Analyse du graphe des cooccurrents de deuxième ordre pour la classification non-supervisée de documents

Aurélien Lauf[1], Mathieu Valette[2], Leila Khouas[3]

[1] ERTIM, INALCO / AMI Software France – alu@amisw.com
[2] ERTIM, INALCO – mvalette@inalco.fr
[3] AMI Software France – lkh@amisw.com

Abstract

This paper is in the context of strategic and competitive intelligence on the Web (medium sized corpora). We propose a linguistic approach for document clustering based on the analysis of a second-order co-occurrences graph. The topic we build can overlap (i.e. a word can be part of more than 1 topic – polysemic words, homographs, etc.) and only include the strongest words. Because of graph theory formalism, we are able to express subtle semantic relations between words within each topic, which are thus not only sets of words. Using these words, we are then able to assign one or more topics to each document.

Résumé

Cette étude s'inscrit dans un contexte concret de veille d'entreprise sur le Web (corpus de taille moyenne). Nous proposons dans ce papier une approche linguistique de classification non supervisée de documents reposant sur l'analyse du graphe des cooccurrents de deuxième ordre (cooccurrents des cooccurrents). La classification est non exhaustive (un mot peut n'appartenir à aucune thématique) et multiclasse (un mot peut appartenir à plusieurs thématiques – polysémie, homographie, etc.). Les thématiques obtenues ne sont pas uniquement des ensembles de mots : le formalisme de la théorie des graphes nous permet d'exprimer concrètement des relations sémantiques fines entre les mots de chaque thématique. Ces mots nous permettent enfin d'assigner à chaque document une ou plusieurs thématiques.

Mots-clés : classification non supervisée, statistiques textuelles, linguistique de corpus, lexicométrie, cooccurrence, théorie des graphes, veille, internet.

1. Introduction

Nos travaux s'inscrivent dans un cadre de veille et d'intelligence économique. Notre objectif est d'assister le veilleur dans deux tâches : 1. dégager des thématiques du corpus (aide à la lecture et à l'interprétation) ; 2. ranger chaque texte dans une ou plusieurs de ces thématiques afin de faciliter le tri et le retour au texte. Cette tâche ouvre par ailleurs la voie à des traitements plus fins, comme l'analyse de l'évolution des thématiques dans le temps.

La littérature dans le domaine est abondante. Nos travaux se démarquent par deux points. Tout d'abord, nous sommes dans un cadre concret de veille sur Internet, ce qui implique certaines

contraintes bien particulières. Par ailleurs, sans remettre en cause la qualité des modèles statistiques, nous souhaitons présenter une alternative linguistique, plus fortement ancrée dans les traditions de la lexicométrie et de la linguistique de corpus ; la méthode que nous proposons permet l'obtention de thématiques qui ne sont pas uniquement des ensembles de mots : le formalisme de la théorie des graphes nous permet d'exprimer des relations sémantiques assez fines entre les mots de chaque thématique.

Dans un premier temps, nous procèderons à un bref rappel sur les *topic models* ainsi que les principaux algorithmes de *clustering*, afin de situer nos travaux. Nous présenterons ensuite notre approche, ainsi que les travaux dont elle s'inspire, avant de décrire le corpus sur lequel nous travaillons et les contraintes qu'il impose. Enfin, nous commenterons nos résultats, puis proposerons les perspectives pour la suite de nos travaux.

2. Etat de l'art et positionnement

Certaines approches statistiques sont fréquemment utilisées pour cette tâche, par exemple la NMF (*Non-negative Matrix Factorization* – (Lee et Seung, 1999)), la *Latent Semantic Analysis* (LSA – (Deerwester, 1990)), ou encore les *topic models*. La *Latent Dirichlet Allocation* (LDA – (Blei *et al.*, 2003)), qui fait suite à la *Probabilistic Latent Semantic Indexing* (PLSI – (Hofmann, 1999)), est le *topic model* le plus connu.

Notre étude s'inscrit dans un cadre concret de veille et d'intelligence économique à partir de données issues du Web. Ce cadre implique des contraintes bien particulières : 1. la nécessité de considérer des corpus de taille moyenne (entre 100 000 et 500 000 mots) ; ces derniers sont réputés trop petits pour les méthodes statistiques, mais trop grands pour être analysés manuellement ; 2. les thématiques que l'on cherche à extraire peuvent être très proches les unes des autres et risquent de partager un nombre conséquent de mots : il ne s'agit en effet pas ici d'opposer des documents médicaux et juridiques par exemple, mais de dégager des thématiques toutes relatives à un même sujet général (ici le nucléaire).

Cependant, la plus grande originalité réside dans la façon d'aborder le problème. Nous cherchons à formaliser la notion de thématique avec un point de vue linguistique, avec un algorithme le plus respectueux possible de certaines propriétés qui, selon nous, font sens d'un point de vue linguistique (Pincemin, 1999) :

1. Le sens n'est pas dans les mots, mais entre les mots. Nous faisons donc l'hypothèse que des thématiques peuvent être modélisées par des regroupements de mots (substantifs pour le moment) apparaissant dans des contextes similaires : des cooccurrences ; la classification repose ainsi principalement sur la répartition des mots dans le corpus, et moins sur leur fréquence.
2. L'algorithme doit être capable de se mettre facilement à l'échelle et être indépendant de la configuration des regroupements de mots. Ce point est important car nous avons constaté que les clusters peuvent grandement varier en taille et/ou en densité selon leur représentativité dans le corpus.
3. Un mot peut n'appartenir à aucune thématique. L'attribution systématique d'un cluster à chaque mot risque de faire perdre en cohérence certains regroupements : on ne s'intéresse qu'aux liens les plus forts.

4. Un mot peut appartenir à plusieurs thématiques à la fois, ce qui peut par exemple traduire une relation de polysémie, d'homographie, ou toute nuance de sens plus fine. Ce point est particulièrement important étant donné la forte intersection lexicale entre les différentes thématiques de notre corpus.

Enfin, à notre connaissance, la notion de thématique n'a jamais été réellement formalisée. Les *topic models*, par exemple, considèrent qu'un *topic* est un ensemble de mots pondérés par des probabilités d'apparition. En prenant le parti de recourir à une approche de classification non supervisée sur des graphes de cooccurrences, nous espérons formaliser la notion de cohérence sémantique de façon plus poussée. En effet, le formalisme de la théorie des graphes[1] nous permet d'exprimer concrètement des relations pondérées entre les mots de chaque thématique ; en d'autres termes, un mot *m1* peut entretenir des liens forts avec les mots *m2* et *m3* mais quasiment nuls avec le reste de la thématique. Par exemple, nous constatons au sein de la thématique autour de Tchernobyl (voir section 5 – Résultats et discussion) que les mots *mémoire*, *monument, victime*, *bougie, mort* entretiennent des liens privilégiés entre eux, tandis que le mot *radioprotection* est naturellement plus proche de *particule*, ou de *césium*. Le type de relations sémantiques symbolisées par les arêtes du graphe reste encore à définir. Des expériences complémentaires devraient nous aider à qualifier plus précisément ces relations.

3. Description de notre approche

Avant de commencer, nous précisons que nous nous refusons tout recours à des ressources sémantiques extérieures afin de qualifier les relations entre les mots (dictionnaires, thésaurus, ontologies – Wordnet par exemple) car elles sont difficilement applicables sur le Web (voir notamment à ce sujet (Slodzian, 2000)). Nous nous basons sur les cooccurrents des mots du corpus car nous estimons qu'ils représentent la forme minimale du contexte (Mayaffre, 2008), et donc du sens (Rastier, 1987). Les regroupements se font donc de façon dynamique et dépendent uniquement du corpus analysé, non de relations universelles définies en amont ; notre approche reste indépendante du domaine, ce qui est crucial lorsque l'on travaille sur le Web. Par ailleurs, cette approche est moins sensible à la fréquence des mots (usages rares) que les méthodes utilisant des mots isolés.

Les algorithmes cumulant une analyse non exhaustive ainsi que la possibilité d'attribuer plusieurs classes à un même élément sont rares. La *Clique Percolation Method* (Palla *et al.*, 2005) répond à ces deux exigences, mais l'utilisation de la notion de cliques (certes légèrement allégée) rend la méthode trop rigide : certains regroupements denses (mais pas assez pour former des cliques) peuvent ainsi être ignorés.

L'algorithme SNN (*Shared Nearest Neighbours)* de (Jarvis et Patrick, 1973) a retenu notre attention car il présente de nombreuses caractéristiques appréciables. Les auteurs évaluent la similarité entre deux points à l'aide du nombre de plus proches voisins que ces derniers partagent ; dès lors, on ne calcule pas uniquement la similarité entre tous les points pris deux à deux (ce type de relations binaires retranscrit mal les interactions linguistiques complexes), mais on met l'accent sur des regroupements de points considérés simultanément. Par ailleurs,

1 (Ferrero i Cancho et Solé, 2001) ont montré que les graphes lexicaux partagent les caractéristiques des graphes « petit monde » (Watts et Strogatz, 1998). Ces régularités structurelles pourraient être la preuve d'une organisation linguistique et cognitive sous-jacente.

le *clustering* reste ainsi indépendant de l'échelle et de la configuration du graphe : on considère le plus proche voisinage, indépendamment de sa densité et de sa taille. Enfin, SNN procède par éclaircissements successifs d'un graphe de similarité (filtrage des relations les moins importantes), afin de ne conserver que les regroupements les plus pertinents. En ce sens, le *clustering* n'est pas exhaustif.

Les trois premiers critères parmi les quatre cités plus haut sont jusqu'ici respectés, mais le chevauchement de clusters n'est pas permis, ce qui n'est pas acceptable pour notre corpus car des mots multiclasses tels que *centrale* ou *EDF* conduiraient à la formation de clusters englobant la quasi-totalité du graphe. On pourrait neutraliser ces mots mais les détecter n'est pas toujours trivial. SNN sera repris par (Ertöz *et al.*, 2003) et (Ferret, 2006), qui vont l'appliquer sur des données textuelles, mais ces travaux partent toujours du postulat qu'un nœud ne peut appartenir qu'à un seul cluster uniquement.

Notre méthode reprend l'algorithme de (Jarvis et Patrick, 1973) et ajoute quelques étapes afin de permettre des regroupements plus stables et de prendre en compte la possible appartenance d'un mot à plusieurs thématiques.

Nous créons une matrice C $v \times v$ de cooccurrences, tel que $C_{ij} = Freq_{i,j}$, sachant que v correspond au nombre de vocables du corpus et que $Freq_{i,j}$ est le nombre de fois que les mots aux indices i et j apparaissent ensemble dans un même contexte. Nous choisissons le paragraphe comme fenêtre d'analyse afin de mettre l'accent sur des relations à longue distance et ainsi éviter les éventuelles « interférences » syntaxiques[2]. Les relations de cooccurrence ne sont pas orientées, car cela n'a réellement de sens que dans les contextes plus petits (la phrase par exemple). Les mots apparaissant moins de 10 fois dans l'ensemble du corpus sont ignorés afin de réduire la combinatoire. Ces valeurs sont ensuite remplacées par la mesure de dissimilarité présentée dans (Véronis, 2003) :

$$W_{A,B} = 1 - \max[p(A|B),p(B|A)]$$

$p(A|B)$ correspond à la probabilité conditionnelle d'observer A dans le même contexte que B, et inversement pour $p(B|A)$; $p(A|B) = Freq_{A,B} / Freq_B$ et $p(B|A) = Freq_{A,B} / Freq_A$. Le score est compris entre 0 (association systématique) et 1 (jamais d'association entre les mots). Dans l'avenir, nous envisageons d'évaluer l'impact de cette mesure sur les résultats, en la comparant avec d'autres mesures (l'information mutuelle de (Church et Hanks, 1990) par exemple).

Cette matrice forme un graphe de cooccurrences où les nœuds sont les mots, et les arêtes représentent les relations (non nulles) de cooccurrences entre ces mots, pondérées selon la formule vue précédemment. On y applique l'algorithme SNN de (Jarvis et Patrick, 1973), qui ramène le problème du *clustering* à celui de la détection de composantes comparativement denses dans un graphe. Pour ce faire, SNN procède par éclaircissements successifs des arêtes les moins fortes, jusqu'à briser le graphe en plusieurs composantes connexes[3] ; chacune de ces composantes connexes correspond à un cluster. Plus précisément, l'algorithme se divise en trois grandes étapes : 1. filtrage de tous les liens sauf des plus forts : on obtient le graphe

2 Nous préférons le paragraphe à une fenêtre glissante de n mots car il s'agit d'un découpage humain qui, a priori, fait sens. L'échelle du document nous semble trop large. Néanmoins, la question de la délimitation du contexte reste encore ouverte.

3 Dans un graphe non orienté G, une composante connexe est un sous-graphe dans lequel il existe un chemin entre toute paire de nœuds.

des plus proches voisins, qui représente les cooccurrents de premier ordre (liste des voisins directs d'un mot – (Grefenstette, 1994)) ; 2. remplacement du poids des arêtes par le nombre de voisins que les nœuds ont en commun afin d'obtenir le graphe des plus proches voisins partagés. A cette étape, il est possible de créer de nouveaux liens (par exemple : les mots *A* et *B* qui ne sont à l'origine pas reliés vont le devenir par transitivité s'ils partagent le voisin *C*). Cette étape est intéressante car elle permet une transition vers des cooccurrents de second ordre (mots partageant un même environnement). On passe donc de relations plutôt syntagmatiques à des relations paradigmatiques propices à des regroupements sémantiques ; 3. on procède à un filtrage des liens inférieurs à un seuil donné, fixé empiriquement à 5 ; les clusters correspondent aux composantes connexes du graphe simplifié ainsi obtenu. En d'autres termes, SNN considère que deux nœuds appartiennent au même cluster s'ils partagent un nombre suffisant de voisins. Cette définition est problématique dans des corpus comme le nôtre qui se caractérisent par un nombre conséquent de mots susceptibles d'appartenir à plusieurs thématiques ; le risque de fusionner plusieurs clusters à cause de ces nœuds est élevé : notre graphe est à ce stade encore connexe[4] et on obtient donc un seul gros cluster.

Nous proposons ci-dessous de nouvelles étapes simples afin de permettre des regroupements plus stables et de prendre en compte le chevauchement de clusters.

1. On remplace à nouveau le poids des liens par le nombre de voisins que les nœuds partagent, ce qui rapproche encore une fois par transitivité certains mots. Nous avons constaté que la répétition de cette étape rend la classification beaucoup plus robuste.
2. Il s'agit désormais d'extraire les composantes comparativement denses du graphe, et d'isoler les mots susceptibles d'appartenir à plusieurs thématiques. Pour ce faire, nous amendons légèrement la définition de SNN d'un cluster : nous considérons désormais que deux nœuds appartiennent au même cluster (deux mots appartiennent à la même thématique) s'ils partagent la majorité de leurs voisins respectifs. En d'autres termes, on ne compare plus uniquement le nombre absolu de voisins communs mais le nombre de voisins communs relativement au nombre total de voisins des nœuds en question, comme indiqué dans l'équation ci-dessous. Les arêtes vers les nœuds multiclasses sont ainsi pénalisées car ces derniers ont un nombre total de voisins largement supérieur au nombre de voisins qu'ils partagent avec chaque thématique prise séparément. Cela présente aussi l'avantage de normaliser le poids des arêtes.

$$\frac{C_{ij}^{\,2}}{(N_i - 1) \cdot (N_j - 1)}$$

C_{ij} correspond au nombre de voisins que partagent *i* et *j* ; N_i et N_j renvoient respectivement au nombre de voisins qu'ont les nœuds *i* et *j*. On retire 1 à chacune de ces valeurs car *i* ne peut évidemment pas partager *j* avec ce dernier. Ce score est compris entre 0 et 1.

3. On filtre les arêtes ayant un poids inférieur à un seuil donné. Ce dernier est choisi volontairement bas afin que la définition d'un cluster ne soit pas trop restrictive (risque de scinder à tort certaines thématiques). Nous avons constaté que *0.5* et *0.6* donnent en général les meilleurs résultats. Nos thématiques sont les composantes connexes de ce nouveau graphe ainsi obtenu.

4 Un graphe non orienté est connexe s'il existe un chemin entre toutes ses paires de nœuds.

4. On réintègre les arêtes n'ayant pas « survécu » à l'étape précédente. Un nœud appartient à un cluster supplémentaire s'il a des liens avec une majorité de mots d'un autre cluster (seuil fixé empiriquement à 80%). Cela permet aussi de fusionner certains petits clusters isolés à tort (quand tous les mots du cluster sont liés à un autre).

Ces 4 nouvelles étapes sont schématisées dans la figure 1. A l'étape 1, on entrevoit deux regroupements : *[hausse,prix,électricité,EDF]* et *[EDF,EPR,militant,moratoire]*. On constate que le nœud *EDF* est à la croisée de ces deux clusters. Chaque nœud partage 2 voisins avec n'importe quel autre nœud du graphe. La pondération des arêtes est modifiée à l'étape 2 afin d'isoler les nœuds multiclasses comme *EDF*. A l'étape 3, on procède à la détection des composantes connexes du graphe, en ignorant les arêtes ayant un poids inférieur à *0.5* : on obtient donc deux clusters : *[hausse,prix,électricité]* et *[EPR,militant,moratoire]*. A l'étape 4, les arêtes ignorées à l'étape précédente reprennent leur place : *EDF* est relié à plus de 80% des nœuds de chacun des clusters détectés à l'étape 3. Il les rejoint donc.

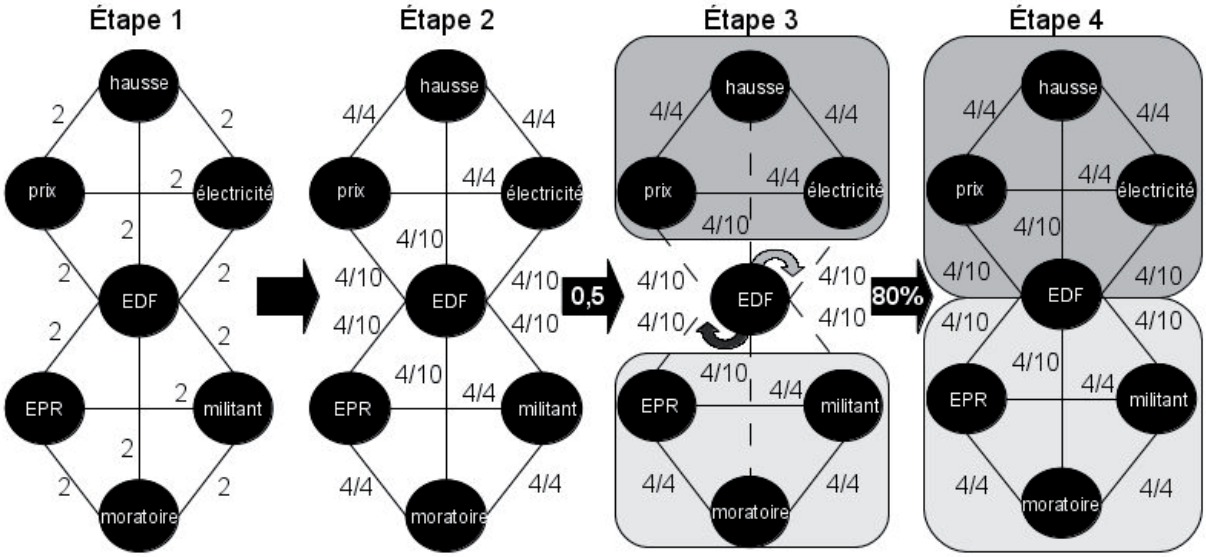

Figure 1 - Les 4 nouvelles étapes ajoutées à SNN.

4. Présentation du corpus

Notre corpus de test a été collecté à l'aide d'un méta-moteur de veille en réponse à la requête *nucléaire*. Nous n'avons considéré que les articles de presse rédigés en français entre le 17/04/2011 et le 16/05/2011 inclus. Cette période a été choisie pour son intérêt dans un cadre de veille : quelle image a le nucléaire en France un mois après l'incident survenu à Fukushima le 11 mars 2011 ? Après filtrage manuel[5], le corpus comporte 471 articles uniques, 170 437 mots et 12 070 vocables. Le corpus a été étiqueté avec *Cordial*[6]. Les mots trop fréquents ou fortement liés au Web ont été filtré. Seuls les substantifs ont été conservés ; ces derniers sont lemmatisés afin de favoriser les rapprochements. Nous avons pris le parti de nous limiter, dans un

5 Textes hors sujet ou devenus inaccessibles (liens morts).

6 http://www.synapse-fr.com/Cordial_Analyseur/Presention_Cordial_Analyseur.htm.

premier temps, aux substantifs car il s'agit de la catégorie la plus « marquée sémantiquement ». Cependant, nous souhaitons dans l'avenir montrer l'influence de chacune des catégories sur les thématiques obtenues : nous projetons donc de prendre en compte, à court terme, les adjectifs ainsi que les verbes. Nous comptons aussi évaluer notre méthode en ignorant les catégories morphosyntaxiques des mots.

La taille de notre corpus peut sembler dérisoire face à ceux utilisés par exemple par les *topic models* ; en effet, un corpus de plusieurs millions de mots entraine des difficultés techniques indéniables qu'il faut être en mesure de gérer. Cependant, le travail sur des corpus de taille moyenne entraine d'autres genres de contraintes ; il est en effet difficile d'extraire des regroupements pertinents avec moins de données en entrée. Bien que cela n'ait jamais, à notre connaissance, été formellement prouvé, il est communément admis que les modèles statistiques ont besoin de corpus de plusieurs millions de mots pour fournir de bons résultats.

Par ailleurs, le corpus que nous avons collecté est un « scénario réel » de collecte d'entreprise ; il est courant d'avoir des corpus de veille dans cet ordre de grandeur (moins d'un million de mots). Il est ainsi primordial de ne pas négliger ce cas de figure. Bien entendu, des tests à plus grande échelle sont tout de même prévus dans l'avenir.

5. Résultats et discussion

5.1. Présentation des thématiques

Rappelons que le but de ces clusters est de permettre une bonne vision d'ensemble du corpus, et servir de premières pistes d'exploration et d'interprétation de la part du veilleur. 11 thématiques sont renvoyées par notre système ; 10 d'entre elles sont facilement interprétables. Nous leur donnons ci-dessous un nom issu de l'interprétation des mots : 1. la hausse des prix de l'électricité en France ; 2. Tchernobyl ; 3. la centrale de Mühleberg et le nucléaire suisse en général ; 4. écologie, société et politique ; 5. reportage *La Zone* à propos des familles vivant aux alentours de Tchernobyl ; 6. incident dans un brise-glace russe ; 7. bourse et entreprises (rachats, fusions, etc.) ; 8. mouvements anti-EPR ; 9. candidature de Nicolas Hulot ; 10. nucléaire iranien.

La thématique 5 ne concerne qu'un seul document dans l'ensemble du corpus. Les très petites thématiques ne sont pas un problème en soi, mais cette dernière est beaucoup trop proche de celle sur Tchernobyl. Dans l'idéal, elles devraient être fusionnées. La thématique 6 concerne un sujet d'actualité qui n'est présent dans la presse qu'un seul jour (4 articles en tout) ; ce niveau de granularité est encourageant, préfigurant la possibilité de considérer la dimension temporelle dans le processus, par exemple : quelles sont les thématiques du jour, comment évoluent-elles, etc.

Faute de place, nous ne présentons partiellement que certaines thématiques. Les mots en gras appartiennent à plusieurs thématiques :

Prix de l'électricité	Mouvements anti-EPR	Tchernobyl	Écologie et société	Nicolas Hulot	Incident dans un brise-glace	Iran
EDF	**EDF**	**centrale**	**centrale**	Nicolas Hulot	Rosatomflot	Iran
électricité	EPR	Tchernobyl	planète	Eva Joly	brise-glace	Téhéran
hausse	Flamanville	Russie	terre	EELV	navire	Ashton
tarif	Greenpeace	drame	**système**	primaire	**système**	programme
euro	Rousselet	liquidateur	évolution	campagne	réacteur	enrichissement
inflation	militant	mort	humain	électeur	fuite	arme
marché	moratoire	monument	**fossile**	**fossile**	incident	sanction
Besson	ASN	mémoire	**écologie**	**écologie**	arctique	discussion
Arenh	chantier	radioprotection	CO2	essence	Barents	dialogue
Nome	site	risque	serre	carbone	mer	négociation

Une bonne connaissance du domaine est bien entendu requise pour bien interpréter ces rapprochements. Le cas échéant, ces ensembles servent de points d'amorce à des recherches plus poussées sur le sujet. L'appartenance d'un mot à plusieurs thématiques est très intéressante dans un cadre de veille. Par exemple, *EDF* est ici fortement lié aux problématiques de la hausse des prix de l'électricité et aux mouvements anti-EPR. Le cas du mot *système* est aussi intéressant. Ce dernier est à la fois présent dans la thématique sur l'écologie et dans celle sur l'incident dans un brise-glace, mais avec un sens diffèrent. Dans un cas, il s'agit principalement du système politique ou économique :

- *Il faut donc agir maintenant et remplacer le **système** économique actuel* (…)
- *Sortir du nucléaire, c'est d'abord sortir du **système** actuel.*
- *Notre **système** est périmé, car bâti sur le principe d'une énergie bon marché (...)*

Dans l'autre, il s'agit du système mécanique :

- *Une faible augmentation de la radioactivité dans l'air a été constatée dans le **système** de ventilation* (…)
- *Le **système** du réacteur sera arrêté et le processus de refroidissement commencera.*
- *La cause probable de l'incident est une perte d'étanchéité des **systèmes** de la première enceinte du réacteur.*

Pour des raisons de lisibilité, les résultats sont présentés sous forme de listes. Rappelons que nos thématiques sont des sous-graphes et que les mots entretiennent donc des relations plus ou moins fortes entre eux ; certains mots entretiennent des relations privilégiées avec d'autres, esquissant ainsi des sous-regroupements intéressants, comme illustré dans la figure 2. Pour Tchernobyl par exemple, 3 sous-thématiques se démarquent clairement[7] : 1. construction du

7 Ces sous-regroupements sont bien séparés lorsque l'on augmente le seuil à *0.6*.

sarcophage de confinement ; 2. radioactivité, santé et pollution ; 3. mort et commémorations. Une analyse plus détaillée de ce graphe est présentée dans la partie 5.2.

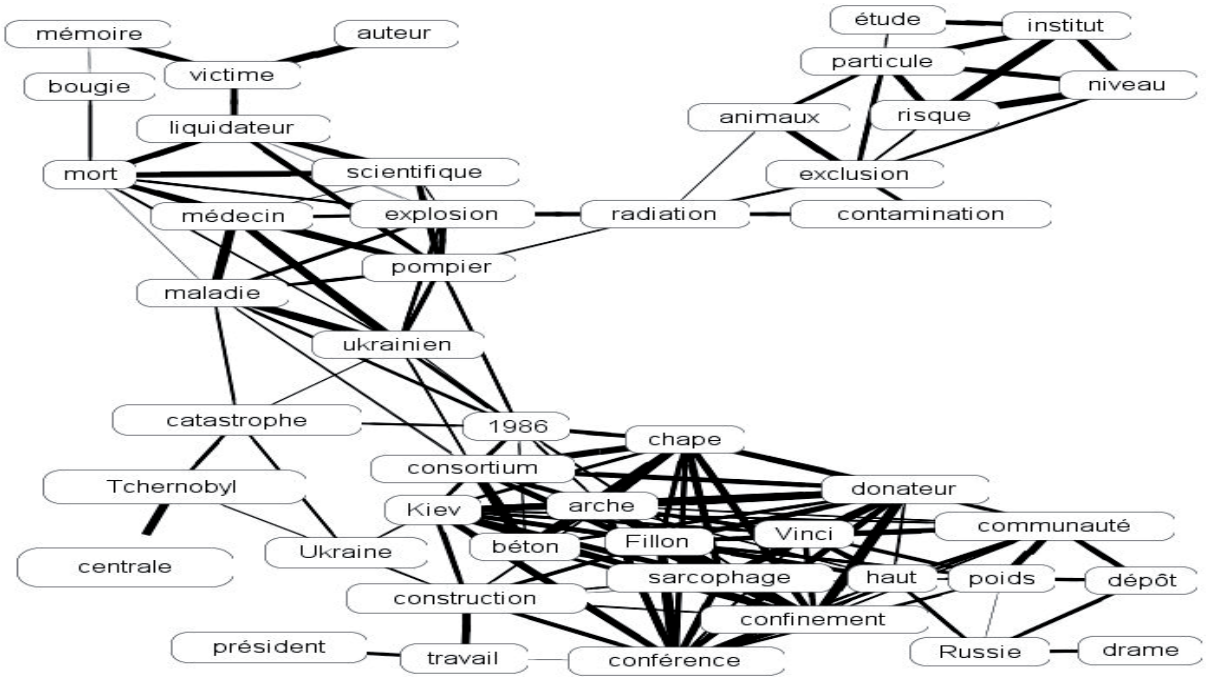

Figure 2 - Thématique sur Tchernobyl sous forme de graphe. Pour des raisons de lisibilité, le graphe a été allégé en retirant tous les mots apparaissant moins de 15 fois. On constate des liens privilégiés entre certains mots, conduisant à des sous-regroupements.

Notons que certaines des 11 thématiques sont extrêmement denses (presque des cliques) et que les liens entre les mots deviennent alors triviaux (voir figure 3) : dans ces cas, la représentation sous forme de graphe n'apporte rien par rapport aux listes classiques telles qu'on peut voir notamment en sortie des *topic models*. Néanmoins, le fait que tous les mots soient fortement liés entre eux peut dans certains cas être une information intéressante.

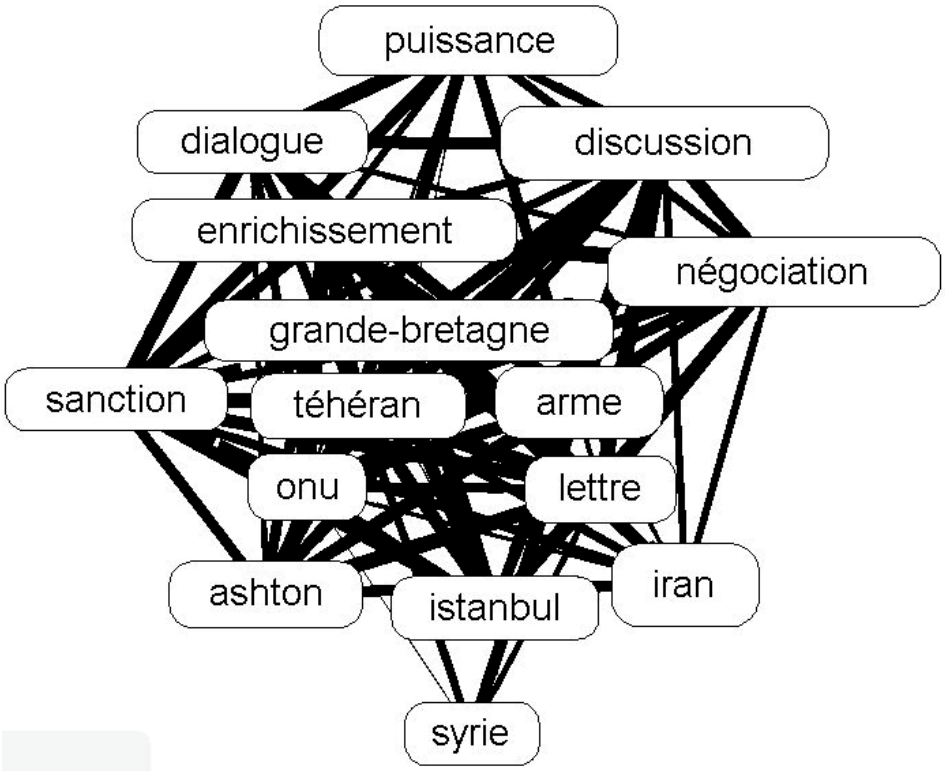

Figure 3 - Thématique sur l'Iran. Tous les mots sont fortement liés entre eux.

La raison pour laquelle nous obtenons un cluster difficilement interprétable parmi les 11 renvoyés est évidente lorsque l'on visualise la configuration du graphe : les thématiques apparaissent toutes nettement en périphérie du graphe tandis que les mots fortement multiclasses se retrouvent au centre (au croisement des thématiques concernées). Nous avons constaté que les nœuds de ce cluster problématique se trouvent tous au centre du graphe : il s'agit donc de nœuds ayant une très forte densité et rattachés à l'ensemble des thématiques, mais reliés à moins de 80% des nœuds de chacune d'entre elles.

5.2. Analyse sémantique des résultats

L'allure générale du graphe présenté dans la figure 2 nous renseigne à la fois sur certaines sous-thématiques mais aussi sur leur degré de maturité. On distingue des zones lexicalement pauvres et d'autres beaucoup plus denses. En périphérie du graphe, des ilots de forte cohérence lexicale donnent à penser qu'il s'agit de formes sémantiques stabilisées tant elles sont aisément restituables (par exemple, bougie et mémoire ; chape et sarcophage) et s'apparentent parfois à des figements syntagmatiques (« centrale [de] Tchernobyl », « Tchernobyl [en] Ukraine », « niveau [de] risque »). L'épaisseur des liens attestent d'ailleurs de fréquences remarquables. A l'inverse, les mots qui opèrent des jonctions entre les sous-thématiques sont des éléments génériques susceptibles d'être partagés par plusieurs formes sémantiques (1986, catastrophe, explosion, radiation, mort). Ils constituent des éléments structurant du graphe ; ils en assurent la cohésion générale, et semblent être les indices, au niveau du corpus et de l'intertexte, d'isotopies génériques. Enfin, la densité de la sous-thématique liée à la construction du sarcophage de

confinement, en bas du graphe, est l'indice de son importance dans le corpus, mais aussi de la relative instabilité des formes sémantiques qui le composent : le vocabulaire est diversifié en raison de la vitalité du thème tandis que les liens, denses et variés, témoignent quant à eux de combinaisons lexicales riches. A l'inverse, les formes sémantiques liées aux niveaux de risques, aux victimes et à la catastrophe proprement dite sont, comme nous l'avons vu, lexicalement appauvries parce qu'elles sont stabilisées, leur lexicalisation est achevée.

5.3. Assignation des thématiques aux documents

Nous avons recours à une méthode simple pour assigner une thématique aux documents de notre corpus ; nous considérons qu'un document appartient à une thématique s'il possède au moins *n* mots du cluster (aucune pondération selon la taille), fidèlement à l'idée selon laquelle tout ensemble de mots apparaissant ensemble identifie un sujet de façon univoque. Un document peut traiter de plusieurs thématiques à la fois, et peut n'appartenir à aucune des thématiques renvoyées par le système.

Afin d'évaluer la qualité de la classification thématique, chacun des textes a été assigné manuellement à une ou plusieurs thématiques parmi celles renvoyées par le système, puis nous avons comparé avec la classification automatique obtenue. La figure 4 indique les scores de rappel[8] et de précision (ainsi que le F-score associé) pour cette tâche de classification thématique. On constate que la précision atteint un très haut score et commence à stagner à partir de $n = 7$. $n = 4$ semble un bon compromis entre rappel et précision. La forte corrélation entre *n* et la précision permet d'assigner des scores de confiance à la classification obtenue.

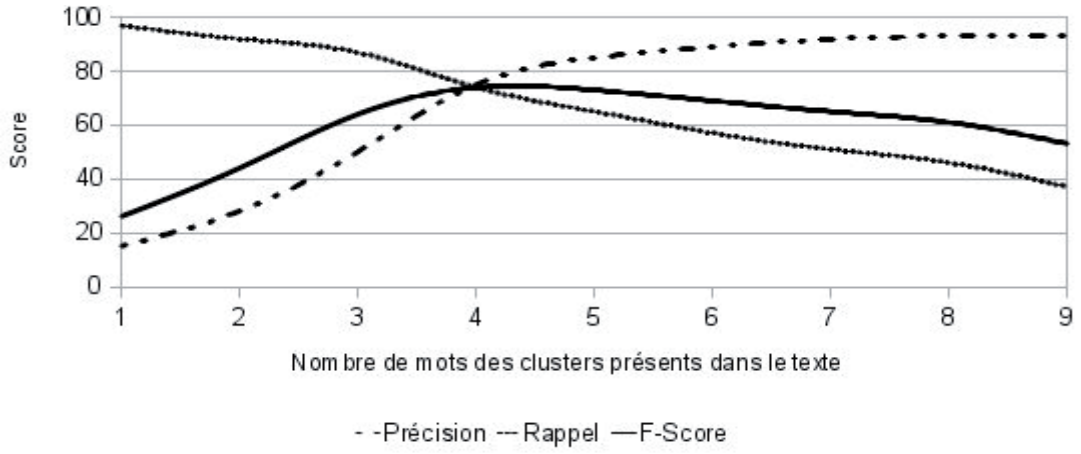

Figure 4 - Précision, rappel et F-Score pour la tâche d'assignation des documents aux thématiques, en fonction du nombre de mots des clusters présents dans le texte.

La majeure partie du silence provient ici des très petits documents (une phrase ou deux – c'est le cas des dépêches) étant donné que la valeur de *n* ne prend pas en compte la taille des textes. Le bruit est souvent dû aux mêmes mots assez peu « discriminants ». C'est par exemple le cas de *dialogue* ou *discussion* dans la thématique autour de l'Iran. Afin de contourner ce problème,

8 Dans un cadre de veille, il peut être préférable de légèrement privilégier le rappel à la précision car il est important de perdre le moins d'information possible.

les mots devraient idéalement être pondérés au sein de chacun des clusters ; on verrait alors apparaitre des mots centraux (sous-classe de mots pour lesquels le sème est inhérent) entouré de mots périphériques (sous-classe de mots pour lesquels le sème est afférent). La meilleure façon de pondérer les nœuds au sein de chaque cluster reste encore à déterminer, mais quelques pistes sont envisagées : nombre de liens forts intra-cluster, nombre de voisins inter-cluster, etc.

6. Conclusion et perspectives

Nous avons proposé une méthode simple, reposant sur les plus proches voisins partagés dans un graphe de cooccurrences, afin de faciliter le parcours interprétatif du veilleur : 1. meilleure vue macroscopique (détection de thématiques), 2. aide au tri, et au retour au texte (catégorisation automatique). La méthode cumule des propriétés qui nous semblent primordiales d'un point de vue linguistique : classification non exhaustive, multiclasse, et selon la répartition des mots. L'algorithme est par ailleurs assez indépendant de la configuration du graphe.

Ces résultats sont très encourageants. Plusieurs pistes pour les améliorer encore sont envisagées : 1. prise en compte de la taille des textes pour la classification ; 2. évaluation de l'impact des catégories morphosyntaxiques sur les résultats en prenant en compte les adjectifs[9] et les verbes ; 3. pondération des nœuds au sein des clusters (mots discriminants *vs* mots périphériques) ; 4. détection en amont de termes : certaines entités nommées sont reconnues par Cordial mais on est le plus souvent confronté à des mots isolés ; 5. révision des valeurs des seuils et adoption, comme (Ferret, 2006), d'un mode unique de fixation s'adaptant à la distribution des valeurs ; 6. évaluation de l'impact de la mesure de dissimilarité employée pour pondérer les relations de cooccurrences dans le graphe d'origine, en la comparant avec d'autres mesures (l'information mutuelle de (Church et Hanks, 1990) par exemple).

Par ailleurs, étant donné que nos travaux se situent dans un cadre de veille et d'intelligence économique à partir de données issues du Web, il serait bon d'évaluer la robustesse de l'approche sur des articles moins bien écrits : les blogs par exemple. Enfin, il faudrait réfléchir à des outils de visualisation adaptés : les graphes de cooccurrences sous forme graphique sont une aide appréciable pour le parcours interprétatif du veilleur.

Références

Blei D., Ng A. and Jordan M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, vol. 3: 993-1022.

Church K. and Hanks P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, vol.16(1): 22-29.

Deerwester S., Dumais S., Landauer T., Furnas G. and Harshman A. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, vol.41(6): 391-407.

Ertöz L., Steinbach M. and Kumar V. (2003). Finding Topics in Collections of Documents: A Shared Nearest Neighbor Approach. *Workshop on Text Mining, held in conjunction with the First SIAM International Conference on Data Mining (SDM 2001)*.

Ferrer i Cancho R. and Solé R. (2001). The small world of human language. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol.268(1482): 2261-2265.

9 Un rapide test avec les adjectifs montre que : 1. les clusters restent souvent presque inchangés ; 2. les mots des clusters sont majoritairement des substantifs.

Ferret O. (2006). Approches endogènes et exogènes pour améliorer la segmentation thématique de documents. *TAL*, vol.47(2).

Grefenstette G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, MA. USA.

Hofmann T. (1999). Probabilistic Latent Semantic Indexing. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 50-57.

Jarvis R. and Patrick E. (1973). Clustering Using a Similarity Measure Based on Shared Near Neighbors. *Computers, IEEE Transactions on,* vol.C-22(11): 1025-1034.

Lee D. and Seung H. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, vol.401(6755): 788-791.

Mayaffre D. (2008). Quand « travail », « famille », « patrie » co-occurrent dans le discours de Nicolas Sarkozy. Etude de cas et réflexion théorique sur la co-occurrence. *Actes du colloque JADT 2008 (9es journées internationales d'analyse statistique des données textuelles)*, pp. 811-822.

Palla G., Derenyi I., Farkas I. and Vicsek T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, vol.435(7043): 814-818.

Pincemin B. (1999). Sémantique interprétative et analyses automatiques de textes: que deviennent les sèmes? *Sémiotiques*, vol.17: 71-120.

Rastier F. (1987). *Sémantique interprétative*. PUF: Paris, France.

Slodzian M. (2000). Wordnet: what about its linguistic relevancy? *Proceedings of the EKAW conference*, Juan-les-Pins.

Véronis J. (2003). Hyperlex: cartographie lexicale pour la recherche d'informations. *Proceedings of TAL 2003*, pp. 265-274.

Watts D. and Strogatz S. (1998). Collective dynamics of 'small-world' networks. *Nature*, vol.393(6684): 440-442.