenant être complétées par plus d'expérimentations, sur différents corpus, et pour différentes configurations. Ce travail est un travail en cours.

4 Travaux relatifs

Avec le développement des réseaux sociaux en ligne, des chercheurs de différents domaines ont étudié les propriétés des réseaux sociaux avec des approches variées. Une grande partie de ce travail a porté sur l'analyse de la structure et des modèles de croissance des réseaux (Gleave

et al., 2009). Une autre direction de recherche similaire a porté sur la prédiction de la force du lien entre une paire d'objets (Kashima et al., 2009).

Les décompositions tensorielles ont également attiré beaucoup d'attention et ont été notamment appliquée à l'analyse du web (Kolda et al., 2005) et à l'analyse de réseaux de communication par e-mail (Bader et al., 2007). Dans ces applications les décompositions tensorielles sont utilisées comme outils d'analyse exploratoire en vue d'effectuer la tâche de prédiction de liens (Acar et al., 2009).

5 Conclusion

Dans cet article, nous avons proposé une nouvelle tâche de prédiction de motifs de liens fondée sur notre modèle LPPT. Via la décomposition CP d'un tenseur, nous transformons la tâche de prédiction de motifs de liens en un problème de complétion de tenseur. Ensuite, nous dérivons un algorithme d'optimisation efficace basé sur la méthode du gradient conjugué pour la décomposition tensorielle pour traiter notre problème. Nos expériences sur des données réelles montrent la précision et l'efficacité de notre modèle.

Dans nos travaux futurs, nous compléterons dans un premier temps la validation expérimentale de notre modèle sur des données à grande échelle. Nous étudierons également l'aspect temporel de l'évolution des motifs de liens.

Références

Acar, E., D. M. Dunlavy, et T. G. Kolda (2009). Link prediction on evolving data using matrix and tensor factorizations. *ICDM Workshops*, 262–269.

Bader, B., M. Berry, et M. Browne (2007). Discussion tracking in enron email using parafac. *in Survey of Text Mining : Clustering, Classification, and Retrieval, Second Edition*, 147–162.

Carroll, J. D. et J. J. Chang (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. *Psychometrika 35*, 283–319.

Gleave, E., H. T. Welser, T. M. Lento, et M. A. Smith (2009). A conceptual and operational definition of 'social role' in online community. *HICSS '09 : Proceedings of the 42nd Hawaii International Conference on System Sciences*, 1–11.

Harshman, R. A. (1970). Foundations of the parafac procedure : Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA working papers in phonetics*, 1–84.

Kashima, H., T. Kato, Y. Yamanishi, M. Sugiyama, et K. Tsuda (2009). Link propagation : A fast semi-supervised learning algorithm for link prediction. *SDM*, 1099–1110.

Kolda, T. G., B. W. Bader, et J. P. Kenny (2005). Higher-order web link analysis using multilinear algebra. *In ICDM 2005 : Proceedings of the 5th IEEE International Conference on Data Mining*, 242–249.

Summary

Many real-world data sets can be considered as a linked collection of objects with multi-type relations in many different applications, such as social network analysis or web mining. Each type of relations may play a distinct role in the multi-relational networks. However, traditional link prediction models provide only a coarse representation of single-type relation. In this work, we analyze the multi-relational data and propose a novel tensor model to discover implicit link patterns between object pairs. Under this model, we formulate the *link pattern prediction* task as a tensor decomposition problem, and derive a conjugate gradient based optimization method. Extensive experiments on real-world multi-relational data sets demonstrate the promise and effectiveness of our model.

Fouille de discussions pour l'identification de rôles sociaux

Mathilde Forestier*, Julien Velcin*,**
Djamel Zighed*,**

*Laboratoire ERIC, Université Lumière Lyon 2, 5 avenue Pierre Mendès-France, 69676 Bron Cedex
mathilde.forestier@univ-lyon2.fr,
eric.univ-lyon2.fr/~mforestier
**julien.velcin@univ-lyon2.fr,
eric.univ-lyon2.fr/~jvelcin
*** abdelkader.zighed@univ-lyon2.fr,
eric.univ-lyon2.fr/~zighed

Résumé. Le Web 2.0 a permis l'apparition de communautés communicantes notamment par le biais des forums. Ces forums sont une source incroyable de connaissances de part l'étendu des sujets qu'ils abordent, mais aussi des individus qui y participent. Cette importante masse de données est très difficile à appréhender pour un seul homme au vu du nombre de messages qu'ils contiennent. Notre travail propose un nouveau cadre formel pour synthétiser l'information contenu sur ces forums. Pour ce faire, nous extrayons un réseau social des participants qui prend en compte la structure et le contenu pour créer les relations. Nous définissons quatre types de relations regroupés selon la sous-thématique abordée par les participants. Grâce à ce réseau social, nous pouvons extraire le rôle social des individus participant à la discussion. Les expériences réalisées sur des discussions web réelles montrent les avantages de notre approche.

1 Introduction

Le Web 2.0 a permis aux internautes de ne plus seulement être spectateur de l'espace virtuel mais acteur sur de nouvelles formes de communication. Il est désormais possible de partager des informations, du contenu, des opinions, de se créer des réseaux d'amitié, professionnels etc. Ces nouvelles formes de communication créent une masse de données considérable difficilement exploitable. Dans cet article, nous nous attacherons à étudier les forums pour plusieurs raisons. La première provient de la diversité des sujets traités par ce médium : il existe des forums sur à peu près tous les sujets existant dans notre monde, que ce soit au niveau politique, économique, d'entraide technique ou sociale etc. De plus, le contenu des forums est disponible depuis internet ce qui permet un rapatriement facile et rapide des données. Enfin, les forums mettent en relation des individus qui interagissent entre eux, donnent leur opinion, partagent des idées, communiquent par l'intermédiaire d'un nouveau support. Ces données nous semblent une source privilégiée d'information pour atteindre notre objectif. Comme dans tout groupe communicant, chaque personne joue un rôle social. Ce rôle met en

exergue l'importance d'un individu dans le groupe, il a donc une importance primordiale dans la compréhension de l'interaction. Le but de ce travail est multiple : il permet à partir d'une discussion virtuelle de comprendre qui parle de quoi, qui répond à qui et qui est qui. Dans ce but, nous parcellons le forum en plusieurs sous-thématiques favorisant ainsi la compréhension des différents sujets abordés dans la discussion. Enfin, nous recherchons les différentes formes de relations entre les individus à partir de la structure du forum et de son contenu pour construire le réseau social, puis nous recherchons les personnes jouant un rôle dans la discussion virtuelle.

Cet article se compose de quatre parties. La première est un état de l'art sur les différents thèmes abordés dans ce travail, puis nous expliquons le cadre formel de cette recherche. Ensuite, nous verrons les expérimentations et les résultats basés sur des données réelles et enfin nous concluons et apportons les perspectives de notre travail.

2 Etat de l'art et définition

2.1 Extraction de réseau social

La réputation des réseaux sociaux n'est aujourd'hui plus à faire. Cet outil utilisé dans les sciences humaines depuis quelques décennies prend un nouvel essor avec Internet. La création de nombreux sites tels que facebook, likendin, Twitter etc. amène les utilisateurs à percevoir le web comme un lieu d'échange d'idées, d'opinion, de contenu. De même, les forums ont permis l'apparition de nouvelles communautés d'individus rassemblés par l'intermédiaire d'un intérêt commun : le sujet de la discussion.

Avec l'émergence de ces différents sites, nous voyons apparaître une nouvelle communauté de chercheurs sur les réseaux sociaux. Cet outil est de plus en plus utilisé pour mettre en exergue les relations non apparentes entre individus. Ainsi de nombreux chercheurs se basent sur les textes pour extraire les relations. Culotta et al. (2004) présentent un système complet intégrant les emails et le contenu du web pour aider les utilisateurs à maintenir à jour de grandes bases de données de contacts. De plus, par l'extraction du réseau social de ces personnes, les auteurs détectent des communautés d'individus. Jing et al. (2007) utilisent des retranscriptions de discours de survivants de l'Holocauste. Par un certain nombre de modules d'extraction d'information les auteurs extraient automatiquement le réseau social des individus ainsi que leurs événements biographiques (date de naissance, décès, mariage etc.). Le réseau social a été en premier lieu extrait par un expert puis comparé au réseau social extrait par le système. Les auteurs utilisent ensuite les mesures de précision et de rappel pour quantifier la performance de leur modèle. Mika (2005) propose un sytème pour l'extraction, l'agrégation et la visualisation de réseaux sociaux en ligne. L'auteur se base sur la communauté de chercheurs du web sémantique et extrait un réseau social basé sur l'analyse des pages personnelles, des profils, des emails et des publications. Ce réseau social permet une analyse des rôles des individus dans la communauté. Jin et al. (2007) reconnaissent l'idée que la force d'une relation entre deux individus peut-être mesurée par le nombre de cooccurences du nom sur le web principalement pour les chercheurs. Toutefois, cette mesure ne peut être applicable aux différentes entités que l'on peut trouver sur internet notamment en ce qui concerne les relations entre entreprises ou entre artistes. Les auteurs extraient à partir des actualités plusieurs types de relations. Par exemple, les entreprises peuvent être liées par une relation d'alliance (connotée positivement) ou

des poursuites judiciaires (connotée négativement). Enfin, McCallum et al. (2007) ont construit un modèle probabiliste permettant l'extraction de thématiques et de rôles à partir de textes. Ce travail bien que très proche du nôtre diffère aussi bien sur la manière de procéder que sur le type de rôles extraits.

2.2 Rôles sociaux

Le concept de rôle social est apparu dans les années 1930. Dans son ouvrage, Coenen-Huther (2005) fait un état des lieux du concept de rôle social en sociologie, il évoque deux positions sociologiques concernant ces rôles : la position structurelle et la position interactionniste. La première, définie par Linton en 1936, explique le rôle social en se basant sur l'individu. Selon lui, un "un individu joue un rôle quand il met en œuvre les droits et les devoirs qui constituent le statut". Quant à l'approche interactionniste, le concept de rôle social est défini par rapport au groupe : "les rôles de groupe différencient les membres suivant les tâches à remplir et les motivations personnelles, quoique de façon relativement positive et régulière". Dans le travail présenté ici, nous ne nous attacherons pas à faire une étude sociologique du concept de rôle social, mais de part la nature de notre travail nous nous baserons sur la définition interactionniste.

Avec l'émergence du web 2.0, des chercheurs se sont intéressés au concept de rôle social dans les communautés virtuelles interactives. En effet, chaque individu va jouer un rôle plus ou moins important dans la discussion à laquelle il participe. Ce rôle peut être mesuré par le temps de parole, la pertinence des remarques, les réactions suscitées etc. Welser et al. (2007) reprennent les termes de Hymnes[1] pour définir la notion de communauté communicante qui est selon l'auteur "un groupe d'individus qui partagent des règles pour la conduite et l'interprétation de la parole, et des règles pour l'interprétation d'au moins une variété linguistique". Ces codes sont une notion très importante puisqu'ils sont la base de l'entente entre les individus. Dans la même optique, un individu ayant acquis les codes du forum dans lequel il participe aura tendance à avoir un rôle positif, à l'inverse l'absence de ces codes va engendrer une réaction négative de la communauté. Les auteurs ont ainsi pu extraire une typologie des rôles des participants dans les communautés communicantes virtuelles. Nous verrons un peu plus loin les définitions que nous utilisons dans le travail présenté dans cet article. Ces définitions reprennent les idées de ces auteurs avec quelques variations que nous expliciterons dans la section 2.3. Golder et Donath (2004) nous donnent deux indicateurs pour appréhender les rôles sociaux dans ces communautés : la participation des auteurs et les réponses perçues. Ainsi, une forte participation implique un fort impact sur la communauté que ce soit du point vue du contenu ou du nombre de messages. Dans la même optique, le nombre de réponses donne une information précieuse sur la manière dont le message a été perçu par la communauté. Toutefois ces deux indicateurs sont fortement corrélés puisque lorsqu'un auteur reçoit des réponses positives à son message, cela l'encourage à publier de nouveau. Le nombre de messages et le nombre de réponses sont donc de très bon indicateurs pour détecter les personnes au comportement positif dans la communauté. Le travail de Fisher et al. (2006) montre que les rôles des individus ne sont pas les mêmes selon le type de forum analysé (forum technique, politique, d'entraide etc.). Les auteurs construisent un réseau social des interactions entre individus dans

[1] Hymnes, Dell. 1974. *Foundations in Sociolinguistics : An ethnographic approach*. University of Pennsylvania Press

le forum (qui répond à qui ?) puis réduisent le graphe à un réseau social égocentrique pour tous les participants. Enfin, par des mesures de connectivité, ils ont pu définir différents types de participation dans les forums.

2.3 Définitions

Dans ce travail, nous avons défini cinq rôles sociaux. Ces rôles proviennent de la littérature sociologique et informatique. Comme nous l'avons précisé plus tôt, les définitions ont été pour certaines modifiées dans le but de correspondre plus précisément aux identités que nous voulons extraire. Les personnes les plus participatives sur l'ensemble du forum sont les **leaders**. Ils correspondent à ce que Welser et al. (2007) nomment la célébrité. Les leaders sont très présents dans la discussion de part le nombre de messages envoyés et du nombre de réponses perçues. Ils apportent des informations pertinentes qui amènent les autres participants à réagir, ils mènent la discussion par leur propos. Les **experts** sont des personnes très actives dans une partie du forum (sous-thématique). Ces personnes ciblent leurs propos sur une sous-thématique intéressante selon leurs critères personnels et vont donc être actifs ponctuellement. La **célébrité** est une personne publique citée dans le corps du message d'un acteur. Elle n'est pas participative dans la discussion mais appelée en tant qu'entité publique reconnue. Le **flammer** a un comportement négatif. Il est présent pour semer le trouble dans la discussion, créer du conflit mais n'apporte aucune idée nouvelle. Enfin, le **rôdeur** ne participe pas activement à la discussion. Il publie assez peu de messages et n'a que très peu de réponses. Il n'apporte pas de nouvelles idées dans la discussion ou ne connaît pas les codes permettant d'interagir avec les autres participants.

3 Cadre formel

Les réseaux sociaux sont de plus en plus utilisés dans la recherche informatique pour comprendre les relations entre individus. Dans cette partie nous nous attacherons à présenter un nouveau modèle permettant d'appréhender les discussions virtuelles.

Les sous-thématiques Tous les forums existants sur internet abordent un sujet précis (ex : Les lunettes de Gandhi vendues aux enchères). Sur de nombreux sites d'information, la page web se compose d'un article de presse suivi d'un forum permettant aux individus d'interagir. Toutefois, comme dans toutes les discussions, qu'elles soient virtuelles ou non, le sujet principal regroupe divers thèmes : c'est ce que nous appelons les sous-thématiques. Ces dernières peuvent être extraites selon plusieurs méthodes : les méthodes probabilistes tel que LDA (Latent Dirichlet Allocation, Blei et al. (2003)), ou des méthodes de type statistique telles que les K-means ou encore les Overlapping K-means (OKM, Cleuziou (2007)). Nous verrons dans la partie 4.1 la méthode utilisée par notre système.

Les relations entre individus La plupart des articles cités précédemment ne prennent en compte qu'un seul type de relation. Toutefois, après lecture d'un certain nombre de forums sur le web, nous nous sommes aperçus qu'il existait en plus de la relation structurelle, d'autres formes de relations. En effet, un auteur peut répondre à plusieurs auteurs dans un même message. Nous en avons défini quatre types :

- la relation structurelle : comme son nom l'indique elle est donnée par la structure même du forum, c'est à dire lorsqu'un utilisateur clique sur un lien pour spécifier à qui il répond.
- la citation du nom : dans le contenu du message, un auteur peut citer soit un autre auteur soit une célébrité en utilisant son nom ou son identifiant.
- la citation d'un texte : cette relation est créée lorsqu'un auteur reprend une partie du texte d'un autre message posté précédemment.
- la similarité textuelle : cette relation est créée lorsque deux auteurs ont posté des messages contenant les mêmes mots. Les deux messages étant très similaire, nous pouvons soupçonner qu'ils parlent de la même chose.

Il existe une relation entre deux auteurs lorsqu'un auteur répond à un autre en utilisant l'un des types de relation sur une sous-thématique, ce qui implique qu'un auteur peut-être lié à un autre de plusieurs façons.

Formalisation Nous avons tout d'abord défini quatre ensembles :

- T : l'ensemble des sous-thématiques extrait à partir du forum avec $T = \{t_1, ..., t_p\}$, p étant défini a priori.
- X : l'ensemble des acteurs et des célébrités avec $X = \{x_1, ..., x_n\}$.
- R : l'ensemble des différents types de relations. R est un ensemble fini de quatre relations possibles.
- D : l'ensemble des documents (chaque message représente un document) avec $D = \{d_1, ..., d_z\}$.

A partir de ces quatres ensembles, nous pouvons définir les applications suivantes :

- Chaque message appartient au moins à une sous-thématique.
  $sousThematique : D \rightarrow \mathcal{P}(T)$
  $d \mapsto T_d$ avec $T_d \subset T$ correspondant aux sous-thématiques extraites.
- Un message est écrit par un seul auteur.
  $auteur : D \rightarrow X$
  $d \mapsto x$

Et les relations binaires suivantes sachant que $\delta \in \{str, c_{texte}, sim\}$[2] :

- $d_a R_{t,\delta} d_b \Leftrightarrow$ le document d_a est en relation de type δ avec d_b sur la sous-thématique t.
- $d_a R_{t,c_{nom}} x \Leftrightarrow$ le document d_a cite le nom de x sur la sous-thématique t.

Ainsi que les relations entre auteurs :

- $x_i R_{t,\delta} x_j \Leftrightarrow \exists d_a, d_b \in D \times D / d_a R_{t,\delta} d_b$ et $auteur(d_a) = x_i$ et $auteur(d_b) = x_j$
- $x_i R_{t,c_{nom}} x_j \Leftrightarrow \exists d \in D / d R_{t,c_{nom}} x_j$ et $auteur(d) = x_i$ [3]

Nous définissons un graphe G = (X,A) où chaque arc a_{ijkl} représente une réponse d'un auteur x_i à x_j sur une sous-thématique $k \in T$ avec une relation de type $r_l \in R$. Les arcs sont pondérés selon la fonction de pondération suivante :

$$\gamma(a_{ijkl}) = f(nb(x_i, x_j, t_k, r_l)) \text{ avec } \forall x_i, x_j \in X, t_k \in T \text{ et } r_l \in R \text{ où } R = \delta \bigcup \{c_{nom}\}$$

Où

[2] str correspond à la relation structurelle, c_{texte} à la relation de citation du texte et sim à la relation de similarité textuelle.

[3] c_{nom} est la citation du nom.

$$nb(x_i, x_j, t_k, r_l) = |\{(d_a, d_b) \leftarrow D \times D / auteur(d_a) = x_i \text{ et } auteur(d_b) = x_j \text{ et } t_k \in sousThematique(d_a) \text{ et } t_k \in sousThematique(d_b) \text{ et } r_l \in R)\}|$$

Cela signifie qu'il existe un arc de x_i à x_j, dont le poids correspond au nombre de fois où x_i répond à x_j sur une sous-thématique selon un type de relation.

Pour illustrer les formules précédentes, nous avons créé l'exemple ci-dessous :

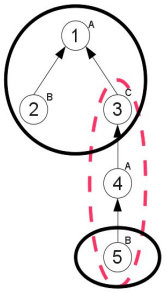

FIG. 1 – *Graphe des messages de la discussion*

La figure 1 expose le graphe des messages d'une discussion. Chaque noeud représente un message publié par un auteur (A, B ou C), les arcs représentent la relation structurelle du forum. Le forum est découpé en deux sous-thématiques représentées par les ellipses : l'ellipse en trait plein contient les messages appartenant à la première sous-thématique, celle en trait pointillé à la deuxième. Notons que les messages peuvent appartenir à plusieurs sous-thématiques. A partir du graphe des messages (figure 1), nous pouvons extraire les relations structurelles entre les trois auteurs mais aussi les relations basées sur le contenu grâce aux textes des messages. Le tableau 1 représente toutes les relations qui pourraient être extraites à partir du forum sur les deux sous-thématiques.

TAB. 1 – *Exemple*

Sous-thématique	Structurelle	Citation du nom	Similarité textuelle	Citation du texte
1	document 2 : B→A document 3 : C→A document 5 : B→A	C → B		C→A
2	document 4 :A→C document 5 : B→A		B → A	

La figure 2 représente le réseau social extrait à partir du forum. Les auteurs sont liés par différents types de relations sur les sous-thématiques. Le poids des arcs est calculé par la formule vue dans la section 3. Dans cet exemple, nous pouvons voir que B répond quatre fois à A : trois fois par une relation structurelle et une fois par similarité textuelle. Deux des trois

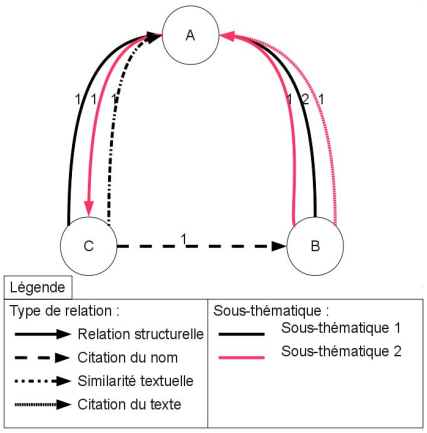

FIG. 2 – *Réseau social extrait à partir du tableau 1*

relations structurelles portent sur la même sous-thématique. Il y a donc trois arcs allant de B à A dont une avec un poids de 2 (même type de relation et même sous-thématique).

Par cette formalisation nous obtenons un p-graphe maximum lorsque p=$n*(n-1)*|T|*4$ où n représente le nombre d'acteurs du forum, $|T|$ le nombre de sous-thématiques, et 4 car les relations sont de quatre types.

Grâce à ce modèle, nous extrayons les acteurs des différents rôles sociaux : pour chaque rôle (à part la célébrité), nous définissons des critères. Ces critères vont permettre de créer un ordre sur les acteurs et à l'aide d'une méthode d'agrégation multicritère nous pouvons extraire les acteurs jouant un rôle dans la discussion. La méthodologie pour extraire les rôdeurs est basée sur des seuils définis a priori.

Le Leader

le leader est défini sur la base de deux critères.

Critère 1 : $leader(x_j) = \sum_{r=1}^{4} \sum_{x_i \in X} \gamma(a_{ijkl})$ mesure le degré de centralité d'un acteur sur le forum entier. Plus l'auteur aura de réponses, mieux il sera placé car ses messages sont bien perçus par la communauté.

Critère 2 : $leader(x_j) = \sum_{k \in T} nb(d_{j,k})$ calcule le nombre de messages postés par l'utilisateur.

L'expert

Critère 1 : $expert(x_j, k) = \frac{\sum_{r=1}^{4} \sum_{x_i \in X} \gamma(a_{ijkl})}{\sum_{r=1}^{4} \sum_{x_i \in X} \sum_{k \in T} \gamma(a_{ijkl})}$ où $x_i \neq x_j$ et r correspond au type de relation et k le numéro de la sous-thématique. Ce critère représente le degré de centralité d'un auteur sur une sous-thématique normalisé par le degré de centralité sur le forum entier. Cette mesure permet de privilégier les individus qui n'ont participé que dans une seule sous-thématique.

Critère 2 : $expert(x_j, k) = \frac{nb(d_{j,k})}{\sum_{k \in T} nb(d_{j,k})}$ où $d_{j,k}$ est un message de l'auteur x_j dans la sous-thématique t_k. Ce critère représente le nombre de messages postés par un auteur dans une sous-thématique normalisé par le nombre de messages du même auteur sur l'ensemble du forum.

Le rôdeur

Critère 1 : rôdeur$(x_j) = \sum \gamma(aijkl) < \alpha$: le rôdeur n'apporte pas de nouvelles idées dans la discussion, personne ne lui répond, ses messages sont jugés sans intérêt.

Critère 2 : rôdeur$(x_j) = nb(d_j) < \varphi$: le rôdeur ne poste pas beaucoup de messages soit par manque d'intérêt pour le sujet, soit par manque de réaction de la communauté. Notons que $\varphi > 0$ car il doit avoir écrit au moins un message (il ne peut pas être une célébrité).

Les paramètres doivent être défini a priori.

La célébrité

La célébrité a un rôle particulier dans la discussion puisqu'elle n'est pas participative mais représente une personne publique citée dans au moins un message. C'est en réalité un problème de détection d'entité nommée. Nous verrons dans la sous section 4.1, la méthode de détection que nous souhaitons utiliser. Toutefois, nous pouvons définir un critère permettant de calculer l'importance de la célébrité : celebrité$(x_j) = nb(a_{ijkl})$

Le flammer

Le flammer est un rôle difficile à extraire puisqu'il faut prendre en compte l'opinion des messages. Nous devons trouver une conversation entre deux acteurs où les messages sont connotés négativement. A ce stade, le rôle n'est pas inclus dans le système. Nous verrons dans la partie conclusion que la fouille d'opinion est une des perspectives de notre travail.

4 Expérimentations et résultats

4.1 Présentation du système

L'extraction du réseau social et des individus jouant un rôle social se déroule en 6 étapes :

1. Parser le forum depuis internet.
2. Extraire les sous-thématiques : Pour cela nous utilisons un clustering conceptuel recouvrant expliqué dans Rizoiu et al. (2010), dans une version toutefois un peu différente qui prend en compte, en plus de la similarité textuelle, la structure du forum. En effet, les messages postés par les auteurs sont de taille relativement réduite et n'utilisent pas forcément le même vocabulaire tout en parlant de la même chose. Rajouter la structure dans le clustering permet de garder les fils de messages dans les sous-thématiques.
3. Extraire les auteurs et les célébrités : en ce qui concerne les auteurs, la tâche est assez simple puisqu'elle consiste à relever les noms depuis le forum parsé. Pour ce qui est des célébrités, nous souhaitons utiliser un outil de taggage morpho-syntaxique. Cet outil (TreeTagger) récupère le lemme de chaque mot du corpus et l'annote selon son type : nom, verbe, nom propre etc. Puis, avec ce fichier annoté nous utilisons un outil du nom de Unitext qui nous permet de retrouver seulement les entités nommées.
4. Extraire les relations : il existe une relation allant de x_i à x_j lorsque x_i répond à x_j sur une sous-thématique selon un des types de relations comme il est expliqué dans la section 3. Le nombre maximal d'arcs entre deux acteurs correspond à 4*$|T|$*2 où $|T|$ représente le nombre de sous-thématiques. Nous utilisons ici le cardinal pour la fonction de pondération f vue dans la section 3.
5. Calculer les critères proposés dans la section 3 pour chaque acteur sur les différents rôles sociaux.

6. Extraire les différents acteurs sur chaque rôle : pour identifier les leaders et les experts des sous-thématiques nous utilisons un front de Pareto (afin d'extraire tous les individus non-dominés sur les différents critères). Le rôdeur comporte deux seuils à définir a priori. Dans les expérimentations suivantes, nous fixons $\alpha = 0$ et $\varphi < 2$. Nous considérons que l'auteur jouant un rôle de rôdeur a posté un message sans réponse.

4.2 Expérimentations

Pour tester notre modèle nous avons utilisé deux forums différents qui proviennent du site d'information *rue89* (www.rue89.com), chaque page du site comprend un article de presse puis un forum de débat public.

4.2.1 Les lunettes de Gandhi vendues à New York, l'Inde s'insurge

Le premier forum[4] date du 3 mars 2009 et porte sur les lunettes de Gandhi vendues aux enchères à New york. Il se compose de 90 messages écrit par 37 auteurs et comprenant 4 sous-thématiques. Notons ici que les sous-thématiques sont données a priori dans le système d'extraction de rôles sociaux. Nous ne jugerons pas dans cet article de la pertinence des résultats du clustering. La première aborde les notions d'Indiens, d'ombre jaune (l'ennemi juré de Bob Maurane) et de paires de lunettes. Elle contient 54 messages écrits par 24 auteurs. La deuxième discute du "Crâne de Napoléon", de lunettes, de valeur (au sens monétaire) et de symbole. Elle contient 47 messages rédigés par 24 auteurs. La troisième est centrée sur les notions d'humanité, d'humain, de génie mais aussi de pauvres et de poubelle. Elle contient 24 messages écrits par 11 auteurs. Enfin la dernière nous parle de Ghandi, du mépris et de poubelle.

Sur ce forum, le système extrait deux leaders. Ces deux résultats sont cohérents au regard du nombre de messages et de réponses suscitées. Le premier, du nom de Cyp apporte de nouvelles idées dans la discussion. Ses propos sont pertinents par rapport à la thématique générale. Son premier commentaire est une réflexion sur l'Inde, Gandhi et la vente de ses lunettes : "Quoi ! L'inde s'offusque ? [...] Je trouve tout à fait normal et d'une parfaite insignifiance que les binocles du petit père Gandhi, que je révère, soient mises aux enchères. C'est logique. Je m'en fous : les connards pleins aux as qui les achèteront n'auront en main qu'un bout de fil de fer et deux petits morceaux de verre sans aucune importance : l'esprit du mâhâtma est ailleurs [...]". Il répond aux interrogations des autres participants : "Oui alors là, N 6, c'est toute une histoire, là aussi... Leurs délires comme quoi le Dalaï Lama aurait été fasciné par le nazisme, c'est vraiment du grand n'importe quoi ! [...]". Au regard de sa participation dans le forum tant par le contenu de ses propos que par le nombre de ses messages, Cyp correspond bien à la définition d'un leader. Nous pouvons remarquer la qualité orthographique et syntaxique de ses propos. Dans ce type de forum, la qualité de la rédaction est un code. Certains auteurs n'hésitent pas à faire des remarques aux auteurs dont le(s) message(s) contienne(nt) des fautes "Nullissime. En quoi le peuple indien vous a-t-il lésé ? [...](dixit... en corrigeant votre faute d'orthographe ..) [...]". Le deuxième leader : Banana Ex de juanitoto est plus contestable. Lorsque l'on regarde le nombre de messages et de réponses, Banana Ex de juanitoto correspond à la définition d'un leader. Toutefois, ses commentaires se résument en une phrase et une image pas toujours en

[4] http ://www.rue89.com/2009/03/03/les-lunettes-de-gandhi-vendues-a-new-york-linde-sinsurge

rapport avec le sujet principal du forum. Sa place en tant que leader est donc discutable. Du point de vue de la temporalité, nous devons être attentif au moment d'intervention des auteurs. Dans le jeu de données présenté, leur participation s'est intensifiée à la fin de la discussion. il serait intéressant d'ajouter une dimension temporelle lors de l'extraction des auteurs jouant un rôle.

En ce qui concerne les différents experts des sous-thématiques, le système en extrait cinq sur la première, trois dans la seconde, et un dans les deux dernières. L'expert DBL8 traite de la sous-thématique portant sur les indiens (sous-thématique 1) : "Vous devriez apprendre à COMPRENDRE ce que vous lisez ! ! Je n'ai pas mis en cause les ouvriers *Indiens* [...]". L'expert extrait dans la troisième sous-thématique reprend presque tous les noms décrivant sa sous-thématique dans son message : "LE *GÉNIE* PAR PROCURATION MONÉTAIRE Et tous les médias de s'enthousiasmer : "la *vente* du siècle ", [...] Ça, c'était pour la *vente* Bergé-St Laurent. Des trésors du patrimoine de l'*humanité*, [...] Où vont-ils échouer, ces chefs-d'œuvres du *génie humain* ? dans quelle forteresse barricadée ? [...] Le *génie* et la grandeur qui leur font défaut ? *Pauvres* petits ! *Pauvres* minables ! Les voilà maintenant qui jettent leur dévolu sur les dépouilles d'un autre *génie humain*, celui de la justice, de la fraternité, [...]". Enfin, Le système détecte six rôdeurs dans ce forum. Sur ces six individus, quatre ne font partie d'aucun fil de discussion : ce sont eux les vrais rôdeurs. En effet, les deux autres, par leur volonté de s'intégrer dans un fil de discussion ne s'apparente pas à des rôdeurs. Il va donc falloir ajouter un critère pour l'extraction de ces individus.

4.2.2 Séisme à Haïti : la Croix Rouge parle de "milliers" de morts

Le deuxième forum [5] extrait à partir de la même source traite du séisme survenu en Haïti. Il date du 13 janvier 2010 et contient 83 messages rédigés par 35 auteurs. Il a été découpé en trois sous-thématiques. La première aborde les thèmes de la famille, du père, des Haïtiens, de la catastrophe et d'Eric Besson. Elle contient 42 messages rédigés par 23 auteurs. La deuxième parle de sable, d'argile, de vacances et d'histoire. Elle contient 32 messages rédigés par 21 auteurs. Enfin, la troisième aborde les notions de séisme, de zone sismique et de pauvreté.

Sur ce forum, le système extrait deux leaders : Batila et Kawaii. En regardant plus précisément les résultats, nous pouvons voir que Batila a posté 15 messages sur les 83 composant le forum et la majorité d'entre eux parlent du séisme : " [...] La solidarité internationale tente de réparer les dégâts. [...]", " [...] Même si indépendamment de cette catastrophe ce que vous dites était vrai, cette catastrophe dépasse largement le sort des réfugiés et autres immigrés de travail [...]", " [...] Pour moi le président haitien est criminel d'avoir laisser autant de gens s'installer dans une zone aussi sismique [...]" etc.

Sur la première sous-thématique, deux experts ont été trouvés : Pierre Harski et caro. Ces deux experts ont bien été détectés car ils abordent tous deux le vocabulaire de la sous-thématique. Par exemple, Pierre Haski écrit : "Ce n'est pas aussi absurde qu'il y parait. Il y a des millions de *Haïtiens* dans la diaspora qui cherchent des informations sur l'ampleur de la *catastrophe*", et dans le même registre caro publie "j'espère que le *père* est toujours vivant et qu'il pourra venir retrouver sa famille. Si le *père* est mort dans le tremblement de terre *besson* est responsable". Sur les cinq experts extraits dans la deuxième sous-thématique, nous pouvons voir que Romi45 joue effectivement ce rôle : "plus de *sable*, il le mange la bas sous

[5] http ://www.rue89.com/2010/01/13/seisme-a-port-au-prince-lappel-a-laide-du-president-haitien-133478-0

forme de galette depuis l'augmentation du prix du riz" ainsi que Chenawon : "De l'*argile* mon cher, de l'*argile* pas du *sable* ! ! J'ai goûté sur place, *argile*, au et un peu de beurre, *histoire* de combler l'estomac et l'impression de faim !". Enfin, morphee 78 joue le rôle d'expert dans la dernière sous-thématique : "Avant de prétendre dire le vrai et le faux , il convient de le dire dans un Français correct . Secousse tellurique , *séisme* , vous avez le choix . Mais secousse *sismique* , cela se mesurer sur l'échelle de Jacob et du pléonasme [...]". Nous remarquons ici aussi l'allusion négative à la qualité rédactionnelle des auteurs.

Les rôles d'experts et de leaders extraits sur les deux forums paraissent pertinents. De plus, les données utilisées dans ce travail sont des messages courts contenant des fautes de frappe, des abréviations etc. ce qui nous amène à penser que nos résultats sont bons. Toutefois, l'extraction des experts peut être améliorée : les individus extraits pour ce rôle ne sont pas ceux qui ont postés le plus grand nombre de messages dans la sous-thématique. En effet, la normalisation du critère comme expliquée dans la section 3 va avantager les individus qui ont peu écrit (souvent un ou deux messages) du fait d'un fort taux de recouvrement entre les sous-thématiques . Il semble donc judicieux de rajouter des critères comme le nombre de fils de discussion introduits dans les sous-thématiques et leurs longueurs. Ainsi, nous pourrons affiner l'extraction des personnes jouant un rôle d'expert. Malgré cela, nous constatons que les experts ont des messages bien ciblés sur leur sous-thématique alors que les leaders parlent du sujet du forum de manière plus générale.

5 Conclusion et Perspectives

Ce travail présente un nouveau modèle pour extraire automatiquement une représentation synthétique d'une discussion virtuelle sur le web. Le modèle créé fait appel à des techniques de fouille de textes afin d'extraire différents types de relations entre les acteurs. Ces relations sous-jacentes basées sur le contenu de la discussion permettent d'enrichir le graphe (construit à partir des relations structurelles) et apporte une précision plus juste des interactions entre individus. De plus, la typologie des rôles proposées dans ce travail permet de prendre en compte la notion de sous-thématique dans une discussion, ces dernières favorisant la compréhension des différents sujets abordés par les participants. Enfin, les premiers résultats obtenus sur la détection des individus jouant un rôle nous encouragent à poursuivre nos recherches dans cette direction.

Par la suite, nous nous attacherons à affiner les critères pour une meilleur extraction des différents rôles et ajouter une analyse temporelle de ces derniers. Une autre perspective est de croiser des discussions provenant de plusieurs sources d'information. Nous souhaitons aussi intégrer un module d'extraction d'opinion sur les messages du forum. Enfin, ce modèle devra être assez général pour se baser sur des sources différentes telles que les blogs, chats et/ou base de données d'email.

Références

Blei, D., A. Ng, et M. Jordan (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research 3*, 993–1022.

Cleuziou, G. (2007). OKM : une extension des k-moyennes pour la recherche de classes recouvrantes. *EGC 07, Namur, Belgique, RNTI-E 9*, 691–702.

Coenen-Huther, J. (2005). *Heurs et malheurs du concept de rôle social*. Librairie Droz.

Culotta, A., A. McCallum, et R. Bekkerman (2004). Extracting social networks and contact information from email and the web.

Fisher, D., M. Smith, et H. Welser (2006). You are who you talk to : Detecting roles in usenet newsgroups. In *Hawaii International Conference on System Sciences*, Volume 39, pp. 59. Citeseer.

Golder, S. et J. Donath (2004). Social roles in electronic communities. *Internet Research 5*, 19–22.

Jin, Y., Y. Matsuo, et M. Ishizuka (2007). Extracting social networks among various entities on the web. *Lecture Notes in Computer Science 4519*, 251.

Jing, H., N. Kambhatla, et S. Roukos (2007). Extracting social networks and biographical facts from conversational speech transcripts. In *Annual Meeting-Association for Computational Linguistics*, Volume 45, pp. 1040.

McCallum, A., X. Wang, et A. Corrada-Emmanuel (2007). Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research 30*(1), 249–272.

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

Rizoiu, M.-A., J. Velcin, et J.-H. Chauchat (2010). Regrouper les données textuelles et nommer les groupes à l'aide des classes recouvrantes. In *10ème Conférence Extraction et Gestion des Connaissances (EGC), Hammamet, Tunisie*, Volume E-19 of *Revue des Nouvelles Technologies de l'Information*, pp. 561–572. Cépaduès.

Welser, H., E. Gleave, D. Fisher, et M. Smith (2007). Visualizing the signatures of social roles in online discussion groups. *Journal of Social Structure 8*(2).

Summary

Forums are an amazing source of knowledge from all the topics they cover but also from the individuals who participate. This huge mass of data is very difficult to understand for a lonely human attending to the number of posts they contained. Our work propose a new formal border to synthesis the information contains in these forums. In fact, we extract a social network from the participant using both the structure and the content to create relashionships. We define a set of four types of relationship clustered by the subtopic they cover. Thanks to the social network, we can extract the social role of individual to know who is who in the conversation. The experiments performed on real web discussions clearly show the benefit of our approach.

REiSO 2010

Atelier REcherche et REcommandation d'information dans les RESeaux sOciaux