Analyse exploratoire de corpus textuels pour le journalisme d'investigation

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

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

Résumé. Nous proposons un outil de visualisation analytique conçu pour et avec une journaliste d'investigation pour l'exploration de corpus textuels. Notre outil combine une technique de biclustering disjoint pour extraire des sujets de haut niveau, avec une méthode de biclustering non-disjoint pour révéler plus finement les variantes de sujets. Une vue d'ensemble des sujets de haut niveau est proposée sous forme d'une treemap, puis une visualisation hiérarchique radiale coordonnée avec une heatmap permet d'inspecter et de comparer les variantes de sujet et d'accéder aux contenus d'origine à la demande.

1 Introduction

Nous présentons un outil de visualisation analytique conçu pour faciliter l'exploration de grand corpus par des journalistes d'investigation. Ces journalistes commencent typiquement par se faire une idée générale du sujet de leur investigation, puis se concentrent sur l'identification de faits et de points de vue qui confirment ou infirment leur hypothèse de travail. Les corpus textuels sont souvent modélisés par des matrices *Termes*×*Documents*, construites avec la pondération *TF-IDF* sur la base des noms et des verbes lemmatisés. On peut en extraire des sujets à l'aide de `Coclus`, une technique de biclustering diagonal basé sur la modularité de graphes (Ailem et al. (2015)). On a souvent recours aux nuages de mots pour représenter un sujet décrit par un ensemble de termes associés aux documents qui en traitent. Nous les affichons dans une carte pondérée des sujets. Après avoir identifié un sujet d'intérêt, l'attention du journaliste se porte sur la compréhension de ses variantes. Il s'agit de biclusters non-disjoints mettant en relation des sous-ensembles de documents qui partagent des cooccurrences de termes. Ces variantes peuvent révéler des faits, des points de vue ou des angles d'analyse partagés par plusieurs sources. Les biclusters non-disjoints ont été visualisés de différentes manières, e.g. sous la forme d'enveloppes non-disjointes dans des diagrammes nœuds-liens, des vues matricielles et des coordonnées parallèles par Santamaría et al. (2008). Dans `BiSet`, Sun et al. (2015) utilisent des graphes bipartites chaînés avec des regroupements sémantiques pour représenter les relations de chaînage entre les biclusters. Pour fournir une vue d'ensemble claire

d'un grand nombre de biclusters non-disjoints, nous proposons une visualisation hiérarchique radiale qui permet d'identifier les termes qui les rapprochent ou les distinguent.

2 Vue d'ensemble de l'outil

La vue *Weighted Topic Map* de la Figure 1 est une vue hybride combinant une treemap où chaque sujet extrait par `Coclus` est représenté par un rectangle de surface proportionnelle à son importance. Chaque rectangle contient un nuage de mots détaillant les termes du sujet. La taille et la couleur des mots reflètent respectivement leur représentativité (*TF-IDF*) et le nombre de documents où ils apparaissent. Une projection MDS calculée à partir de la matrice de similarité des biclusters de `Coclus` fournit des positions 2D qui servent à placer les rectangles de la vue *Weighted Topic Map* Ghoniem et al. (2015). Ainsi, les sujets similaires se retrouvent dans des rectangles voisins. L'indice de Jaccard est utilisé pour afficher interactivement les liens entre un sujet cible et les cinq sujets les plus similaires. L'affichage de ces liens vise à atténuer les effets du partitionnement strict de `Coclus`. Quand l'analyste sélectionne un

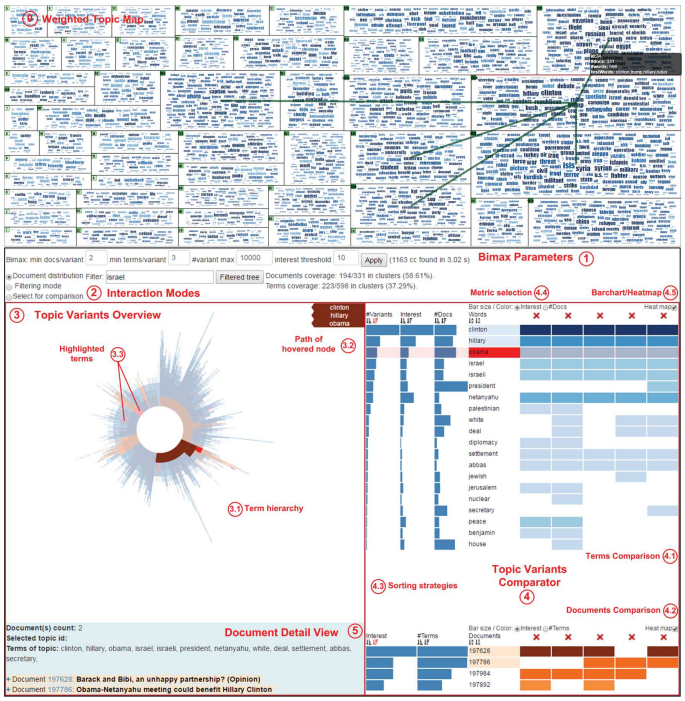

FIG. 1 – *Sujet relatif aux élections présidentielles américaines sélectionné parmi 3 992 articles de presse en ligne compilés entre le 2 et le 16 novembre 2015. Cinq variantes concernant Hillary Clinton sont placés dans le comparateur (*`https://youtu.be/rj9YrTMPClQhttps://youtu.be/rj9YrTMPClQ`*).*

sujet pour l'examiner, `Bimax` (Prelić et al. (2006)), un algorithme de biclustering non-disjoint

à base de motifs, en extrait les variantes. Bien que l'exhaustivité de `Bimax` soit conforme aux besoins de l'analyste, cet algorithme produit une myriade de biclusters. Pour comprendre le résultat de `Bimax`, nous créons une hiérarchie de biclusters sur la base de leurs termes communs à l'aide de l'algorithme `FPTree` (Han et al. (2000)). L'arborescence qui en découle est représentée à l'aide d'une vue *Sunburst* (3.1 dans la Figure 1). Les termes les plus communs ont un degré de chevauchement plus élevé, et sont placés à proximité de la racine, alors que les termes plus spécifiques sont placés plus en périphérie. Chaque chemin allant de la racine jusqu'à une feuille, décrit les termes d'un bicluster. À mesure que l'on s'éloigne de la racine le long de ce chemin, la combinaison de termes devient plus spécifique et caractérise de moins en moins de documents. Au niveau d'une feuille, on retrouve les documents correspondant à un seul bicluster. À l'aide de cette vue et de la vue *Variant Comparator* (4), le journaliste peut se concentrer sur un aspect spécifique du sujet et afficher les liens entre les documents pertinents, pour identifier les faits et les points de vues relatifs à son hypothèse de travail. Le texte des documents est accessible via la vue *Document Detail* (5). De plus, nous fournissons plusieurs modes d'interaction permettant de filtrer les variantes par mots clés et d'analyser la dispersion de ses documents. En survolant un terme de l'arborescence, toutes ses occurrences sont surlignées en rouge (3.3 dans la Figure 1) et la séquence de termes correspondante est affichée à droite (3.2). Le comparateur de variantes permet l'analyse des termes communs et distinctifs, ainsi que la distribution des documents à travers les variantes de sujet sélectionnées. Différents critères de tri sont proposés pour faciliter l'identification des termes les plus informatifs.

Le nombre de biclusters `Bimax` augmente avec la taille et la densité des blocs extraits par `Coclus` et peut excéder les 10 000 biclusters. Pour réduire ce nombre, nous permettons à l'utilisateur de modifier les paramètres de `Bimax` : le nombre minimal de termes ou de documents par bicluster (*MinT*, *MinD*) et le nombre maximal de biclusters (*MaxB*). Comme `Bimax` s'applique à des matrices binaires, nous autorisons aussi l'utilisateur à changer le seuil de binarisation (*Thr*) appliqué à la matrice de poids *TF-IDF*. L'augmentation de ce seuil sélectionne, pour chaque document, les termes les plus représentatifs et réduit la densité et les dimensions de la matrice. La Figure 2, montre l'effet des variations des paramètres sur la hiérarchie de termes associée au sujet concernant les élections présidentielles américaines. Après chaque variation de paramètre, le nœud racine « Obama » est sélectionné systématiquement pour apprécier en orange la distribution de ses documents. Avec les paramètres par défaut ($MinT = 3$, $MinD = 4$, $Thr = 5$), seuls les premiers niveaux des 13 000 biclusters sont visibles dans la *Sunburst*. Augmenter *Thr* ou *MinT* réduit la dispersion des documents concernant « Obama », en préservant mieux la morphologie de la hiérarchie avec *Thr*. Enfin, lorsque *MinD* augmente, la cardinalité des termes des biclusters tend à décroître mais la dispersion des documents sélectionnés demeure jusqu'à ce que le nœud « Obama » disparaisse.

3 Conclusion

Cet outil adopte une approche multi-résolution pour explorer des corpus textuels. La carte pondérée des sujets aide à comprendre des dizaines de sujets rapidement et à apprécier leur importance relative, pour ensuite se focaliser sur un sujet d'intérêt. Une analyse plus fine des variantes de sujet permet d'explorer différents angles ou points de vue partagés par plusieurs documents. Nous envisageons de mener une étude utilisateur pour évaluer, à travers nos visualisations, la faisabilité des tâches en comparant `Coclus` avec `LDA`.

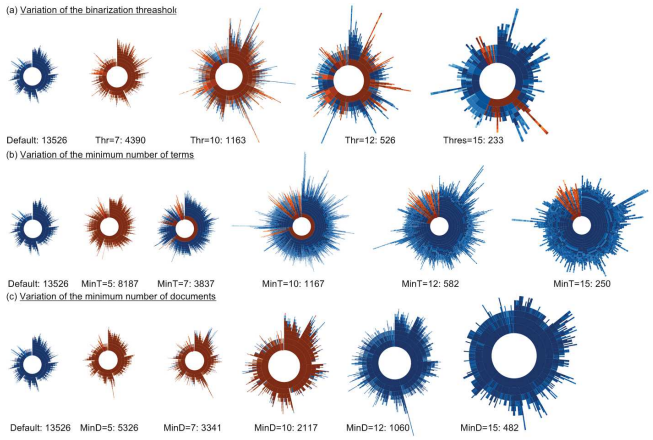

FIG. 2 – *Nombre de biclusters lorsque les paramètres de* `Bimax` *varient.*

Références

Ailem, M., F. Role, et M. Nadif (2015). Co-clustering Document-term Matrices by Direct Maximization of Graph Modularity. In *Proc. of the 24th ACM International on CIKM*, CIKM '15, pp. 1807–1810. ACM.

Ghoniem, M., M. Cornil, B. Broeksema, M. Stefas, et B. Otjacques (2015). Weighted maps : treemap visualization of geolocated quantitative data. In *IS&T/SPIE Electronic Imaging*, pp. 93970G–93970G. Int. Soc. for Optics and Photonics.

Han, J., J. Pei, et Y. Yin (2000). Mining Frequent Patterns Without Candidate Generation. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD '00, pp. 1–12. ACM.

Prelić, 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.

Santamaría, R., R. Therón, et L. Quintales (2008). A visual analytics approach for understanding biclustering results from microarray data. *BMC Bioinformatics 9*(1), 247.

Sun, M., P. Mi, C. North, et N. Ramakrishnan (2015). BiSet : Semantic Edge Bundling with Biclusters for Sensemaking. *IEEE TVCG PP*(99), 1–1.

Summary

We propose a visual analytics tool to support investigative journalists in the exploration of large text corpora. Our tool combines graph modularity-based diagonal biclustering to extract high-level topics with overlapping bi-clustering to elicit fine-grained topic variants. Our coordinate and multi-resolution views allows explorin high-level topics, inspecting their variants while accessing the original content on demand.