Sous-graphes de cooccurrences pour la détection de thématiques dans un corpus de taille moyenne

Aurélien Lauf*|**

** ERTIM*[1] *(Équipe de Recherche en Textes, Informatique, Multilinguisme)*
INALCO - 49bis avenue de la Belle Gabrielle - 75012 PARIS

*** AMI Software France*
Immeuble "Le Cristal", 1475 avenue Albert Einstein - 34000 Monptellier
alu@amisw.com

RÉSUMÉ. *Ce papier aborde la question de la classification non supervisée de documents, dans un contexte de veille sur le Web (corpus de taille moyenne). Notre but est d'assister le veilleur dans deux tâches : 1. dégager des thématiques à partir du corpus ; 2. ranger chaque texte dans une ou plusieurs de ces thématiques. Nous proposons une approche linguistique, reposant sur les plus proches voisins partagés dans un graphe de cooccurrences. Nos thématiques peuvent se chevaucher (partager des mots) et 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. Les premiers résultats sont très encourageants.*

ABSTRACT. *This paper deals with document clustering in the context of strategic and competitive intelligence on the Web (medium-sized corpora). Our goal is to assist the user with the following tasks: 1. find topics within collected textual data; 2. put each document in one or more of these topics. We present a corpus linguistics approach, using shared nearest neighbors within a cooccurrence graph. The topics we build may overlap (i.e. share many words) and are not only set of words: using graph theory formalism, we are able to express subtle semantic relations between words within each topic. First results are quite satisfying.*

MOTS-CLÉS : *classification non supervisée, cooccurrence, linguistique de corpus, graphe, veille.*

KEYWORDS: *document clustering, cooccurrence, corpus linguistics, graph, competitive intelligence.*

1. Directeur de thèse : Mathieu Valette.

1. Introduction

Nos travaux s'inscrivent dans un cadre de veille sur Internet. Notre but est de dégager des thématiques du corpus et d'y ranger les documents. Nous proposons une alternative linguistique aux méthodes statistiques. Notre approche repose sur des sous-graphes de cooccurrences et permet d'exprimer des liens entre les mots de chaque thématique.

Notre corpus est un scénario réel de collecte d'entreprise. Il est composé d'articles de presse rédigés en français entre le 17/04/2011 et le 16/05/2011, collectés par un méta-moteur de veille pour la requête *nucléaire*. Le corpus compte 471 articles, 170 437 mots et 12 070 mots uniques. Le corpus a été étiqueté avec Cordial[1]. Pour le moment, seuls les substantifs lemmatisés ont été conservés. La principale contrainte de notre corpus est la forte intersection lexicale entre les thématiques : il s'agit d'extraire des thématiques proches toutes relatives au sujet général qu'est le nucléaire.

2. Positionnement

De nombreuses approches statistiques sont utilisées en clustering de documents, par exemple la LSA (Deerwester *et al.,* 1990) ou les *topic models* (LDA (Blei *et al.,* 2003), PLSI (Hofmann, 1999)). Nous abordons la question sous un autre angle, celui de la lexicométrie et de la linguistique de corpus. Les *topic models* notamment considèrent qu'une thématique est un ensemble de mots pondérés par des probabilités d'apparition. En utilisant une approche de classification non supervisée sur des graphes de cooccurrences, nous espérons formaliser la notion de cohérence sémantique de façon plus poussée ; la théorie des graphes nous permet d'exprimer concrètement des relations pondérées entre ces mots, qui peuvent ensuite être observées à l'aide d'outils de visualisation dynamique, permettant une meilleure interprétation des résultats.

3. Description de notre approche

Nous nous basons uniquement sur les cooccurrences du corpus pour qualifier les relations entre les mots ; nous faisons l'hypothèse que des thématiques peuvent être modélisées par des regroupements de mots apparaissant dans des contextes similaires : ici le paragraphe. A la différence d'approches reposant sur des syntagmes (Hammouda *et al.,* 2004) (Wang *et al.,* 2011), nous mettons l'accent sur des relations à grande distance. La question de la délimitation du contexte reste néanmoins ouverte ; nous pensons que les deux types de relations sont complémentaires.

Un mot peut n'appartenir à aucune thématique, et peut être dans plusieurs à la fois (polysémie, homographie, etc.). L'algorithme SNN (*Shared Nearest Neighbours*) de (Jarvis *et al.,* 1973), repris notamment par (Ertoz *et al.,* 2003) et (Ferret, 2006), a re-

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

tenu notre attention car il permet une classification non exhaustive, et indépendante de l'échelle. Nous créons une matrice C $v \times v$ de cooccurrences, tel que $C_{ij} = Freq_{ij}$, sachant que v est le nombre de mots uniques du corpus et $Freq_{ij}$ le nombre de fois que les mots aux indices i et j cooccurrent dans le même paragraphe. Ces valeurs sont remplacées par la mesure de dissimilarité présentée dans (Véronis, 2003)[2]. Cette matrice forme un graphe de cooccurrences où les noeuds sont les mots (occurrant au moins de 10 fois dans le corpus), et les arêtes les relations de cooccurrences. On y applique SNN : 1. filtrage des liens les moins forts : on obtient le graphe des plus proches voisins, qui représente les affinités de premier ordre (Grefenstette, 1994) (liste de voisins directs d'un mot) ; 2. remplacement du poids des arêtes par le nombre de voisins que les noeuds ont en commun afin d'obtenir le graphe des plus proches voisins partagés, ce qui permet une transition vers des affinités de second ordre (mots partageant un même environnement : passage de relations plutôt syntagmatiques à des relations paradigmatiques propices à des regroupements sémantiques) ; 3. filtrage des liens inférieurs à un seuil donné ; les clusters correspondent aux composantes connexes de ce graphe. Néanmoins, notre graphe est à ce stade encore connexe (un seul cluster) à cause des mots partageant de nombreuses thématiques. Nous proposons de nouvelles étapes (voir figure 1) afin de permettre des regroupements plus stables et de permettre le chevauchement de clusters.

– On répète l'étape 2, ce qui rapproche à nouveau par transitivité certains mots. Cette étape rend les résultats plus robustes.

– Il s'agit désormais d'isoler les composantes denses ainsi que les mots multiclasses. Deux noeuds appartiennent au même cluster s'ils partagent la majorité de leurs voisins respectifs ; on ne compare plus le nombre absolu de voisins communs mais le nombre de voisins communs relativement au nombre total de voisins des noeuds en question, comme indiqué dans la formule 1[3]. Les arêtes vers les noeuds multiclasses sont ainsi pénalisées.

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

– On filtre les arêtes inférieures à un seuil donné (0.5 ou 0.6[4]). Nos thématiques sont les composantes connexes de ce nouveau graphe.

– On ré-intègre les arêtes n'ayant pas « survécu » à l'étape précédente. Un noeud appartient à un cluster supplémentaire s'il a des liens avec 80% des mots d'un autre cluster. Cela a aussi pour effet de fusionner certains petits clusters isolés à tort.

---

2. Nous projetons dans l'avenir de comparer avec d'autres mesures.
3. ${C_{ij}}^2$ est le nombre de voisins que partagent i et j ; N_i et N_j sont respectivement le nombre de voisins qu'ont les noeuds 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.
4. Un seuil trop bas empêche les thématiques de se séparer. A l'inverse, un seuil supérieur à 0.7 fragmente le graphe en petits groupes triviaux.

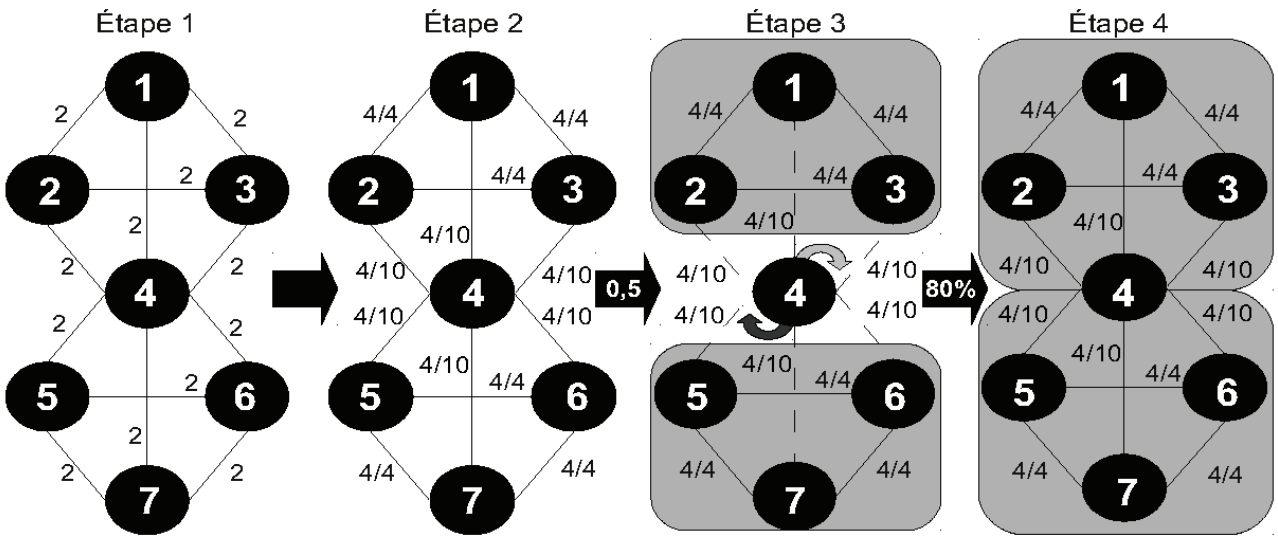

Figure 1. *Etapes ajoutées à SNN. 1. on entrevoit deux groupes :* $[1, 2, 3, 4]$ *et* $[4, 5, 6, 7]$ *; 2. le poids des arêtes est modifié (voir formule 1) pour isoler les noeuds multiclasses ; 3. on cherche les composantes connexes en ignorant les arêtes inférieures à* 0.5 *; on obtient deux clusters :* $[1, 2, 3]$ *et* $[5, 6, 7]$ *; 4. on reprend les arêtes ignorées précédemment :* 4 *est relié à plus de 80% de chaque cluster. Il les rejoint.*

Prix énergie	**Tchernobyl**	**Bourse**	**Nicolas Hulot**	**Iran**
EDF	centrale	bourse	Nicolas Hulot	Iran
électricité	Tchernobyl	titre	Eva Joly	Téhéran
hausse	Russie	EBITDA	écolo	Ashton
tarif	drame	point	EELV	programme
euro	liquidateur	part	primaire	enrichissement
inflation	mort	offre	campagne	arme
marché	monument	affaire	électeur	sanction
Besson	mémoire	dollar	proposition	discussion
Arenh	radioprotection	objectif	essence	dialogue
Nome	risque	entreprise	carbone	négociation

Tableau 1. *Extrait de 5 thématiques. Une bonne connaissance du domaine est requise pour les interpréter. Le cas échéant, ils peuvent servir de bons points d'amorce.*

10 sur les 11 thématiques renvoyées sont parlantes[5]. Le tableau 1 en présente certaines, sous forme de listes pour des raisons de lisibilité. Rappelons que nos thématiques sont des sous-graphes et que les mots entretiennent des relations pondérées entre eux, esquissant des sous-regroupements, comme illustrés dans la figure 2. Néanmoins, certaines thématiques sont très denses, rendant les liens triviaux.

5. L'une d'entre elles ne concerne qu'un jour. Ce niveau de granularité est encourageant et préfigurerait la prise en compte de la dimension temporelle dans le processus.

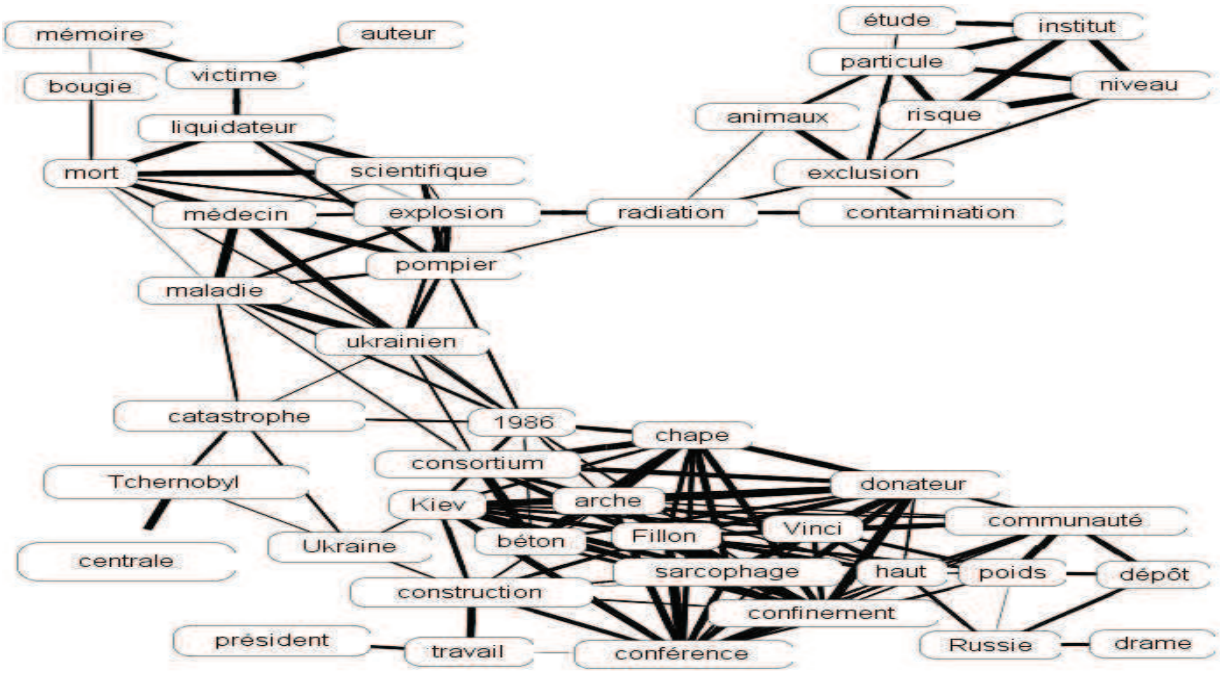

Figure 2. *Thématique sur Tchernobyl sous forme de graphe (seuil de* 0.5*). 3 sous-thématiques se démarquent : 1. enceinte de confinement ; 2. radioactivité ; 3. mort et commémorations. Le type de relations symbolisées par les arêtes reste à définir.*

Pour comparaison, nous avons soumis le corpus à la LDA. Les résultats sont très proches pour 10 clusters[6]. En augmentant à 15, les découpages sont plus fins, avec notamment les trois sous-thématiques de la figure 2. Les mots y sont pondérés, mais les relations entre eux sont absentes. Des comparaisons plus poussées sont prévues.

3.1. *Assignation des thématiques aux documents*

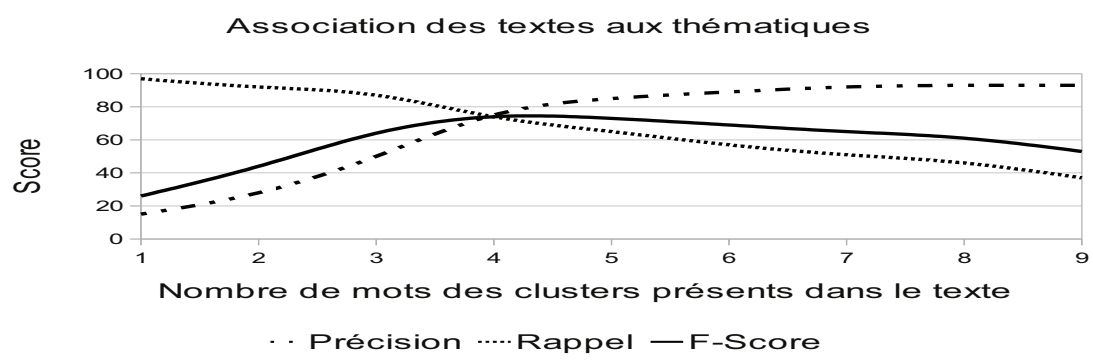

Figure 3. *Précision, rappel et F-Score pour l'assignation des documents aux thématiques, en fonction du nombre de mots des clusters dans le texte.*

Nous assignons une thématique à un document s'il contient au moins n mots du cluster. Chaque texte a été assigné par des annotateurs à une ou plusieurs thématiques

6. La LDA requiert de spécifier le nombre de clusters.

du système, puis nous avons comparé avec la classification automatique (figure 3). Le silence provient surtout des petits textes. Le bruit est souvent dû aux mots peu discriminants comme *dialogue* dans la thématique sur l'Iran. Idéalement, les mots devraient être pondérés (mots centraux et périphériques). Quelques pistes sont envisagées : nombre de liens forts intra-cluster, nombre de voisins inter-cluster, etc.

4. Conclusion et perspectives

Nous avons proposé une méthode reposant sur les plus proches voisins partagés d'un graphe de cooccurrences pour dégager des thématiques et les assigner aux documents. Les thématiques peuvent se chevaucher. Les résultats sont encourageants. Plusieurs pistes pour les améliorer sont envisagées : 1. considération de la taille des textes pour la classification ; 2. évaluation de l'impact des catégories morpho-syntaxiques ; 3. pondération des noeuds des clusters ; 4. détection en amont de termes ; 5. réflexion sur les valeurs des seuils ; 6. comparaison avec d'autres mesures de similarité. Par ailleurs, il faudrait évaluer la robustesse de l'approche sur des articles moins bien écrits comme des blogs. Enfin, nous souhaitons nous concentrer par la suite sur la question de la délimitation du contexte, et sur l'évolution des thématiques dans le temps.

5. Bibliographie

Blei D., Ng A., Jordan M., « Latent Dirichlet Allocation », *Journal of Machine Learning Research*, 2003.

Deerwester S., Dumais S., Landauer T., Furnas G., Harshman R., « Indexing by Latent Semantic Analysis », *Journal of the American Society of Information Science*, 1990.

Ertoz L., Steinbach M., Kumar V., « 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)*, 2003.

Ferret O., « Approches endogène et exogène pour améliorer la segmentation thématique de documents », *TAL*, 2006.

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

Hammouda K. M., Kamel M. S., « Efficient Phrase-Based Document Indexing for Web Document Clustering », *IEEE Transactions on knowledge and data engineering*, 2004.

Hofmann T., « Probabilistic Latent Semantic Indexing », *Proceedings of the 22nd ACM SIGIR conference on Research and development in information retrieval*, 1999.

Jarvis R., Patrick E., « Clustering Using a Similarity Measure Based on Shared Near Neighbors », *Computers, IEEE Transactions on*, 1973.

Véronis J., « Hyperlex : cartographie lexicale pour la recherche d'informations », *Proceedings of TALN 2003*, 2003.

Wang Y., Ni X., Sun J.-T., Tong Y., Chen Z., « Representing document as dependency graph for document clustering », *Proceedings of the 20th ACM CIKM'11*, 2011.