Vers une approche Visual Analytics pour explorer les variantes de sujets d'un corpus

Nicolas Médoc[*,**] Mohammad Ghoniem[**] Mohamed Nadif[*]

[*]LIPADE, Paris-Descartes
[**]Luxembourg Institute of Science and Technology
nicolas.medoc@list.lu,
mohammad.ghoniem@list.lu,
mohamed.nadif@mi.parisdescartes.fr

Notre objectif, à terme, est de proposer un outil de visualisation analytique (*Visual Analytics*) permettant d'explorer les différents points de vue ou les variantes des sujets traités dans un corpus tel que des articles de presse. Appliqués sur le modèle de sac de mots (matrice *termes x documents*), les méthodes probabilistes d'extraction de sujets du type *Latent Dirichlet Allocation* calculent une distribution des termes dans un nombre prédéfini de sujets, pour les regrouper par proximité sémantique. Il en résulte ensuite une distribution des documents dans ces mêmes sujets. D'autres méthodes du type co-clustering (Govaert et Nadif, 2013) considèrent simultanément les vecteurs des termes et des documents pour produire des bi-clusters regroupant les termes sémantiquement proches et les documents qui les partagent.

Les visualisations habituelles des sujets avec des nuages de mots permettent d'interpréter les sujets à travers les N termes les plus représentatifs. Dans le contexte d'un corpus d'articles de presse, cette approche met en exergue ce qui est majoritaire et déjà connu. Un analyste cherche, au contraire, à identifier des points de vue alternatifs et inédits. Dans ce but, nous proposons de structurer le corpus en regroupant les documents par points de vue partagés, selon différentes combinaisons de mots-clés colocalisés dans les documents. Nous nous appuyons sur *Bimax* (Prelic et al., 2006), une méthode de bi-clustering non-disjoint qui extrait, à partir d'une matrice binaire, tous les bi-clusters (blocs constitués uniquement de 1) vérifiant une contrainte d'*inclusion maximale*. Cette contrainte impose qu'aucun bi-cluster ne soit entièrement inclus dans un autre. *Bimax* est adapté aux matrices sparses (c'est le cas pour le texte) et permet d'extraire tous les bi-clusters optimaux. Dans une matrice *termes x documents*, discrétisée avec un seuil sur des poids de type TF-IDF, les bi-clusters regroupent des documents de manière unique selon les multiples co-occurences possibles de mots-clés. Nous faisons l'hypothèse que les bi-clusters ainsi obtenus constituent l'ensemble des points de vue concernant les sujets d'un corpus. Cependant, *Bimax* produit une grande quantité de bi-clusters contenant beaucoup de redondances au niveau des termes et des documents, mais aussi quelques spécificités (termes ou documents exclusifs à un bi-cluster). De plus, la représentation visuelle et l'interprétation des bi-clusters non-disjoints restent des tâches difficiles (Sun et al., 2014).

Pour faciliter l'exploration des bi-clusters, nous cherchons à hiérarchiser les éléments des deux dimensions selon leur degré de redondance dans les bi-clusters, en agrégeant ces derniers par points communs. Nous proposons de décomposer la matrice *termes x documents* en un ensemble de blocs disjoints regroupant les cellules appartenant à une intersection unique de