Mesurer l'intrication sémantique dans une collection de documents

Benjamin Renoust[*,**], Guy Mélançon[**], Marie-Luce Viaud[*]

[*]Institut National de l'Audiovisuel
{brenoust, mlviaud}@ina.fr,
[**]CNRS UMR 5800 LaBRI & INRIA Bordeaux Sud-Ouest
{benjamin.renoust, guy.melancon}@labri.fr

Résumé. Avec l'explosion des ressources en ligne, l'exploration efficace de collections de documents est une tâche quotidienne pour bon nombre de professions, mais paradoxalement reste un enjeu pour la recherche. Nous proposons trois outils pour aider à percevoir le contenu sémantique d'un ensemble de documents : un réseau pondéré d'interaction des termes associés au jeu, une mesure d'intensité et une mesure d'homogénéité d'intrication sémantique reflétant les caractéristiques de partage des termes par les différents items de la collection. De plus, ces mesures sont utilisées pour catégoriser les collections.

1 Introduction

La plupart des utilisateurs de système d'information explore quotidiennement des collections de documents. Les scientifiques parcourent aussi chaque jour des bases de données de publications. Les journalistes interrogent des bases de données à la recherche de documents publiés autour d'un évènement. Ainsi les sociétés ou instituts spécialisés (tels que l'AFP, Reuters, l'Ina, la BnF...) offrent des services de requête et de recherche dans leur archives documentaires. La réalisation de nouvelles techniques ou l'amélioration de techniques d'exploration et de recherche des collections de documents satisfait une réelle demande des utilisateurs tant grand public qu'acteurs industriels.

La plupart des moteurs de recherche classe les documents correspondant à une requête suivant un ordre de pertinence [Croft et al. (2009)], parmi ces indices, Pagerank [Brin et Page (1998)] est sûrement le plus connu et le plus utilisé. Cependant ce classement ne prend pas en compte directement la sémantique des documents retournés et renvoie un ordre souvent peu compréhensible à l'utilisateur.

Notre contribution vient en complément des approches d'indexation et d'identification classiques. Nous introduisons un *réseau pondéré d'interaction des termes* permettant d'introduire plusieurs mesures d'intrication sémantique d'une collection. Ce réseau est généré à partir de termes associés à la collection quelque soit leur provenance (annotations manuelles, statistiques, LSA, LDA...). Nous conduisons une analyse spectrale inspirée des analyses de réseaux sociaux [Burt et Scott (1985)] afin de définir un *indice d'intrication* associé à chaque terme et deux mesures globales d'intrication sémantique. Notre objectif est de permettre à l'utilisa-

teur d'évaluer sémantiquement la cohésion d'une collection ou d'un sous-ensemble désigné ou d'identifier des zones de forte intrication.

Ce papier présente brievement l'état de l'art avant d'introduire en section 2 le réseau d'interaction des termes et les indices d'intrication. Une étude de cas est ensuite discutée afin de montrer le potentiel d'utilisation de ces outils (section 3). Nous montrons ensuite le graphe d'interaction des termes et les indices d'intrication sur des données obtenues à l'Ina [1]. Nous rapportons enfin les résultats d'une étude utilisateurs conduite avec des documentalistes experts. Nous concluons avec une discussion en section 5.

1.1 État de l'art

L'analyse de collections de documents considère souvent une matrice de co-occurrence à partir de laquelle différents indices peuvent être dérivés. Un indice connu est le tf-idf [Salton (1983)] qui détermine un poids pour des termes. Les documents d, d' peuvent alors être considérés comme des vecteurs de poids indexés par terme, correspondant à une ligne dans la matrice de co-occurrence. Ces vecteurs peuvent être utilisés pour évaluer les (dis)similarités entre documents. La similarité cosinus $\cos(d, d') = \frac{\langle d, d' \rangle}{\|d\| \|d'\|}$ est une autre mesure connue et très utilisée.

Depuis les travaux pionniers de Salton (1983), les chercheurs ont proposé des améliorations au model "bag-of-words" (par exemple [Beaza-Yates et Ribeiro-Neto (1999)]). L'analyse sémantique latente (LSA) [Dumais et al (1988)] ou l'indexation sémantique latente (LSI) [Deerwester et al. (1990)] exploite l'idée que les mots ayant un sens similaire apparaissent souvent de manière proche dans un texte. Ces méthodes évaluent la proximité sémantique en appliquant une décomposition en valeurs singulières sur une matrice d'occurences de termes. L'indexation latente sémantique probabiliste (PSLI) [Thomas (1999)] s'appuie sur une décomposition mixte dérivée d'un modèle latent classique qui peut être ajusté par un algorithme d'esperance-maximisation (EM).

L'allocation de Dirichlet latente (LDA) [Blei et al. (2003)] est un modèle de sujets similaire à PLSI, où chaque document est vu comme un mélange de sujets variés. LDA suppose que chaque document est un mélange d'un petit nombre de sujets, où la présence des mots dans les documents est attribuable à l'un des sujets du document. LDA utilise une modélisation probabiliste des fréquences d'occurrence de termes dans les documents pour définir un sujet, ayant pour but de trouver les termes ou sujets les plus pertinents dans un jeu de documents. Les documents peuvent ainsi être décrits en utilisant un vecteur pondéré ou une distribution probabiliste par terme, ce qui permet à l'utilisateur de calculer des similarités entre documents pouvant ainsi nourrir différents algorithmes de recherche et/ou d'agrégation sur des collections de documents [Kohonen et al. (2000) ou Zhao et al. (2005)]. Il est necessaire ici de bien distinguer un sujet d'un terme. En pratique, LDA calcule une distribution de *sujets* en utilisant une collection de documents, ce qui correspond à une distribution de probabilités associée à un jeu de mots présents dans la collection de documents. Nous nous intéressons seulement aux *termes* qui correspondent aux mots trouvés dans des documents ou issus d'un vocabulaire contrôlé utilisé pour l'indexation des documents.

Notre approche diffère de ces techniques d'indexation sur plusieurs points. Nous considérons d'abord le réseau d'interaction commme un élement central duquel sont dérivés nos

1. www.ina.fr

indices d'intrication et à partir duquel nous pouvons tirer nos conclusions. Notre approche peut être semblable à celle de Arun et al (2010) en ce qu'elle considère un réseau termes-documents plutôt que la matrice termes-sujets utilisée par LDA. Arun et al (2010) utilisent une matrice de sujets-documents pour estimer le véritable nombre de sujets présents dans une collection de documents. Nos objectifs sont différents puisque nous visons à établir si un groupe de documents forme un groupe cohésif selon un jeu de termes d'indexation donné. Cependant la topologie du reseau d'interaction (composantes connexes, densités) permet de faire émerger les différents sujets (donc leur nombre) emmêlés dans la collection de documents (section 3)

L'indice d'intrication peut être déterminé sur *n'importe* quel groupe de documents et à partir de *n'importe* quel jeu de termes indexant ces documents. Notre technique apparaît alors comme une procédure a posteriori, offrant un retour sur une quelconque procédure d'agrégation ou d'indexation. Les indices d'intrication sont basés sur les interactions qui prennent place entre les termes (section 2) et qui exploitent totalement la topologie du réseau d'interaction.

2 Intrication sémantique

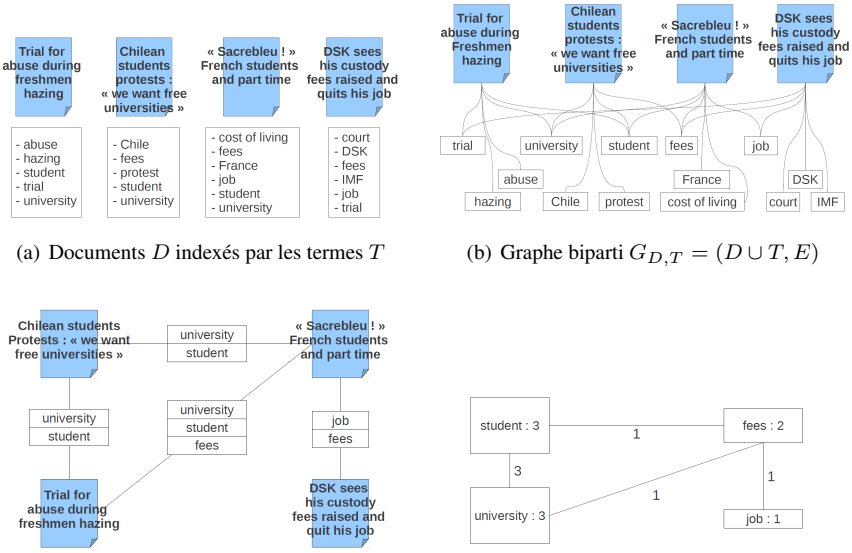
Nous allons maintenant définir un indice d'intrication sur la base d'une analyse spectrale d'un réseau d'interaction des termes. Soit D une collection de documents $d \in D$, chacun indexé par des termes $t \in T$, où T denote une collection de termes. Les *termes* ici *indexent* les documents et correspondent à des mots issus d'un vocabulaire contrôlé (thésaurus) ou extraits des documents. Nous supposons ici que les termes ont déjà étés identifiés et/ou déterminés, ainsi tous les documents sont associés à un jeu de termes. Soit $M = (m_{d,t})_{d \in D, t \in T}$ la matrice d'occurrence, avec $m_{d,t}$ le nombre d'occurrences du terme t dans le document d. Le document d peut alors être vu comme un vecteur de poids indexés par les termes $t \in T$, soit, $d = (m_{d,t})_{t \in T}$ correspond à une ligne dans la matrice d'occurrence M.

2.1 Réseau d'interaction des termes

On peut aussi définir les relations documents-termes avec une représentation sous forme de graphes. La matrice d'occurrence correspond en effet à un graphe $G_{D,T} = (V, E)$, dont les noeuds sont soit des documents soit des termes, $V = D \cup T$ et les arêtes $e = \{d, t\} \in E$ connectent les documents aux termes. Ce graphe est de toute évidence *biparti*, puisque les arêtes ne connectent jamais deux documents ou deux termes directement. La figure 1 illustre cette construction depuis un jeu de 4 documents différents indexés par des termes (1 (a)). La figure 1 (b) correspond au graphe biparti defini à partir de ces documents et termes.

Les arêtes $e \in E$ peuvent être pondérées, $\omega : E \to \mathbb{R}$. Une fonction de poids évidente serait $\omega(e) = m_{d,t}$. Si les documents sont indexés par des termes générés par LDA, le poids $\omega(e) = P(t|d)$ pourra être la probabilité donnée par le modèle. N'importe quelle autre fonction de poids résultant de l'indexation des documents peut aussi être utilisée. La littérature présente plusieurs techniques et algorithmes permettant d'exploiter ce graphe biparti afin de rechercher ou segmenter la collection de documents [Xu et al. (2003)], le jeu de termes [Slonim et Tishby (2000)] ou bien les deux en même temps [Dhillon (2001)].

Ce graphe biparti est utilisé pour dériver le graphe $G_D = (D, E_D)$, reliant directement les documents. Le graphe est construit à partir de $G_{D,T}$ en projetant les chemins $d - t - d'$ sur les arêtes $e = \{d, d'\} \in E_D$. La figure 1 (c) illustre comment G_D est obtenue à partir de



(a) Documents D indexés par les termes T

(b) Graphe biparti $G_{D,T} = (D \cup T, E)$

(c) Graphe document-document $G_D = (D, E_D)$

(d) Réseau d'interaction des termes $I_T = (T, F)$

FIG. 1 – *A partir d'une collection de documents indexés par des termes (a), nous construisons un graphe biparti liant documents et termes (b). Nous considérons ensuite le graphe document-document avec les termes en attributs des arêtes (c) et en dérivons le* réseau d'interaction des termes *(d) (ici pondéré par le nombre d'occurrences et co-occurrences des termes dans G_D).*

$G_{D,T}$ (il *n'inclue aucune boucle* reliant un document à lui-même). L'arête $e = \{d, d'\}$ entre les documents d et d' sur $G_{D,T}$ est induite s'ils partagent les termes distincts $t, t', \ldots$. Ces termes deviennent des attributs *associés* à l'arête e. Notre analyse portant sur l'interaction des termes, nous ne considérons dans le graphe G_D qu'uniquement les arêtes partageant *au moins 2 termes distincts*.

Nous construisons maintenant un *réseau d'interaction des termes*. Considerons le graphe $I_T = (T, F)$, avec pour noeuds les termes $t \in T$. L'arête $f \in F$ relie les termes $t, t' \in T$ lorsqu'ils sont *ensemble associés* à l'arête $e = \{d, d'\} \in E_D$, c'est à dire, lorque deux documents distincts $d, d' \in D$ sont tous les deux indexés par les termes t, t'.

Le graphe $I_T = (T, F)$ n'est pas tout à fait construit à partir de $G_{D,T}$ en projetant les chemins $t - d - t'$ sur les arêtes $e = \{t, t'\}$, mais plutôt à partir des arêtes de G_D. La figure 1 (d) illustre comment I_T est obtenu à partir du graphe G_D. Le *réseau d'interaction des termes*, défini *après* que les documents ont été indexés, est notre principal objet d'intéret pour explorer l'espace des documents.

L'idée de la construction d'un réseau d'interaction des termes est empruntée à l'analyse des réseaux sociaux [Burt et Scott (1985)]. Nous calculons *les matrices d'interaction* $N_I = (n_{t,t'})$ et $C_I = (c_{t,t'})$ (où les indices correspondent aux termes $t, t' \in T$). Soit $e \in E_D$ une arête de G_D et $\tau(e) \subset T$ le jeu de termes associés à e. De la même manière, soit $\tau^{-1}(t)$ le jeu d'arête $e \in E_D$ ayant pour terme associé $t \in T$. Nous écrivons $n_t = |\tau^{-1}(t)|$ la cardinalité de ce

jeu. Nous définissons alors $n_{t,t'}$ le nombre d'arêtes $e \in E_D$ avec $\{t, t'\} \subset \tau(e)$, c'est à dire, $n_{t,t'}$ est égale au nombre d'arêtes $e \in E_D$ portant à la fois les termes t et t'. En d'autres mots, $n_{t,t'} = |\tau^{-1}(t) \cap \tau^{-1}(t')|$.

Définissons $c_{t,t} = \frac{n_{t,t}}{|F|}$, et $c_{t,t'} = \frac{n_{t,t'}}{n_t}$. Si la matrice N_I est symétrique, la matrice C_I elle, ne l'est pas. Les entrées de la diagonale $c_{t,t}$ dans la matrice C_I peuvent être considérées de manière non formelle comme la probabilité qu'une arête de E_D soit associée au terme t. Les autres entrées $c_{t,t'}$ correspondraient alors aux probabilités conditionnelles qu'une arête soit associée au terme t sachant qu'elle est associée au terme t'.

Considérons les matrices N_I (à gauche) et C_I (à droite) ci-dessous. Ces matrices sont déterminées à partir de la clique de 5 termes de la figure 2, section 2.3, indexant une collection de 18 documents partageant 103 liens. En regardant la diagonale, $n_{1,1} = 71$ on voit que les liens sont associés avec le premier terme (sécurité routière), et $n_{2,2} = 48$ avec le second (prévention des accidents). Le nombre de liens associés à la fois avec le premier et le second terme est donc $n_{1,2} = n_{2,1} = 35$. En regardant la première entrée de C_I, le premier terme est associé à $c_{1,1} = 69\%$ pour tous les liens, et il y a $c_{1,2} = 73\%$ de chance de trouver un lien associé avec le second terme parmi tous ceux associés au premier terme, lorsque seuls $c_{2,1} = 49\%$ des liens qui sont associés avec le second terme le sont aussi avec le premier.

$$\begin{bmatrix} 71 & 35 & 61 & 46 & 28 \\ 35 & 48 & 35 & 41 & 15 \\ 61 & 35 & 78 & 42 & 45 \\ 46 & 41 & 42 & 67 & 21 \\ 28 & 15 & 45 & 21 & 45 \end{bmatrix} \quad \begin{bmatrix} 0.69 & 0.73 & 0.78 & 0.69 & 0.62 \\ 0.49 & 0.47 & 0.45 & 0.61 & 0.33 \\ 0.86 & 0.73 & 0.76 & 0.63 & 1 \\ 0.65 & 0.85 & 0.54 & 0.65 & 0.47 \\ 0.39 & 0.31 & 0.58 & 0.31 & 0.44 \end{bmatrix}$$

2.2 Indice d'intrication

Nous souhaitons maintenant mesurer *l'indice d'intrication* soit la participation d'un terme t dans l'intrication globale du groupe. Cette notion d'intrication est adaptée directement d'une notion similaire, l'ambigüité des relations sociales [Burt et Scott (1985)]. Posons λ l'indice d'intrication maximal parmi tous les termes et γ_t la fraction d'intrication correspondant au terme t. L'indice pour le terme t peut alors être déterminé comme $\gamma_t \cdot \lambda$. Comme l'intrication d'un terme est renforcée par ses interactions avec d'autres termes intriqués, et comme les entrées $c_{t,t'}$ de la matrice C_I offrent une interprétation probabiliste, nous pouvons poser l'équation suivante définissant les valeurs γ_t.

$$\gamma_{t'} \cdot \lambda = \sum_{t \in T} c_{t,t'} \gamma_t \tag{1}$$

Le vecteur $\gamma = (\gamma_t)_{t \in T}$, collecte alors les valeurs pour tous les termes t, formant ainsi le vecteur propre *à droite* de la matrice transposée C'_I. A partir de l'eq. (1) nous pouvons lever l'équation matricielle $\gamma \cdot \lambda = C'_I \cdot \gamma$. L'indice maximal d'intrication sera égal à la valeur propre maximale de la matrice C'_I. Nous nous intéressons ainsi aux valeurs relatives γ_t. La section suivante présente à partir du vecteur γ_t et de la valeur λ *l'intensité d'intrication* et *l'homogénéité d'intrication* comme mesures globales de l'intrication d'un graphe.

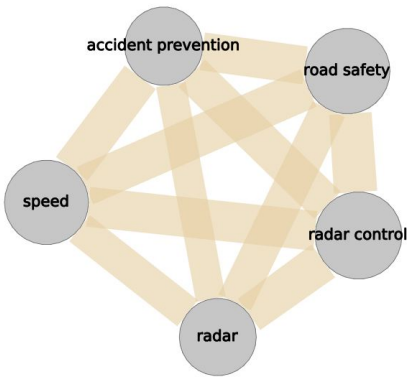


FIG. 2 – *Un groupe de documents avec une intrication optimale a pour réseau d'interaction des termes une clique où chaque terme interagit avec les autres identiquement.*

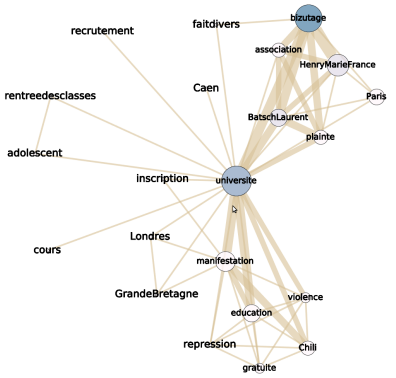


FIG. 3 – *Un réseau d'interaction sous forme d'étoile regroupe (a) un(des) terme(s) central(aux), avec un(des) indice(s) d'intrication plus élevé. Les termes périphériques peuvent former des sous-groupes denses avec des indices d'intrication* localement *plus élevés.*

2.3 Profils d'intrication

La topologie du réseau d'interaction des termes $I = (T, F)$ fournit une information complémentaire sur la façon dont les termes contribuent à l'intrication sémantique globale d'un groupe ou sous-groupe de documents. L'attention se porte ici sur les interactions entre les termes au sein d'une collection

L'archétype d'un *groupe de documents optimal cohésif* est une configuration dans laquelle tous les documents sont indexés exactement par les mêmes termes (et par conséquent tous les termes sont intriqués entre eux de manière maximale). Le graphe $I = (T, F)$ correspond alors à une *clique*. Dans ce cas, toutes les entrées $n_{t,t'}$ de la matrice N_I coïncident, et toutes les entrées de la matrice C_I valent 1. La valeur propre maximale C'_I est dans ce cas égale à $\lambda = |F|$, et tous les indices γ_t coïncident. La théorie de Perron-Frobenius sur les matrices non-négatives (Ding et Zhou, 2009, Chap. 2) montre en plus que $\lambda = |F|$ est la valeur propre maximale possible pour une telle matrice avec ses entrées comprises dans l'intervalle $[0, 1]$.

La situation opposée est celle où le réseau d'interaction des termes n'est pas connexe. Les termes se séparent en plusieurs sous réseaux qui n'interagissent jamais entre eux. C'est une force de la représentation du réseau de pouvoir identifier directement les composantes connexes (qui doivent être analysées chacune indépendamment). En effet, lorsque $I = (T, F)$ est non connexe, la matrice C_I est considérée comme *réductible* ; à l'inverse, lorsque $I = (T, F)$ est connexe, la matrice C_I est *irreductible*. Dans ce cas, la théorie des matrices non négatives nous apprend que la matrice C_I a une valeur propre maximale $\lambda \in \mathbb{R}$ avec un seul vecteur propre associé γ_t. Ce vecteur propre a des entrées réelles et non négatives [(Ding et Zhou, 2009, Theorem 2.6)]. Nous considérons par la suite que C_I est irréductible.

Un autre cas intéressant apparaît typiquement lorsque quelques termes sont centraux et le

reste des termes périphériques : d'une part, tous les documents partagent quelques termes en commun, se regroupant donc autour de quelques termes centraux ; d'autre part, ces documents forment des sous-groupes autour de termes ou sous-sujets secondaires. Le réseau d'interaction des termes présente alors une structure en forme d'étoile (figure 3). L'indice d'intrication est plus élevé pour les termes centraux alors que les autres termes ont des indices bien plus faibles (les noeuds de la figure 3 sont colorés en fonction de leur indice d'intrication). Les termes périphériques peuvent aussi former des sous-réseaux plus denses entre eux. Lorsque l'on calcule les indices d'intrication restreints à ce sous-réseau de termes, les valeurs se rapprochent de celles du scénario de la clique. Les études de cas présentées ci-dessous développent cette situation.

Nous pouvons déterminer l'intrication au niveau du groupe de document par comparaison avec les valeurs de l'archétype de la configuration optimale. Nous avons vu auparavant que la valeur propre maximale est délimitée par $|F|$, ainsi le ratio $\frac{\lambda}{|F|} \in [0,1]$ mesure combien intense est l'interaction des termes dans un groupe de documents. Ce ratio nous fournit une mesure de *l'intensité d'intrication* parmi tous les documents, *sous considération du jeu de termes* T.

De la même manière, lorsque les poids $c_{t,t'}$ sont identiques nous obtenons un vecteur propre avec des entrées identiques. Ce vecteur propre détermine alors l'espace diagonal généré par le vecteur diagonal $1_T = (1, 1, \ldots, 1)$. Ceci motive la définition d'une seconde mesure liée à l'homogénéité de la distribution de l'intrication parmi les termes. Nous déterminons ainsi une similarité cosinus $\frac{\langle 1_T, \gamma \rangle}{\|1_T\| \|\gamma\|} \in [0,1]$ qui exprime la proximité du groupe de document avec la situation d'intrication optimale. Nous l'appellerons *l'homogénéité d'intrication*.

3 Etude de cas

Cette section présente deux cas d'usage illustrant l'utilisation des mesures d'intrication et du réseau d'interaction des termes pour l'exploration d'un groupe de documents.

Chacun de ces cas d'usage a été construit à partir des extraits de journaux télévisés (JTs) couvrant plusieurs sujets sur une période de 100 jours. Les documents ont été manuellement indexés par l'INA. Les groupes de documents ont été identifiés en utilisant des approches de clustering classiques (ces approches n'étant pas le sujet de ce papier).

3.1 La sécurité routière et les radars

Nous considérons tout d'abord un jeu de 20 documents, tous liés à la sécurité routière. Bien que petit, cet ensemble de documents présente des caractéristiques intéressantes. La sécurité routière est devenue un sujet d'intéret suite à la promotion par le gouvernement des radars automatiques, impliquant une augmentation du nombre d'amendes. Ce sujet a bien sûr été abordé par chacun des JTs. Les documents ont été par la suite annotés avec des termes tels que *arrestation, automobiliste, conduite, danger, prévention des accidents, radar, sécurité routière, société, vitesse, etc.* La figure 4 montre le réseau d'interaction des termes en résultant.

Les noeuds les plus sombres ont un indice d'intrication plus élevé. La taille des noeuds correspond au nombre de liens associés au terme *dans le graph* G_D. Le dessin du graphe nous montre que les noeuds centraux *prévention des accidents* et *vitesse* ont des indices d'intrication plus élevés (respectivement 0.38 et 0.44). L'intensité d'intrication pour tout le réseau $I = (T, F)$ est de $\lambda/|F| = 0.33$, et l'homogénéité (similarité cosinus 2.3) est de 0.81.

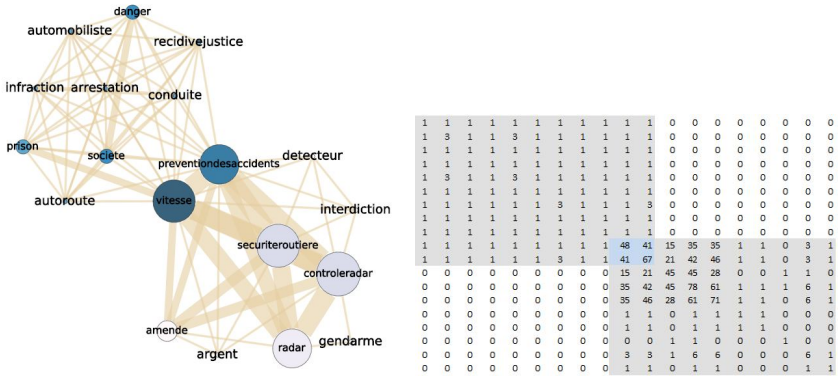


FIG. 4 – *Réseau d'interaction des termes déterminé à partir d'un jeu de documents liés à la sécurité routière et aux radars. Le réseau se divise en deux composantes organisées autour des termes centraux* prévention des accidents *et* vitesse*, ainsi menant à une évidente structure en blocs de la matrice d'interaction (une fois ordonnée).*

La figure 4 montre que les termes se divisent en deux groupes indiquant pourquoi l'intrication sémantique maximale ne peut pas être observée. La matrice d'interaction N_I présente de même une structure en blocs (en grisé), avec des valeurs hors diagonale nulles. Les termes centraux sont dans la partie bleue de la représentation matricielle, interagissant avec tous les autres termes. La partie supérieure de la matrice correspond à la partie supérieure du réseau d'interaction et montre que tous ces termes interagissent ensemble de manière peu fréquente (c.a.d. les termes indexent un petit sous-ensemble de tous les documents). La partie inférieure de la matrice présente un comportement complètement différent, où cinq termes interagissent de manière plus intensive ensemble, mais pas avec tous les autres termes de la composante.

La topologie du réseau suggère d'examiner de plus près les documents liés aux termes de la partie inférieure, positionnés sous les termes centraux *prévention des accidents* et *vitesse*. Nous considérons un sous-graphe I' formé à partir des termes *amende*, *gendarme*, *radar*, *sécurité routière*, etc. Les termes de I' indexent tous les documents à l'exception d'un seul. Pour ce sous-réseau I', nous mesurons une intensité de 0.31 et une homogénéité de 0.72, ce qui est légèrement inférieur aux mesures du réseau complet I. Ces valeurs sont plus faibles car beaucoup de termes comme *amende*, *argent* et *detecteur* n'indexent que peu de documents (comme leur taille le suggère) et les termes de I' se distribuent de manière inégale sur les arêtes entre documents en comparaison de la distribution de tous les termes de I (comme suggéré par les entrées nulles de la partie en bas à droite de la matrice). Notons que la fonction $\cos(-)$ est non linéaire pour bien interpréter toute variation dans la mesure d'homogénéité.

Nous pouvons maintenant nous concentrer sur la clique de 5 noeuds de la figure 2 (section 2.3) associée à 18 des 20 documents. Comme prévu elle atteint des valeurs d'intrication plus élevées avec 0.6 en intensité et 0.98 en homogénéité, ce qui confirme que ces 18 documents forment bien un tout cohésif autour de ces 5 termes.

Nous terminons ce premier exemple en regardant la partie supérieure du réseau, formée par les termes positionnés au dessus des termes centraux *prévention des accidents* et *vitesse*. Nous

considérons un sous-graphe composé des termes *danger*, *automobiliste*, *autoroute*, *prison*, etc. Nous obtenons le sous-réseau I'' en haut de la figure 4, où les termes indexent 19 des 20 documents originaux. En restreignant l'analyse à ce sous-réseau I'', l'intensité et l'homogénéité d'intrication sont à 0.57 et 0.98. Après suppression des termes centraux *prévention des accidents* et *vitesse*, les mesure d'intensité et d'homogénéité montent à 0.71 et 0.99. Nous trouvons ainsi que ces documents forment une unité cohésive sémantiquement formant incidentellement une clique avec des poids d'interaction inégaux $n_{t,t'}$. Cette conclusion doit malgré tout être modérée parce que les termes de ce second sous-réseau ne concernent que 4 documents. On constate ici que de fortes valeurs en intensité et en homogénéité sont bien plus faciles à atteindre sur des ensembles réduits de documents et de termes.

3.2 À propos des profils d'intrication

Nous retournons maintenant sur la notion de profils d'intrication en considérant l'exemple discuté précédement. Nous avons utilisé l'intensité $\frac{\lambda}{|F|}$ et l'homogénéité $\frac{\langle 1_T,\gamma\rangle}{\|1_T\|\|\gamma\|}$ comme deux mesures distinctes afin de fournir une information complémentaire au réseau d'interaction des termes. Cet exemple montrent des situations où ces mesures varient beaucoup. Bien que l'on suspecte que ces quantités ne varient pas indépendament, nous pouvons néanmoins placer ces variables dans un plan où l'intensité serait notée sur l'axe x et l'homogénéité sur l'axe y (figure 5). N'importe quel réseau d'interaction peut alors être placé comme un point sur ce plan $(x, y) = (\frac{\lambda}{|F|}, \frac{\langle 1_T,\gamma\rangle}{\|1_T\|\|\gamma\|})$, ainsi nous pouvons diviser ce plan en zones correspondant à des profils différents. Une supposition naïve est de diviser ce plan en 4 zones plus ou moins rectangulaires (lignes en pointillés). C'est assez loin de la réalité, on suspecte que les véritables zones suivent des formes plus complexes et le problème reste à étudier plus amplement, mais nous observons malgré tout 4 catégories.

La clique qui était présentée comme l'archétype du réseau d'interaction optimal se situe dans la partie la plus en haut à droite du plan $(x, y) = (1, 1)$. La zone supérieure droite correspond aux réseaux relativement denses et homogènes. La clique de 5 noeuds du "Radar" de la figure 2 appartient à cette catégorie de profil tout comme le sous réseau supérieur de la "Sécurité routière" (figure 4, section 3.1).

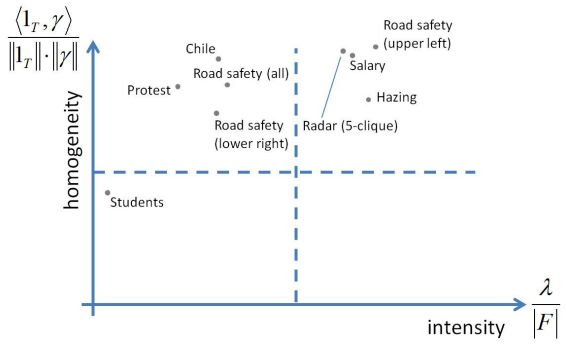


FIG. 5 – *Les profils d'intrication peuvent être grossièrement catégorisés en combinant intensité d'intrication et homogénéité d'intrication pour identifier 4 zones critiques.*

La partie supérieure gauche correspond à une homogénéité relativement élevée avec une intensité relativement faible : les termes interagissent pratiquement tous entre eux, mais pas autant que le graphe des documents G_D ne l'autorise en théorie. Les matrices N_I ne sont pas creuses, et ont une diagonale plutôt large avec des entrées hors diagonales plus faibles. La partie inférieure du réseau sur la "Sécurité Routière" (Figure 4), en est un bon exemple.

La situation de la partie inférieure droite apparaît lorsque les termes sont imbriqués un peu comme s'ils exprimaient des concepts similaires à différents niveaux de généralité. Cette situation se traduit par des inclusions consécutives des termes dans les arêtes des documents (c.a.d. les liens de G_D associés avec un terme t incluent tous les liens associés au terme t' avec en plus quelques autres liens).

La partie inférieure gauche rassemble des réseaux avec de faibles intensités et homogénéités d'intrication. C'est un cas commun qui bien souvent rassemble des documents et des termes avec peu d'interaction. Dans cette situation, beaucoup de termes sont satellites de termes centraux, avec parfois plusieurs centres. Un jeu de termes couvrant un paysage sémantique large induit inévitablement ce genre de situation. Un réseau typique de cette situation présente une faible densité (peu d'arêtes), menant à une matrice N_I creuse avec des entrées d'ordre ϵ. Le réseau d'interaction de départ de notre exemple tombe dans cette catégorie.

Bien que ces zones nous fournissent une grille intéressante pour l'évaluation des profils de réseaux, nous avons besoin de plus d'expérimentations pour confirmer ces catégories et déterminer les limites de ces zones et pour estimer de quelle manière elles sont peuplées.

4 Étude utilisateur

Notre objectif est de fournir à l'utilisateur des outils pour évaluer plus rapidement et avec plus d'assurance le contenu d'une collection de documents. Nous avons testé nos outils auprès de 4 documentalistes experts (2 séniors et 2 juniors) de l'INA. Bien que restreinte, cette expérience a été conçue autour d'un protocole strict : 4 jeux de notices documentaires de reportages de JT concernant 4 évènements médiatiques ont été générés avec des techniques de clustering classique, avec des tailles et des valeurs d'intensités et d'homogénéités d'intrication variées.

Les utilisateurs disposaient de deux interfaces : la première propose une liste de documents permettant d'inspecter librement les titres, leur contenu et leurs termes d'indexation ; la seconde interface proposait une représentation interactive et synchronisée du graphe de documents et du réseau d'interaction des termes précédemment introduit. Les indices d'intrications sont calculés sur les sous-graphes sélectionnés par l'utilisateur.

Les utilisateurs avaient 10 minutes pour se familiariser avec la tâche et les interfaces. Les jeux de données et les interfaces ont été distribués aléatoirement de manière à éviter les biais. Un entretien et un questionnaire clôturaient cette évaluation, d'une durée moyenne de 2h30.

Les tâches demandées aux utilisateurs étaient les suivantes :

- évaluer la cohésion sémantique globale de chacun des groupes de documents ;
- éliminer les documents "bruitant" une collection afin de renforcer la cohésion sémantique globale de la collection et indiquer les documents qui leur apparaissaient mal indexés (mauvais termes) ;
- dans une collection donnée, trouver des documents correspondant à une requête donnée ;
- raconter l'histoire expliquant le contenu d'une collection de documents (qui devrait alors plus ou moins correspondre avec l'évènement couvert par les JT) ;

– exprimer leur confiance dans leur analyse (par exemple, avoir écarté les *bons* documents, et raconté la *bonne* histoire).

La compilation des interviews réalisées rapportent les observations suivantes : les utilisateurs ont apprécié le réseau d'interaction pour son utilité à identifier/discriminer les différents termes, pour la concision de sa représentation. Ils ont confirmé son utilité quant à la compréhension d'une collection de documents dans son ensemble. Les commentaires recueillis lors des entretiens révèlent que les utilisateurs pouvaient développer une bonne intuition du type de collection de documents à partir de la forme du réseau, et utiliser cette intuition pour identifier ses caractéristiques importantes (comme les termes centraux et périphériques, le découpage d'un thème en plusieurs petites histoires). Les utilisateurs ont aussi confirmé le lien entre la cohésion sémantique perçues et les mesures d'intrication retournées. Les documents annotés par des termes de faible indice d'intrication correspondent bien à des documents " à la limite " des sujets principaux des jeux et cet indicateur est donc très pertinent. Finalement, les utilisateurs ont trouvé que l'utilité apportée par le réseau d'interaction des termes grandissait avec l'augmentation de la taille du jeu, ce que semblent confirmer nos premières mesures de temps (bien qu'on ne puisse le dissocier pour le moment d'un effet d'apprentissage de l'interface).

5 Conclusion

Ce papier présente des outils facilitant l'exploration et l'évaluation approfondies du contenu sémantique d'une collection de documents. Deux mesures d'intrication et un graphe de termes associés à la collection sont introduits pour permettre une exploration interactive récursive des données. L'étude de cas (section 3) montre l'apport du graphe d'interaction, dont la topologie exprime certaines caractéristiques sémantiques de la collection analysée. Les mesures d'intensité et d'homogénéité d'intrication permettent des analyses plus complètes qu'une simple proximité sémantique et offrent à l'utilisateur une nouvelle appréciation sur la sémantique de la collection ou d'une sous collection sélectionnée. Ces mesures nous ont permis de distinguer 4 profils génériques d'intrication sémantiques. L'étude utilisateur confirme l'intérêt de ces approches pour une étude de contenu détaillée d'une collection de documents.

Les tests ont été effectués sur des échantillons de petite taille, limités à quelques centaines de documents. Néanmoins, une analyse humaine approfondie ne peut s'exercer sur des collections de milliers de documents à la fois. Cet outil trouve sa place en complément de procédure d'agrégation à grande échelle, dans des contextes de qualité d'accès aux collections spécifiques, comme les bibliothèques ou les centres documentaires.

Nos travaux futurs porterons sur l'intégration de notre prototype dans un contexte d'usage réel afin de mesurer les retours utilisateurs dans une situation d'exploitation réaliste. Les mesures d'intrication étant génériques, elles peuvent s'appliquer à tous types de réseau mais avec une interprétation spécifique. Nous effectuons actuellement des tests sur des données issues des échanges collaboratifs.

Références

Arun, R. et al (2010). On finding the natural number of topics with latent dirichlet allocation : Some observations advances in knowledge discovery and data mining. Volume 6118 of

Lecture Notes in Computer Science, pp. 391–402. Springer.

Beaza-Yates, R. et B. Ribeiro-Neto (1999). *Modern Information Retrieval*. ACM Press.

Blei, D., A. Ng, et M. Jordan (2003). Latent dirichlet allocation. *Journal of Machine Learning Research 3*(5), 993–1022.

Brin, S. et L. Page (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems 30*(1-7), 107–117.

Burt, R. et T. Scott (1985). Relation content in multiple networks. *Social Science Research 14*, 287–308.

Croft, B., D. Metzler, et T. Strohman (2009). *Search Engines : IR in Practice*. Addison Wesley.

Deerwester, S., S. Dumais, G. W. Furnas, T. K. Landauer, et R. Harshman (1990). Indexing by latent semantic analysis. *Journal of the Society for Information Science 41*(6), 391–407.

Dhillon, I. S. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. In *7th ACM SIGKDD*, San Francisco, California, pp. 269–274. ACM.

Ding, J. et A. Zhou (2009). *Nonnegative Matrices, Positive Operators and Applications*. Singapore : World Scientific.

Dumais, S. T. et al (1988). Using latent semantic analysis to improve access to textual information. In *SIGCHI*, pp. 281–285. ACM.

Kohonen, T. et al. (2000). Self organization of a massive document collection. *Neural Networks, IEEE Transactions on 11*(3), 574 –585.

Salton, G. (1983). *Automatic Text Processing : The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley.

Slonim, N. et N. Tishby (2000). Document clustering using word clusters via the information bottleneck method. In *23rd ACM SIGIR*, SIGIR '00, pp. 208–215. ACM.

Thomas, H. (1999). Probabilistic latent semantic indexing. In *22nd ACM SIGIR*, Berkeley, California, United States, pp. 50–57. ACM.

Xu, W., X. Liu, et Y. Gong (2003). Document clustering based on non-negative matrix factorization. In *26th Annual International ACM SIGIR*, SIGIR '03, pp. 267–273. ACM.

Zhao, Y., G. Karypis, et U. Fayyad (2005). Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery 10*, 141–168.

Summary

With the booming growth of online resources, efficiently exploring document collections has become a daily task for most users, and yet still remains a challenge for research. In this article, we present three tools to help users perceive the semantic content of a set of documents : a weightened interaction network of terms associated with documents, as well as intensity and homogeneity measures of semantic intrication reflecting the main features of term distribution within a set of documents. These measures can also help to identify different types of set of documents.