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

bi-clusters. Le bloc d'intersection avec le plus haut degré de chevauchement contient les éléments les plus redondants, autrement dit, les points communs entre le plus grand nombre de bi-clusters. Ce type de bloc est remonté au niveau d'une racine de la structure *poly-hiérarchique* (structure hiérarchique dont les enfants peuvent avoir plusieurs parents). Un bloc d'intersection d'un nœud enfant concerne un ensemble de bi-clusters inclus dans celui du bloc parent. Ses éléments décrivent des points d'articulation guidant l'utilisateur vers les bi-clusters qui l'intéressent. Un bloc exclusif à un bi-cluster contient des éléments décrivant sa spécificité. Il est placé au niveau d'une feuille. Cette structure hiérarchique peut être modélisée par un graphe orienté acyclique que nous représentons visuellement par un diagramme nœuds-liens basé sur un modèle de champs de force. De manière générale, cette approche peut faciliter la visualisation et l'interprétation des résultats des méthodes de bi-clustering non-disjoint. Dans le cas d'un corpus de textes, elle permet d'identifier les termes communs entre les différents bi-clusters, décrivant ainsi des sujets de haut niveau. L'exploration des termes, des racines jusqu'aux feuilles, permet ensuite de guider l'utilisateur et de comprendre la spécificité des points de vue ou des variantes de sujets. La sélection des documents à chaque nœud (union des documents de tous les bi-clusters de l'intersection) est réduite à mesure que l'utilisateur explore en profondeur les points de vue spécifiques. L'approche décrite ci-dessus peut ainsi faciliter l'identification de points de vue alternatifs et inédits, relatifs à l'actualité.

Avec un corpus d'articles de presse agrégés sur une journée, nos premières analyses à travers les visualisations renforcent notre intuition que cette hiérarchie permet d'explorer différents points de vue. Cependant, appliquer *Bimax* directement sur une matrice *termes x documents* fait apparaitre beaucoup de bi-clusters, de racines et de nœuds, ce qui nuit à la découverte des points de vue intéressants. En pré-traitement, nous envisageons d'extraire les sujets de haut niveau via une méthode de co-clustering en diagonale. Ensuite, en traitant chaque sujet individuellement, *Bimax* peut bénéficier du regroupement des termes avec les documents qui les partagent, pour proposer un ensemble de points de vue plus réduit et mieux ciblé.

Références

Govaert, G. et M. Nadif (2013). *Co-Clustering* (Wiley ISTE ed.).

Prelic, A., S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, et E. Zitzler (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics 22*(9), 1122–1129.

Sun, M., C. North, et N. Ramakrishnan (2014). A five-level design framework for bicluster visualizations. *IEEE TVCG 20*(12), 1713–1722.

Summary

Our purpose is to implement a Visual Analytics tool for exploring topic variants in text corpora. The overlapping bi-clustering methods extract multiple topics from the documents, but the interpretation of the results remains difficult. We make the assumption that bi-cluster overlaps are articulation points between high-level topics, and their multiple variants and viewpoints. We propose to extract and visualize a hierarchical structure of bi-cluster overlaps, allowing to explore the corpus and to discover unsuspected viewpoints.