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