Exploration visuelle de variantes de sujets par une approche hybride de biclustering

Nicolas Médoc, Mohammad Ghoniem, Mohamed Nadif

Exploration visuelle de variantes de sujet par une approche hybride de biclustering

Nicolas Médoc
LIPADE, Université Paris-Descartes
ERIN-eScience, LIST
nicolas.medoc@list.lu

Mohammad Ghoniem
ERIN-eScience, LIST
L-4422, Belvaux, Luxembourg
mohammad.ghoniem@list.lu

Mohamed Nadif
LIPADE, Université de Paris-Descartes
75006, Paris, France
mohamed.nadi@parisdecartes.fr

RÉSUMÉ
Dans des corpus textuels volumineux, les journalistes analytiques cherchent des documents et des récits qui corroborent des faits, en les examinant sous tous les angles. Nous présentons un outil de visualisation analytique leur permettant de vérifier, d'affiner et de générer des hypothèses sans avoir à lire la totalité des contenus. Notre système repose sur une approche hybride de biclustering. Les sujets de haut niveau sont présentés via une carte pondérée de sujets, reflétant à la fois leur importance et leur similarité relative. Pour chaque sujet, une vue hiérarchique et interactive dresse un aperçu de toutes ses variantes, de manière à identifier les documents traités sous un même angle ou partageant des faits communs. Des vues multiples et coordonnées permettent une analyse plus fine, en filtrant, sélectionnant et comparant les variantes de sujet, au regard des motifs de co-occurrence de termes les plus intéressants. L'utilité de l'outil est montrée par un scénario d'usage, puis évaluée qualitativement par un journaliste analytique.

Mots Clés
Visualisation de textes ; biclustering ; journalisme analytique.

ABSTRACT
In large text corpora, analytic journalists need to identify facts, verify them by locating corroborating documents and survey all related viewpoints. This requires them to make sense of document relationships at two levels of granularity : high-level topics and low-level topic variants. We propose a visual analytics software allowing analytic journalists to verify and refine hypotheses without having to read all documents. Our system relies on a hybrid biclustering approach. A new *Weighted Topic Map* visualization conveys all top-level topics reflecting their importance and their relative similarity. Then, coordinated multiple views allow to drill down into topic variants through an interactive term hierarchy visualization. Hence, the analyst can select, compare and filter out the subtle co-occurrences of terms shared by multiple documents to find interesting facts or stories. The usefulness of the tool is shown through a usage scenario and further assessed through a qualitative evaluation by an expert user.

Author Keywords
Text Visualization ; biclustering ; analytic journalism.

ACM Classification Keywords
H.5.2. Information Interfaces and Presentation: Graphical user interfaces (GUI)

INTRODUCTION

Les journalistes analytiques font face à un dilemme : alors que les sources d'information croissent continuellement, le temps accordé à l'investigation est toujours plus réduit par les rédactions. Nous proposons une approche de visualisation analytique, conçue pour et avec des journalistes analytiques, permettant une analyse exploratoire de corpus de textes libres collectés durant leur travail de terrain. Les délais accordés ne permettent pas aux journalistes de lire des volumes importants de documents. Ils cherchent malgré tout à produire un travail exhaustif. Nous proposons de répondre à ce défi avec un outil au service de deux tâches : 1) rechercher des documents vérifiant une hypothèse donnée, et 2) identifier des angles et points de vue inédits, dans le but d'affiner ou de générer de nouvelles hypothèses qui collent véritablement aux faits trouvés.

L'un des défis consiste à éviter de lire la totalité des documents pour y trouver des actifs utiles à l'enquête et à la narration. Nous proposons d'aborder ce défi en résumant le corpus avec des techniques d'extraction de sujets. Un sujet est décrit par un ensemble de termes associés aux documents qui en traitent. Un corpus de plusieurs milliers de documents peut receler une cinquantaine de sujets (cf. scénario d'utilisation). L'analyste doit pourtant localiser rapidement ceux qui l'intéressent et comprendre leurs liens. Nous proposons ainsi une nouvelle visualisation sous forme d'une *carte pondérée des sujets*. Basée sur des nuages de mots incrustés dans une *Treemap*, cette visualisation reflète l'importance et la similarité relative des sujets. Ensuite, les journalistes cherchent à multiplier les sources contenant des faits, des points de vue ou des angles d'analyse en commun. Nous proposons donc un outil permettant d'analyser chaque sujet par les relations de co-occurrences de termes partagées par des sous-ensembles de documents. Ces relations sont appelées des *variantes de sujet*. L'analyse d'un sujet du point de vue de ses variantes

http://dx.doi.org/10.1145/3004107.3004116

Figure 1. La carte pondérée de 50 sujets (nuages de mots) extraits de diverses sources d'actualité en ligne du 2 au 16 novembre 2015. Leur taille et leur proximité reflètent leur importance et leur similarité relative. L'interaction de survol affiche 4 liens réciproques de proximité.

permet de focaliser le travail sur une pré-sélection de documents potentiellement utiles aux tâches des journalistes. Nous adoptons alors une approche multi-résolution alternant aussi bien des analyses descendantes sur le modèle vue d'ensemble, zoom et filtrage, détails à la demande [36], que des analyses ascendantes, associant des éléments de détails (termes et documents) de proche en proche par des relations de plus haut niveau (variantes de sujets, sujets).

Les techniques de biclustering [21, 31] appliquées à des matrices *Termes*×*Documents* considèrent simultanément la dualité entre les vecteurs des documents et de leurs termes. Notre système profite des avantages de deux structures imbriquées reposant sur deux techniques de biclustering distinctes (Figure 2). La première exploite la modularité de graphe pour identifier des biclusters diagonaux [1] et constituer les sujets de haut niveau. La seconde technique, `Bimax` [33], est une méthode de biclustering non-disjoint. Appliqué à chaque sous-matrice constitutive d'un sujet, cet algorithme révèle toutes les combinaisons de termes partagées par un ensemble de documents représentant les variantes de sujets. Un autre défi se situe alors au niveau de l'exploration et de l'interprétation du grand nombre de biclusters non-disjoints (*variantes de sujet*) produits par `Bimax`. Nous proposons de les visualiser en les disposant dans une hiérarchie de termes construite sur la base des chevauchements de biclusters. Ainsi, chaque *variante* est représentée par les termes d'une branche de l'arbre. L'objectif est d'explorer les *variantes* en partant des termes les plus redondants (aux racines), représentant le sujet dans ses grandes lignes, vers les termes les plus spécifiques (aux feuilles), tout en sélectionnant les documents concernés par chaque séquence de termes. Des vues multiples et coordonnées proposent à l'expert des interactions de tri, de filtrage et de comparaison pour identifier les termes les plus informatifs et sélectionner des variantes pertinentes.

Nos contributions, intégrées dans un système d'exploration de corpus de textes, sont multiples : 1) Nous combinons dans une structure imbriquée une méthode de biclustering diagonal, basée sur la modularité de graphe, avec `Bimax`, une méthode de biclustering non-disjoint. 2) Nous proposons des vues multiples et coordonnées pour explorer et analyser un grand nombre de biclusters non-disjoints. 3) Nous proposons une nouvelle carte pondérée de sujets, sous la forme de nuages de mots incrustés dans une *Treemap* reflétant leur taille et leur similarité relative.

Dans la suite de cet article, nous décrivons les tâches et les données prises en compte pour la conception du système. Ensuite, nous dressons un état de l'art qui commence par les outils existants de visualisation analytique de corpus textuels, se poursuit par la visualisation de textes et de sujets, et se termine par les méthodes de biclustering et leur visualisation. Puis, nous décrivons chaque composant de notre système, comprenant la chaîne des traitements analytiques ainsi que les choix de conception visuelle. Nous poursuivons par notre évaluation suivie d'une discussion sur les limites et avantages de notre système, et des travaux futurs envisagés. Enfin, nous dressons notre conclusion.

ABSTRACTION DES TÂCHES ET DES DONNÉES

Nous adoptons le modèle imbriqué de Munzner [32]. Nous décrivons ci-dessous la caractérisation du problème, des tâches et des données, et dans des sections dédiées, les algorithmes et l'encodage visuel.

Caractérisation du problème. Notre analyse a débuté par une étude de terrain. Nous avons conduit un entretien semi-dirigé avec une journaliste analytique pour comprendre ses méthodes de travail et ses difficultés face à l'analyse de grands corpus. En général, elle a exprimé un sentiment de frustration, n'ayant pas les moyens d'être exhaustive lors de ses enquêtes, par manque de temps. Elle est souvent forcée de réduire le nombre de documents collectés lors de son travail de terrain. Un outil permettant de résumer les contenus et d'en extraire toutes les relations pourrait faciliter et accélérer l'identification d'actifs sans nécessiter une lecture complète du corpus. Nous nous sommes appuyés aussi sur l'ouvrage de Lee Hunter et al. [29] stipulant que « une hypothèse est une histoire et une méthode pour la vérifier ». Nous avons alors extrait un workflow composé de trois processus adoptés par les journalistes de manière alternée :

— **Cartographie du sujet :** le journaliste cherche à obtenir une vue d'ensemble du sujet d'investigation.

— **Focalisation :** le journaliste réduit son angle d'analyse à un aspect spécifique pour identifier des faits et des points de vue qui valident ou réfutent ses hypothèses.

— **Diversification :** le journaliste élargit son angle, en regardant autour de l'objet de recherche, en vue de trouver des informations inattendues ou des angles d'analyse inédits. Cette étape permet de s'assurer de ne pas rater de faits remettant en question les hypothèses initiales.

Abstraction des tâches. Lors de notre analyse préliminaire, nous avons identifié le besoin d'analyser un grand corpus de textes libres via un outil exploratoire conçu pour trois tâches subdivisées en sous-tâches :

T1 Résumer le corpus, identifier les sujets d'intérêt ainsi que les aspects à investiguer
T1.1 Comprendre et localiser les sujets
T1.2 Comprendre et identifier les variantes des sujets
T2 Chercher les documents qui vérifient une hypothèse
T2.1 Rechercher des variantes de sujet par mots-clés
T2.2 Comparer des variantes de sujets
T2.3 Accès aux contenus, voir les termes dans leur contexte
T3 Identifier des angles d'analyse inédits et des points de vue qui incitent les journalistes à affiner ou générer de nouvelles hypothèses conformes aux faits trouvés
T3.1 Découvrir les sujets similaires, leurs relations
T3.2 Suggérer des termes pour affiner la recherche
T3.3 Suggérer des variantes partageant des documents/termes

Abstraction des données. Un modèle vectoriel représente le corpus sous la forme d'une matrice *Termes*×*Documents*. Chaque document est représenté par un vecteur de termes distincts dont les valeurs sont définies par le schéma de pondération *Term Frequency - Inverse Document Frequency (TF-IDF)* [37]. Nous introduisons ci-dessous les notations utilisées dans cet article. Pour un ensemble I donné de n documents, et un ensemble J de m termes, la matrice *Termes*×*Documents* est définie par $X = \{e_{ij}, i \in [1..n], j \in [1..m]\}$. Grâce aux poids *TF-IDF*, chaque cellule e_{ij} mesure à quel point le terme j est représentatif du document i, tout en évitant les mots vides de sens qui apparaissent dans beaucoup de documents (*IDF*). Nous exploitons ce schéma de pondération aussi bien dans les traitements analytiques que dans les visualisations. Les sujets latents et les variantes de sujet sont modélisés par des biclusters formalisés dans la section *Traitements analytiques*.

ÉTAT DE L'ART

Visualisation analytique pour des corpus textuels

Feature Lens [13] est dédié à l'investigation de corpus textuels par l'exploration des motifs fréquents de mots ou de n-grams. Divers modes de tri font émerger des motifs localisables dans leurs documents. Cette solution présente les motifs sous forme de listes sans en offrir une vue d'ensemble. Overview [7] est un outil de visualisation analytique consacré à l'analyse de corpus textuels par des journalistes d'investigation. Basé sur une méthode de clustering hiérarchique, cet outil résume de grands corpus et couvre deux tâches similaires aux nôtres : vérifier et générer des hypothèses. Les clusters de documents sont représentés par les termes les plus fréquents. Notre approche permet de focaliser l'analyse sur un sujet et d'explorer les multiples angles et points de vue avec la précision et l'exhaustivité offerte par Bimax. L'outil d'Alexander et al. [3] permet l'exploration de corpus textuels notamment par des approches analytiques à la fois descendantes et ascendantes favorisant la sérendipité. Lors de notre analyse des méthodes de travail des journalistes, nous avons identifié des processus de focalisation et de diversification, qui s'y apparentent. Nous avons mis en œuvre ces deux processus en plaçant l'analyse au niveau des *variantes de sujet*, un niveau d'abstraction entre les sujets de haut niveau et les éléments de base du corpus que sont les documents et leurs termes. Nous proposons alors une exploration multi-résolution basée sur une structure hybride de biclusters.

Visualisation de textes et de sujets

La visualisation de textes est abordée soit en considérant leurs éléments constitutifs, tels que les termes ou les documents, soit en considérant des données dérivées telles que les sujets extraits des contenus. Dans la première catégorie, les techniques de projection comme l'analyse en composantes principales (*ACP*) ou le positionnement multidimensionnel (*MDS*) sont largement utilisées pour visualiser les clusters de documents via des nuages de points, comme dans IN-SPIRE [47]. L'étude de Brehmer et al. [7] indique que les journalistes préfèrent une navigation hiérarchique, approche que nous adoptons pour l'exploration des *variantes de sujet*. Pour comprendre les contenus textuels, sans pour autant lire la totalité du corpus, les nuages de mots [44] mettent en exergue les mots les plus importants par leur taille et leur couleur. Plusieurs extensions ont récemment été proposées pour y intégrer la dimension temporelle [27] ou une hiérarchie [18, 45]. L'espace peut être subdivisé verticalement en un nombre limité de sujets, les termes communs pouvant être reliés horizontalement [9].

Dans la seconde catégorie d'approches, de nombreux outils proposent des solutions d'exploration de sujets. Basés sur Latent Dirichlet Allocation (LDA) [5] ou sur des techniques dérivées telles que hLDA [22] ou HDP [43], Tiara [46], ParallelTopics [14], TextFlow [10] et LeadLine [15] montrent l'évolution temporelle des sujets via des variantes de Theme River [25] où chaque couche représente un sujet. HierarchicalTopics [16] propose une organisation hiérarchique pour faciliter l'exploration de sujets nombreux. Mais chaque sujet est représenté par une liste composée des N termes les plus importants. TopicPanorama [30] donne une vue d'ensemble des sujets extraits, tout en distinguant les éléments communs ou les spécificités entre des sources de différentes origines. Conçu pour des corpus de très grande taille, cet outil offre plusieurs niveaux de résolution, chacun représenté par un graphe permettant d'apprécier la similarité entre les sujets. Si certains sujets sont étiquetés par deux ou trois mots, un nuage de mots plus exhaustif est déployé interactivement. Notre carte pondérée de sujets permet une interprétation directe et exhaustive de tous les sujets à travers des nuages de mots reflétant leur taille et leur similarité relative. Nous proposons en sus une exploration des *variantes de sujet* capturées par les relations subtiles de documents basées sur les motifs de co-occurrences de termes.

Biclustering et visualisation

La majorité des solutions de visualisation analytique de corpus textuels se base sur les modèles de sujets (*Topic Models*) tels que *LDA* [5] et Non-Negative Matrix Factorization [28, 8]. La sémantique des sujets est souvent représentée par les *N* termes les plus fréquents. Cependant, Alexander et al. [2] ont montré qu'un sujet est bien plus que ses dix premiers mots. Ils ajoutent que la qualité sémantique des sujets dépend de la capacité des modèles à refléter les motifs subtils de co-occurrence de termes.

Le biclustering, aussi connu sous le nom de co-clustering, est largement utilisé dans le domaine de la bio-informatique [31, 21, 33] et a été étendu aux systèmes de recommandation et à l'analyse de texte. En somme, ces méthodes s'appliquent à toute matrice dont les cellules représentent une relation entre les lignes et les colonnes. Si les techniques de biclustering considèrent simultanément la dualité qui existe entre les documents et leurs termes, les techniques dérivées de LDA reposent entièrement sur un modèle génératif pour la dimension des documents [35]. Notre système utilise donc les techniques de biclustering pour obtenir des biclusters cohérents groupant des documents similaires avec leurs termes les plus représentatifs. Notre système combine deux méthodes. Une méthode de biclustering diagonal [1] à partitionnement strict, basée sur la modularité de graphe, a montré des performances supérieures à l'approche spectrale de Dhilon et al. [12] sur des données textuelles. Nous l'utilisons pour extraire les sujets de haut niveau. Leurs relations de similarité sont visualisées, compensant en partie les limites du partitionnement strict, comme décrit dans la section discussion. Prelić et al. [33] évaluent plusieurs méthodes de biclustering sur des données d'expression de gênes. Ils proposent un algorithme de biclustering non-disjoint nommé `Bimax` vérifiant une contrainte d'inclusion maximale. Nous l'utilisons pour extraire toutes les variantes d'un sujet donné.

L'exploration de la pléthore de biclusters non-disjoints produits par `Bimax` nécessite un outil de visualisation analytique conçu à cet effet. Dans leur revue de littérature, Sun et al. [42] proposent un cadre de conception pour la visualisation de biclusters sur cinq niveaux de relations : les relations d'entités (1-1), de groupes (1-n), de biclusters (n-m), les relations chaînées (n-m-...-z), et les schémas relationnels. Nos travaux s'inscrivent essentiellement au niveau des relations de biclusters (n-m). BicOverlapper [34] visualise des biclusters produits par `Bimax` avec un diagramme nœuds-liens, un modèle basé sur les champs de force et l'ajout de contours. Mais dans les zones denses avec beaucoup de chevauchements, les entrecroisements et superpositions ne permettent pas de distinguer les parties communes et distinctives des biclusters.

Les visualisations matricielles et les coordonnées parallèles s'avèrent plus efficaces que les diagrammes nœuds-liens pour l'analyse des relations de biclusters [42]. De plus, bien que moins intuitives, les vues matricielles sont plus lisibles lorsque les données sont volumineuses et denses [20]. Ainsi, beaucoup de solutions en bio-informatique proposent des vues coordonnées combinant des *heatmap* matricielles et des coordonnées parallèles [4, 26, 34]. Les éléments des deux dimensions y sont placés séparément, et ordonnés en ligne et en colonne de manière à faire apparaître les biclusters individuellement. Mais ces réarrangements linéaires ne permettent pas de dresser une vue d'ensemble des biclusters non-disjoints, sans dupliquer leurs éléments [26, 40]. En réponse à ces problèmes, Streit et al. [40] proposent une visualisation hybride dans laquelle les biclusters sont représentés par des blocs matriciels, reliés par les lignes et colonnes qu'ils ont en commun. Bixplorer [17] suit une approche similaire pour des données textuelles afin d'explorer les relations chaînées entre différentes catégories d'entités nommées. Suivant une approche analytique ascendante, le système est efficace pour une sélection réduite de biclusters. Pour permettre une approche analytique descendante, BiSet [41] propose une visualisation interactive avec des coordonnées parallèles similaire à Jigsaw [39]. Les relations chaînées (n-m-...-z) sont représentées par des faisceaux de liens sémantiques formés par les biclusters groupant les éléments de deux coordonnées adjacentes. Pour les relations de biclusters (n-m) entre les termes et les documents, nous organisons les biclusters selon une hiérarchie de termes basée sur leurs chevauchements. Une vue matricielle coordonnée permet ensuite de comparer une sélection réduite d'intérêt.

VUE D'ENSEMBLE DU SYSTÈME

L'outil présenté dans cet article permet une exploration des sujets selon différentes résolutions, des sujets de haut niveau jusqu'aux contenus bruts des documents. Il comprend quatre composants visuels couvrant l'ensemble des tâches T1 à T3 : la carte pondérée de sujets dans la Figure 1, puis dans la Figure 4, la vue d'ensemble des variantes de sujet (3), le comparateur de variantes (4) et la vue détaillée des documents (5). Le journaliste commence son travail en scrutant la carte pondérée des sujets, obtenant ainsi une vue d'ensemble des sujets (**T1.1**) extraits du corpus par la méthode de biclustering diagonal. En sélectionnant un sujet d'intérêt, les *variantes de sujets* sont caractérisées par *Bimax* et organisées hiérarchiquement sur la base des termes partagés dans la Figure 4 (3.1). L'analyste peut ensuite explorer toutes les variantes de sujet via une visualisation radiale d'arbre telle que *Sunburst* (**T1.2**). Celle-ci est utilisée pour démarrer le processus de focalisation afin de vérifier les aspects spécifiques des hypothèses de travail. L'expert peut ensuite filtrer les variantes de sujet par mots-clés pour évacuer les variantes inintéressantes (**T2.1**). Ensuite, dans la Figure 4, un sous-ensemble de *variantes de sujet* est choisi pour une inspection approfondie dans le comparateur (4). Cette vue place les variantes en colonne dans une matrice afin d'apprécier la distribution des termes (4.1) et des documents (4.2) (**T2.2**). Divers modes de tri (4.3) fournissent des perspectives alternatives sur les données aidant ainsi le journaliste à trouver les termes les plus informatifs. Enfin, le niveau de détail le plus bas (5) affiche le texte intégral apportant la sémantique précise des termes utilisés dans leur contexte. Le document est évidement le matériel utilisé par les journalistes comme actif pour constituer les éléments de preuve en lien avec les hypothèses.

Le processus de diversification est mis en œuvre par des interactions combinées dans le but de promouvoir la sérendipité. Sur la carte pondérée des sujets, l'expert navigue

dans le voisinage d'un sujet d'intérêt pour suivre les liens suggérés vers des sujets similaires et pousser plus loin son investigation (**T3.1**). Puis dans la Figure 4, la vue d'ensemble des *variantes de sujet* (3), suggère des termes provenant des documents que l'expert n'a pas forcément à l'esprit afin d'étendre la portée de son analyse (**T3.2**). Enfin, la vue d'ensemble des variantes de sujet met en exergue ces dernières lorsqu'elles partagent des termes ou des documents explorés aussi bien dans le comparateur que dans la vue d'ensemble elle-même (**T3.3**). Notre système est construit sur un empilement de traitements analytiques décrits ci-dessous.

TRAITEMENTS ANALYTIQUES

Notre outil repose sur une structure imbriquée combinant une méthode hybride de biclustering présentée dans la Figure 2. Notre outil implémente une architecture Web alliant les technologies Java, Scala et Python pour le backend et javascript avec *D3.js* pour les visualisations.

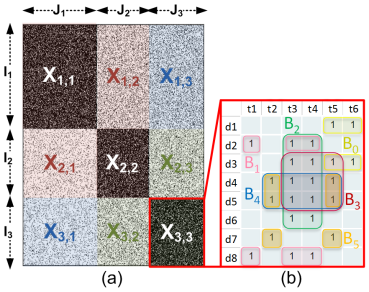

Figure 2. a) Les biclusters diagonaux (sujets) et les blocs de confusion utilisés par l'équation (2), b) les biclusters en chevauchement (variantes de sujet) produits par `Bimax` pour chaque sujet, p. ex. $X_{3,3}$.

Traitement du texte

D'abord, le texte est traité par une librairie de Traitement Automatique des Langues (*Stanford CoreNLP*). On y extrait notamment la nature des mots, pour ne garder que les noms et les verbes. Ces derniers sont suffisamment informatifs pour interpréter les sujets et leurs variantes. Ensuite, la matrice *Termes*×*Documents* est créée. Seuls les 10 000 termes ayant les plus grandes valeurs *TF-IDF* sont gardés. Ce nombre est configurable et la matrice ainsi réduite est traitée par le biclustering diagonal.

Extraction des sujets avec le biclustering diagonal

Extraction des sujets

Les méthodes de biclustering exploitent la dualité qui existe entre les documents et leurs termes pour grouper les termes sémantiquement proches et les documents où ils apparaissent ensemble. Pour extraire les sujets du corpus, nous utilisons l'algorithme de biclustering diagonal proposé par Ailem et al. [1]. Cet algorithme s'appuie sur la modularité de graphe, adaptée pour y incorporer simultanément les partitions en ligne et en colonne. Les biclusters sont décrits comme suit. Pour un nombre donné K de biclusters, et une matrice X de taille $n \times m$ avec I l'ensemble des n documents et J l'ensemble des m termes, le bicluster $B_k, k \in [1..K]$ est la sous-matrice $I_k \times J_k$ avec $I_k \subseteq I$ et $J_k \subseteq J$. Le partitionnement strict des lignes et des colonnes est garanti par la contrainte suivante : $\forall k,l \in [1..K]$ avec $k \neq l$, $I_k \cap I_l = \emptyset$ et $J_k \cap J_l = \emptyset$. Pour un corpus donné, le nombre de sujets (K) est un paramètre important qui doit être choisi attentivement. Une approche consiste à faire varier K pour garder la valeur optimale selon le critère d'optimisation utilisé par l'algorithme [1]. Pour chaque valeur de K, nous exécutons un nombre significatif de tests (200) car l'algorithme initialise les partitions de manière aléatoire et converge vers un optimum local.

Relations entre les sujets

L'affichage des relations entre les sujets repose sur une mesure de similarité. Hanczar et Nadif [24] utilisent la mesure suivante basée sur l'indice de Jaccard pour des biclusters non-disjoints :

$$Sim(B_k,B_l) = \frac{|B_k \cap B_l|}{|B_k \cup B_l|} = \frac{|B_k \cap B_l|_I + |B_k \cap B_l|_J}{|B_k \cup B_l|_I + |B_k \cup B_l|_J} \quad (1)$$

où $|.|_I$ est la cardinalité de l'ensemble des lignes (les documents) et $|.|_J$ est la cardinalité de l'ensemble des colonnes (les termes). Or, pour les biclusters diagonaux, $|B_k \cap B_l|_I = 0$ et $|B_k \cap B_l|_J = 0$. Cette mesure de similarité doit donc être adaptée, par exemple en incluant les informations recelées dans les blocs adjacents aux blocs diagonaux. La partition en diagonale subdivise la matrice X en $K \times K$ sous-matrices $X_{l,k}$ (Figure 2), $k \in [1..K]$ se réfère à la partition en ligne et $l \in [1..K]$ à la partition en colonne. À chaque bicluster B_k correspond un bloc diagonal $X_{k,k}$. Pour une paire donnée de biclusters diagonaux (B_k, B_l), $X_{k,l}$ et $X_{l,k}$ forment les blocs de confusion indiquant si les biclusters partagent des lignes ou des colonnes. Notre mesure de similarité calcule ainsi l'intersection en ligne et en colonne entre les biclusters et les blocs de confusion. Notons $I_{k,l} = \{i \in I_k : \exists j \in J_l, e_{ij} \in X_{k,l}\}$ et $J_{k,l} = \{j \in J_l : \exists i \in I_k, e_{ij} \in X_{k,l}\}$ les ensembles des lignes, respectivement des colonnes, non vides de $X_{k,l}$. La mesure de similarité devient :

$$Sim'(B_k,B_l) = \frac{|I_{k,k} \cap I_{k,l}| + |I_{l,l} \cap I_{l,k}| + |J_{k,k} \cap J_{l,k}| + |J_{l,l} \cap J_{k,l}|}{|I_k \cup I_l| + |J_k \cup J_l|} \quad (2)$$

Nous l'utilisons pour construire la matrice de similarité des sujets permettant de visualiser leurs relations et de calculer leur proximité spatiale dans la carte pondérée des sujets. L'étape de traitement suivante construit la structure du second niveau de détail dédié aux variantes de sujet.

Extraction des variantes de sujet avec `Bimax`

Extraction des variantes de sujet

Un journaliste vérifie ses hypothèses en multipliant les sources relatant les mêmes faits ou les mêmes récits. Nous supposons que les corpus agrégés par les journalistes contiennent de telles redondances. Ainsi, extraire les relations de co-occurrence entre les documents via les techniques de biclustering permet, à la fois, d'identifier les termes relatifs à des faits ou des récits, et de retrouver l'ensemble des documents qui les partagent. Les termes et les documents pouvant être liés à plusieurs faits ou récits, les méthodes de biclustering non-disjoints sont appropriées.

Prelić et al. [33] proposent un tel algorithme nommé *Bimax*, avec une implémentation en java [4]. Il s'applique à des matrices binaires pour identifier des blocs dont les cellules sont uniquement composées de 1 (Figure 2b). Prelić et al. formalisent une contrainte d'inclusion maximale [33], garantissant qu'aucun bicluster n'est complètement recouvert

par un autre. À l'inverse, l'algorithme étend au maximum chaque bicluster en ligne et en colonne. Sur une matrice *Termes*×*Documents* binarisée, `Bimax` identifie toutes les combinaisons distinctes de termes partagées par un ensemble de documents. Les biclusters de `Bimax`, que nous appelons aussi *variantes de sujet*, peuvent être représentatifs de faits, d'angles inédits ou de points de vue partagés par plusieurs documents.

Human in the loop

Dans les blocs diagonaux de la structure englobante, on constate des variations de taille et de densité selon les sujets. Mais avec *Bimax*, le nombre de biclusters augmente avec la taille et la densité de la matrice.

`Bimax` s'appliquant à des matrices binaires, nous avons donc défini un seuil de binarisation configurable, appelé « seuil d'intérêt », qui s'applique aux valeurs *TF-IDF* de la matrice *Termes*×*Documents*. Non seulement ce seuil réduit la densité de la matrice, mais aussi sa taille lorsque des vecteurs sont totalement mis à 0. Ensuite, `Bimax` possède trois paramètres : le nombre minimum de lignes et de colonnes par bicluster (*MinRows* et *MinCols*), et un nombre maximum de biclusters fixé pour arrêter l'algorithme (*MaxBC*). Augmenter *MinCols* ignore les motifs de co-occurrence ayant trop peu de termes. Augmenter *MinRows* ignore les motifs de co-occurrence trouvés dans trop peu de documents. Par définition des biclusters, $MinRows \geq 2$ et $MinCols \geq 2$. Ensuite, nous avons observé que, pour des données textuelles, les temps d'exécution augmentent drastiquement lorsque le nombre de biclusters atteint 10 000. Nous choisissons cette valeur pour *MaxBC*. Mais lorsque cette limite est atteinte, les résultats sont difficilement exploitables. La densité de la sous-matrice doit donc être réduite en augmentant le *seuil d'intérêt* pour ne garder que les termes les plus significatifs. Ces paramètres permettent à l'expert de piloter `Bimax` jusqu'à obtenir des résultats interprétables et intéressants.

CONCEPTION DES VISUALISATIONS

Carte pondérée des sujets

Afin de dresser une vue d'ensemble des sujets traités, nous proposons une nouvelle visualisation, la carte pondérée des sujets (Figure 1). Elle combine : 1) une projection MDS des sujets dans un plan 2D générant des coordonnées spatiales utilisées par 2) une visualisation nommée *Weighted Map*, une variante des *Treemaps* garantissant une cohérence spatiale, où chaque nœud est représenté par 3) un nuage de mots. Par exemple, la Figure 1 représente 50 sujets extraits de sites d'information en ligne entre le 2 et le 16 novembre 2015. La taille des sujets (rectangles) est proportionnelle au nombre de termes et au nombre de documents qu'ils contiennent. Leur proximité dans le plan 2D reflète leur similarité. Nous discutons des différents éléments de la carte pondérée des sujets ci-dessous.

Les nuages de mots

L'étape de biclustering diagonal extrait des sujets, chacun regroupant un ensemble de termes sémantiquement cohérents par rapport aux documents. Au lieu d'afficher les N termes les plus fréquents de chaque sujet, nous considérons tous leurs termes et montrons leur importance relative. Dans ce but, nous utilisons des nuages de mots. La taille des termes correspond à un critère d'intérêt calculé par la somme des poids *TF-IDF* sur tous les documents du sujet dans lesquels ils apparaissent ($\log(1+\sum_{i\in I_k} e_{ij}), \forall j \in J_k$). Une transformation logarithmique permet de ne pas écraser le bas de l'échelle de valeurs. L'intensité de la couleur correspond au nombre de documents contenant le terme dans le sujet, la couleur s'assombrissant lorsque le nombre de documents augmente. Avec cet encodage visuel, le journaliste peut identifier rapidement des motifs d'intérêt variés. Nous utilisons le modèle de nuage de mots implémenté par Davies avec D3.js [11]

Taille et positionnement des sujets dans la carte

Nous allouons aux sujets des poids relatifs à leur taille (le produit du nombre de termes et du nombre de documents). Une projection MDS de la matrice de similarité (équation 2) génère des coordonnées 2D pour les sujets. Les poids et les coordonnées 2D servent à générer une vue *Weighted Map*, une variante de *Treemap* proposée par Ghoniem et al. [19] pour des données hiérarchiques géo-référencées.

De cette manière, les nuages de mots sont incrustées dans les rectangles de la vue *Weighted Map*. Leur taille encode l'importance de leur sujet et leur proximité traduit leur similarité relative. Grâce à cette visualisation, l'analyste peut embrasser l'ensemble des sujets couverts dans le corpus, localiser ceux qui l'intéressent, et naviguer dans le voisinage pour découvrir des sujets similaires. De plus, le survol d'un sujet affiche par transparence des liens vers les 5 sujets les plus proches. La couleur des liens encode la force de la relation, donnée par la métrique de similarité de l'équation 2. Mais, avec cette dernière, les sujets les plus grands tendent à « attirer » les plus petits. Nous appliquons alors un premier filtre pour sélectionner les 20 sujets les plus proches sur la base des liens réciproques (pour une paire de sujets chacun doit apparaître dans le top 20 de l'autre). Puis, nous affichons seulement les 5 premiers parmi les candidats restants. À ce stade, l'analyste peut cliquer sur un sujet pour l'analyser en détail et explorer toutes les relations entre les documents dans la vue d'ensemble des variantes de sujet. Leur nombre est d'ailleurs indiqué en haut à gauche de chaque sujet.

Vue d'ensemble des variantes de sujet

Les *variantes de sujet* captent dans chaque sujet les relations entre les documents extraites par `Bimax`. Sous la contrainte d'inclusion maximale, un bicluster regroupe un ensemble unique de documents partageant une combinaison unique de termes. Ainsi, `Bimax` identifie tous les biclusters optimaux assurant l'exhaustivité recherchée par les journalistes. En contrepartie, cette exhaustivité se traduit par un nombre important de biclusters se chevauchant par les termes ou les documents, causant des problèmes d'interprétation majeurs. Pour donner du sens à toutes ces variantes de sujet et leurs nombreuses intrications, nous avons conçu une visualisation hiérarchique (Figure 4). En effet, l'exploration hiérarchique des sujets est plus efficace qu'une exploration plate basée sur des listes [16] et préférée par les journalistes utilisant Overview [7].

Hiérarchie de termes

L'interprétation sémantique des *variantes de sujet* passe principalement par l'analyse de leurs termes. Nous pro-

posons alors une hiérarchie de termes qui organise les *variantes de sujet* (les biclusters) sur la base des chevauchements de biclusters selon leurs termes (Figure 3). Nous observons qu'un terme avec un degré de chevauchement élevé (apparaissant dans un grand nombre de biclusters) tend à décrire le sujet de haut niveau dans ses grandes lignes. De tels termes (ex : « Clinton, Obama, Israël, Netanyahu ») apparaissent aussi dans beaucoup de documents, sélectionnés par les biclusters en chevauchement. Il est donc pertinent de placer ces termes génériques dans les premiers niveaux de la hiérarchie. À l'inverse, les termes avec un degré de chevauchement faible sont plus spécifiques aux *variantes de sujets* (ex : « jewish, secretary »), voire sont exclusifs à une variante. Nous les plaçons dans les niveaux plus profonds des branches de la hiérarchie.

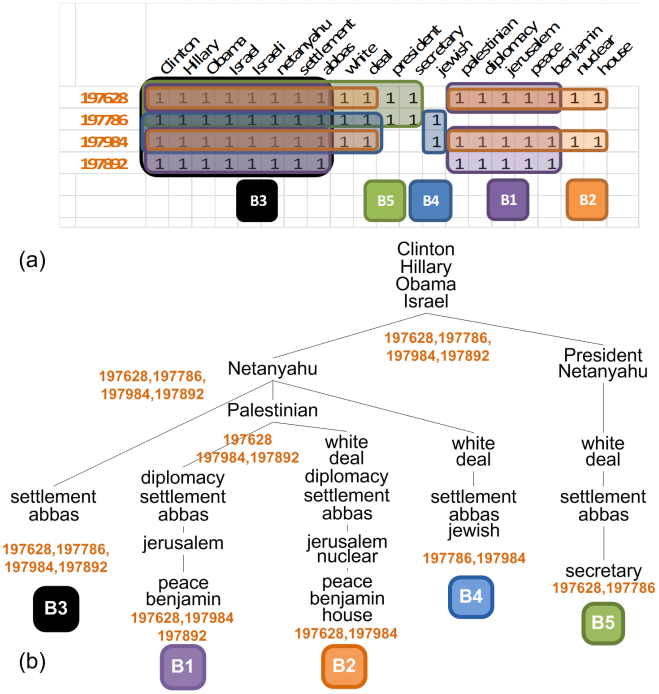

Figure 3. a) Représentation matricielle des biclusters (B1-B5) sélectionnés dans le comparateur de la figure 4. b) La hiérarchie de termes : chaque bicluster est inséré en associant ses termes triés par degré de chevauchement. Les numéros oranges identifient les documents. Le nœud *Palestinian* a deux biclusters en chevauchement : B_1,B_2. L'union des documents de ces biclusters est *197628,197984,197892*. Ces documents partagent l'ensemble de la séquence de termes définie par le chemin *« Clinton, ... , Israël, Netanyahu, Palestinian »*.

Les nœuds de la hiérarchie sont formés par les termes des *variantes de sujet* (biclusters). Une branche est une séquence unique de termes décrivant une *variante*, commençant par les termes les plus génériques au niveau des racines et se terminant par les termes les plus spécifiques au niveau des feuilles. Cette hiérarchie est construite en utilisant l'algorithme *FPTree* [23]. Les termes du sujet sont d'abord triés par degré de chevauchement dans l'ordre décroissant. En cas d'égalité, un tri alphabétique est appliqué pour garantir un ordre global unique à chaque sujet. Chaque bicluster est inséré dans l'arbre à partir des racines et en associant ses termes un à un en suivant l'ordre global. L'insertion d'un terme nouveau provoque la création d'une nouvelle branche ancrée à partir du dernier terme associé. Le bicluster est placé sur la feuille de cette branche.

Ainsi, à chaque niveau de la hiérarchie se trouvent des termes ayant différents degrés de chevauchement. Les termes regroupant plusieurs branches peuvent être vus comme des points d'articulation guidant progressivement l'analyste pour trouver des *variantes de sujet* en adéquation avec ses besoins. En outre, chaque nœud de la hiérarchie sélectionne tous les documents de ses biclusters. Lorsque l'analyste se déplace vers les feuilles le long d'une branche, les documents considérés sont progressivement filtrés. Ils partagent des séquences de termes plus longues mais plus spécifiques. Cette navigation permet de focaliser progressivement l'analyse autour d'un angle précis.

Visualisation interactive

La hiérarchie décrite ci-dessus est représentée par une visualisation radiale *Sunburst* [38], implementée dans D3.js [6]. Dans la Figure 4 (3.1), chaque branche représente une séquence complète de termes d'une *variante de sujet*. Lorsque l'analyste survole un nœud, le terme associé apparaît dans une info-bulle et la séquence complète du chemin est affichée verticalement sur la droite (3.2). De plus, le terme survolé est coloré en rouge dans toutes les branches (3.3) où il apparaît. En cliquant sur un nœud, les documents associés sont listés dans la vue détaillée des documents (5). Un premier objectif est de trouver des termes porteurs d'informations qui confirment ou infirment les hypothèses de travail. Un deuxième objectif est de suggérer des combinaisons de termes aidant l'analyste à trouver des points de vue inattendus ou à exprimer des requêtes qui concordent avec le contenu.

Nous proposons trois modes d'interaction (2) facilitant la réalisation des tâches **T1** à **T3**. 1) Le mode *Filtrage* distingue en bleu les chemins contenant les termes saisis dans un champs de recherche (**T1.2, T2.1**), les autres branches restant dans les nuances de gris. Les *variantes de sujet* grisées peuvent être cachées ou montrées sur simple clic. 2) Le mode *Distribution des documents* permet à l'analyste de distinguer en orange tous les chemins contenant au moins l'un des documents sélectionnés par le nœud cliqué. Cela aiguille l'analyste vers de nouvelles *variantes de sujet* qui partagent les documents sélectionnés (**T3.3**). Ces variantes peuvent aussi ramener de nouveaux documents révélant de nouveaux angles, points de vues ou faits partagés avec des documents déjà connus. 3) Le mode *Sélection pour comparaison* sélectionne les chemins envoyés au comparateur de variantes (**T2.2**).

Comparateur de variantes

Le comparateur de variantes de sujet offre un espace de travail dans lequel l'analyste peut ajouter ou supprimer les *variantes de sujet* de son choix. Il fournit des interactions avec différents modes de tri, offrant des perspectives multiples pour identifier des termes et des documents porteurs d'information (**T3.2**). Dans la Figure 4, le comparateur dispose les *variantes de sujet* en colonne dans une visualisation matricielle divisée en deux parties. Dans la partie supérieure (4.1), les lignes correspondent aux termes des variantes. Dans la partie inférieure (4.2),elles correspondent aux documents des variantes. La palette de couleurs de la

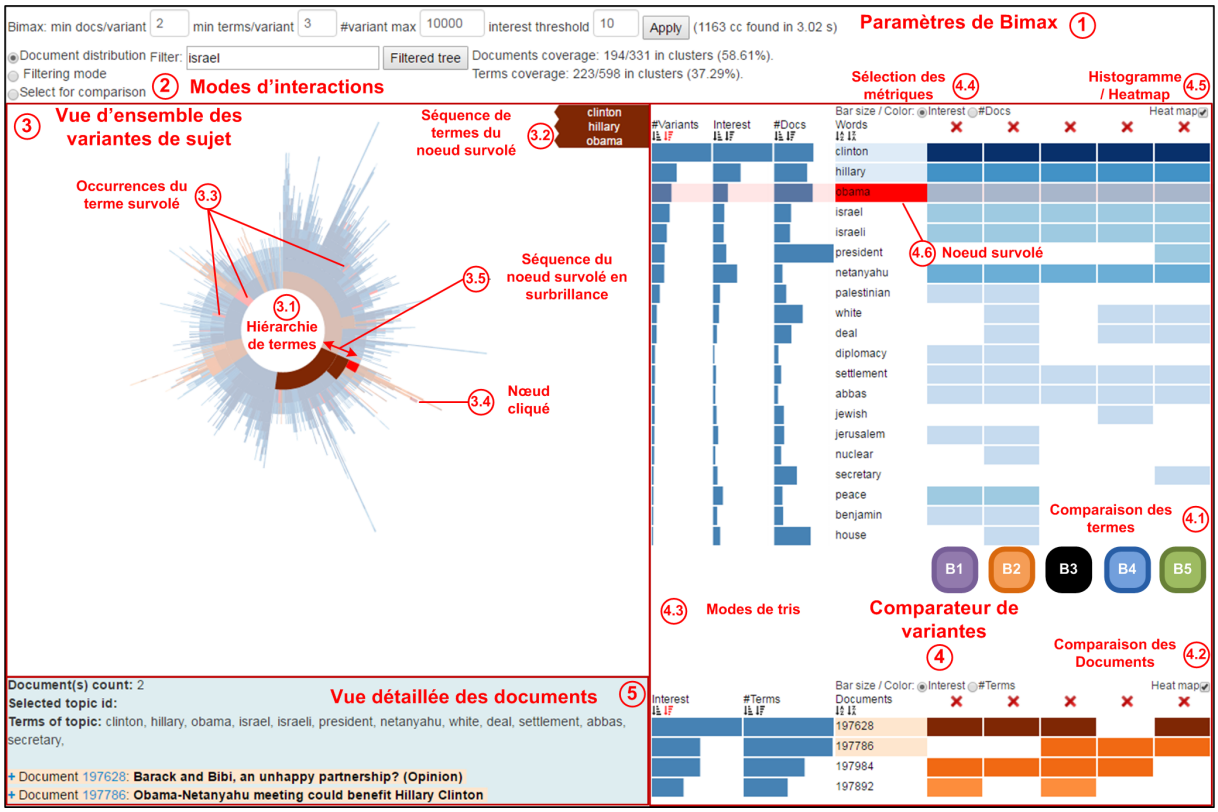

Figure 4. Trois composants permettent d'explorer les variantes du sujet concernant la campagne présidentielle des États-Unis : *la vue d'ensemble* agrégeant 1 163 variantes de sujets (3), *le comparateur de variantes* (4) et *la vue détaillée des documents* (5). Cinq *variantes de sujet* ont été sélectionnées pour être comparées B1-B5. La feuille « secretary » de la variante B5 a été cliquée dans la vue d'ensemble (3.4), sélectionnant deux documents (5). Le terme « obama » est survolé dans le comparateur (4.6) colorant en rouge toutes ses occurrences (3.3). La séquence se terminant par le nœud survolé est mise en surbrillance (3.5).

vue d'ensemble est réutilisée : les informations concernant les termes sont en bleu et celles concernant les documents en orange. Pour chaque terme, trois métriques sont calculées : 1) le degré de chevauchement des biclusters (#Variantes), permettant d'ordonner les termes à l'identique des branches de la hiérarchie, 2) le degré d'intérêt des termes, basé sur les poids *TF-IDF* et 3) le nombre de documents du sujet où le terme apparaît. Ces métriques, représentées à gauche par des histogrammes, permettent différents modes de tri (4.3). De plus, l'analyste peut choisir d'afficher l'une de ces métriques dans les cellules de la matrice (4.4), au choix sous forme d'histogrammes ou de heatmap (4.5). Le comparateur et la vue d'ensemble des variantes sont coordonnés : les termes survolés sont colorés en rouge dans les deux vues (**T3.2, T3.3**) et leur chemin partant de la racine est coloré en bleu dans la colonne *Words* du comparateur ou surligné dans la vue d'ensemble. En cliquant sur une cellule, les documents sélectionnés sont listés dans la vue détaillée des documents (5) (**T2.3**).

La partie inférieure de la matrice (4.2) montre la distribution des documents dans les variantes de sujet. Seulement deux métriques sont proposées : 1) une mesure d'intérêt basée sur les poids *TF-IDF* agrégés par ligne et 2) le nombre de termes du sujet pour chaque document. Dans cette partie de la matrice, les interactions s'appliquent par colonne et concernent la variante complète. Les titres des documents (5) sont colorées en orange dans la liste s'ils apparaissent dans le comparateur. Ces interactions sont conçues pour faciliter la comparaison des variantes de sujet (**T2.2**) et pour identifier les termes et les documents partagés ou spécifiques (**T3.2, T3.3**). On peut aussi accompagner le processus de diversification. Un terme peut être survolé pour identifier de nouvelles variantes (**T3.3**) à ajouter dans l'espace de travail et compléter l'investigation.

ÉVALUATION

Scénario d'utilisation

Nous montrons par un scénario d'utilisation comment l'outil permet d'explorer un grand corpus, en démontrant ses capacités à générer, affiner et vérifier des hypothèses. Les données sont composées de 3 992 articles d'actualité collectés en ligne à partir de sources multiples (BCC, CNN, Reuters, France24, Egypt Independent and Der Spiegel) du 2 au 16 novembre 2015. Nous avons extrait 50 sujets ($K = 50$), visibles dans la carte pondérée des sujets de la Figure 1. Sur les 200 tests effectués pour chaque valeur de $K \in \{10, 20, ..., 100, 150, ..., 500\}$, la partition optimale selon le critère de modularité d'Ailem et al. [1] a été obtenue avec $K = 50$.

Sur la gauche, les sujets de grande taille se rapportent aux événements prépondérants de la période couverte (un avion russe s'est écrasé en Égypte le 31 octobre 2015, la campagne présidentielle américaine, la guerre au Moyen Orient, la crise des migrants, et les attaques terroristes à Paris). Le sujet lié aux élections américaines attire notre

attention. Le nuage de mots contient des termes foncés de grande taille («president», «debate», «candidate», «clinton», «trump») et aussi des termes plus clairs et de grande taille («netanyahu», «israel», «palestinian»). Dans la Figure 4, ce sujet possède 1 163 *variantes de sujets*. Le survol des nœuds au centre de la hiérarchie révèle les termes les plus partagés : « republican, clinton, rubio, israel, trump, debate, candidate ». Les termes sont tous liés à la campagne électorale américaine sauf « israel ». Notre hypothèse à vérifier est donc que les candidats débattent de ce sujet.

D'abord, le mode *Distribution des documents* révèle en orange les variantes qui partagent des documents avec le nœud cliqué. Le nœud central « israel » colore ainsi des variantes avec une séquence commune « clinton, hillary, obama, israel, israeli, netanyahu », faisant le lien entre Israël et les élections américaines. Ensuite, nous filtrons les variantes pour focaliser l'analyse sur celles qui contiennent « israel ». Parmi les variantes restantes, nous ajoutons dans le comparateur celles contenant « clinton ». Nous choisissons d'afficher le degré d'intérêt et de trier les termes par leur nombre de variantes. Les termes partagés par toutes les variantes aident à se focaliser sur le sujet : « clinton, hillary, obama, israel, israeli, netanyahu, settlement et abbas ». La variante B3 regroupe tous ces termes et tous les documents du comparateur. Leurs titres informent que Netanyahu rencontre Obama à la Maison-Blanche. Deux variantes (B1 et B2) contiennent des termes spécifiques « palestinian, diplomacy, peace ». Ces variantes apportent trois documents évoquant la diplomatie difficile entre Obama et Netanyahu. Le terme « nuclear » de la variante B2 se réfère à leur différend à propos du nucléaire iranien. Sur la droite, deux variantes proposent d'autres termes spécifiques : « president, secretary, jewish ». Elles apportent un nouveau document de CNN (id=197786) titré « Obama-Netanyahu could benefit Hillary Clinton ». L'auteur anticipe qu'une rencontre entre Obama et Netanyahu pourrait influencer les votes juifs au bénéfice de la candidate Hilary Clinton, secrétaire d'état aux États-Unis. Ce nouveau document nous amène à affiner notre hypothèse : « une bonne diplomatie entre Obama et Netanyahu bénéficie aux candidats démocrates ». Ce scénario d'utilisation montre la capacité de notre outil à analyser dans le détail un sujet, à générer et affiner des hypothèses, à identifier des relations entre documents révélant des faits et des récits tout en distinguant des angles et points de vues multiples.

Évaluation qualitative avec une experte du domaine

Notre outil de visualisation analytique a été conçu en collaboration avec Warda Mohamed, une journaliste analytique professionnelle et rédactrice à Orient XXI. Elle écrit aussi pour de nombreux médias français dont Le Monde diplomatique et Mediapart. Par la suite, nous la désignons sous le terme « experte ». Nous avons rencontré l'experte trois fois, durant deux à trois heures à chaque fois. Nous avons commencé par un entretien semi-dirigé pour comprendre les besoins des journalistes analytiques et identifier les tâches de haut niveau. Lors du second entretien, nous avons présenté une première version de notre système pour recueillir ses retours et valider/ajuster notre définition des tâches. Lors du troisième entretien, nous avons procédé à une évaluation qualitative que nous avons enregistrée.

Dans une première partie durant 30 minutes, nous avons montré comment utiliser l'outil sur un ensemble de 9 documents ayant servi à l'experte pour publier un article sur les bavures policières survenues récemment en France. L'objectif était de confronter nos découvertes avec les siennes afin de valider l'outil avec une forme de vérité terrain connue de l'experte. Nous l'avons incitée à poser des questions et à commenter les résultats présentés. Pour ce petit corpus, la carte pondérée des sujets n'est d'aucune utilité. Nous nous intéressons donc à la vue d'ensemble des variantes de sujet. Après un rapide coup d'œil, l'experte valide la pertinence des termes de la hiérarchie en disant : « *Je connais le sujet et tous les points importants apparaissent* ». Par exemple, le terme « bras » se rapporte à la façon dont Ali Ziri a été plié durant son arrestation. Le terme « fugitif » a été largement débattu dans les médias. Et le terme « avril » correspond au mois où Amine Bentounsi a été tué en 2012, faisant débat entre les deux tours des élections présidentielles françaises.

Lors de la seconde partie de l'évaluation durant 1h30, l'experte a manipulé les visualisations sur le corpus plus volumineux de notre scénario d'utilisation. Nous dirigions ses manipulations par des questions couvrant toutes les tâches **T1** à **T3**. Nous avons invité l'experte à commenter ce qu'elle comprenait, trouvait intéressant, ainsi que les difficultés rencontrées et les points perfectibles. Nous avons donc commencé par lui demander de trouver, dans la carte pondérée, deux sujets concernant les thèmes du football, de l'astronomie, de la santé/médecine et de l'Asie. L'experte a localisé les trois premières paires de sujets en quelques secondes en expliquant leurs différences. Pour l'Asie, un seul sujet a été trouvé car le thème sur la Chine est associé au thème des réfugiés dans un même sujet (entouré en rouge dans la Figure 1). Globalement, l'experte apprécie la carte pondérée des sujets : « *J'aime bien cet outil, parce que 3 000 documents c'est énorme pour moi* ». « *Même si on voit du bruit dans les sujets, il y a toujours un lien entre les termes. La couleur et la taille des termes dans les sujets a du sens et les liens proposés sont pertinents.* » L'experte trouve cependant étrange que tous les sujets importants apparaissent à gauche. Leur nombre important de termes et de documents induit de fortes similarités qui les rapprochent dans la carte.

L'experte souhaite explorer plus en détail le sujet sur la Chine et les réfugiés. Certaines variantes concernent les pays européens, ainsi que l'Érythrée. L'experte commente : « *L'Érythrée n'est pas un sujet majeur par rapport à la crise des réfugiés en Europe. Cela montre que les variantes de sujet couvrent des aspects variés et très précis. Par exemple, il en ressort aussi bien les pays concernés par la crise que les débats tels que les accords de Schengen ou les demandes d'asile.* » Ensuite, l'experte ajoute être un peu déboussolée par cette visualisation, tout en comprenant son intérêt. Cela nécessite une amélioration de l'étiquetage de la vue sunburst, comme proposé dans la section discussion.

Pour expliquer le lien entre la Chine et la crise des réfugiés en Europe, nous guidons l'experte vers le mode d'interaction de *Distribution des documents*. Alors que les termes « Chine » et « réfugié » colorent en orange des sous-arbres distincts de la hiérarchie, un autre terme, « île », colore en

orange une majorité de variantes. Un examen approfondi des documents révèle que le terme « île » associé à « Chine » se réfère à Taiwan, alors qu'associé à « réfugié » il renvoie aux îles grecques. L'experte commente alors : « *On voit que les termes sont rassemblés sous certains angles très précis, qui font sens ou non. Mais c'est bien que le système montre ces liens, cela suscite la curiosité* ».

Ensuite, nous lui avons demandé d'identifier les variantes pertinentes grâce au comparateur. Nous avons remarqué que l'experte suivait un schéma analytique répété. Après quelques manipulations avec les modes de tri, l'experte examine les documents pour comprendre la sémantique exacte des termes dans leur contexte. L'experte précise : « *Cet outil me fait gagner beaucoup de temps. Même avec des sujets non familiers, si je sais que les premiers termes sont les plus pertinents, je peux me focaliser sur les 10 premiers seulement*. » Nous avons également remarqué que l'experte n'utilisait pas spontanément la distribution des documents dans la partie inférieure du comparateur. Elle nous explique finalement que des redondances apparaissent dans beaucoup de documents et seuls les documents étant au cœur du sujet à son angle doivent être gardés pour construire son fichier « *master* » rassemblant tous les actifs utiles pour son récit. Nous pensons que le comparateur de documents peut être le précurseur de ce fichier « *master* », utilisé par beaucoup de journalistes.

Finalement, l'experte nous suggère certaines améliorations de l'outil. D'abord, une fonction de sauvegarde de l'espace de travail permettrait d'aborder le corpus selon différents angles et revenir sur les résultats précédents. Ensuite, avec une allocation stricte des termes dans les sujets, l'experte peut passer à côté d'aspects importants.

DISCUSSION ET TRAVAUX FUTURS

Le scénario d'utilisation montre que notre outil permet d'explorer la multiplicité des angles et points de vue partagés par des documents, et prouve ainsi la faisabilité des tâches **T1** à **T3**. L'évaluation qualitative reste préliminaire. Elle repose sur une seule experte et une exploration semi-dirigée. Néanmoins, sa participation à la conception garantit une bonne caractérisation du problème et des tâches [32]. Si cette évaluation donne une première appréciation prometteuse de l'utilité de l'outil, elle doit être complétée pour en tirer des conclusions plus générales.

En effet, l'experte est seule juge du caractère inédit des variantes ou de la validation de ses hypothèses. Pour l'évaluer, nous sommes confrontés à la difficulté que les journalistes se limitent à des corpus à taille humaine. À défaut de grands corpus et d'une vérité terrain connue, il nous faut observer l'adoption à long terme de l'outil par plusieurs journalistes sur de nouvelles enquêtes, en évitant le biais possible d'une approche semi-dirigée. Nous envisageons à l'avenir des évaluations quantitatives avec un échantillon suffisamment grand d'étudiants en Web Journalisme. Nous explorerons la faisabilité des tâches à travers nos visualisations en comparant nos algorithmes de biclustering hybride avec hLDA [22], puis la visualisation Sunburst avec d'autres visualisations telles que Word Tree [45].

Pour dresser une vue d'ensemble des variantes (biclusters) contenant beaucoup de redondances de termes, nous sommes face à un choix entre 1) éviter de dupliquer les termes redondants, obligeant à relier explicitement les termes par des liens ou des contours qui génèrent beaucoup d'occlusions (vues nœuds-liens et coordonnées parallèles [34], graphes bipartites [41]) compliquant ainsi l'identification des biclusters ; 2) éviter de dupliquer les termes redondants pour représenter ces relations (vues matricielles [26, 40]), ce qui complique l'identification des termes communs et distinctifs. Notre hiérarchie propose un compromis en évitant les occlusions tout en réduisant les duplications. Les variantes sont identifiables individuellement ainsi que leurs termes communs et distinctifs. Les interactions donnent une appréciation de la distribution des éléments dupliqués, et la matrice du comparateur positionne les variantes verticalement pour éviter les duplications sur une sélection réduite de variantes d'intérêt.

Ensuite, une des limites de notre outil vient du partitionnement strict du biclustering diagonal. En effet, aussi bien les termes que les documents peuvent concerner plusieurs sujets. D'autre part, les approches probabilistes produisant les partitions non-disjointes sont difficiles à appréhender par un public non averti [2]. Nous avons déjà proposé quelques solutions dans la carte pondérée des sujets en affichant les relations de similarité entre les sujets, aidant par ailleurs les journalistes dans leur processus de diversification. À l'avenir, nous envisageons de mieux exploiter les informations de chevauchement captées par les blocs de confusion (Figure 2) pour aider l'utilisateur à façonner ses propres sujets, en proposant des interactions de fusion partielle ou complète entre deux ou plusieurs sujets.

Enfin, l'évaluation qualitative a montré que la visualisation *Sunburst* peut être difficile à utiliser pour des journalistes. Une des raisons provient de l'étiquetage partiel en vigueur en réaction au survol des nœuds. Même si la séquence complète des termes est affichée sur la droite, l'utilisateur focalise son attention sur les positions pointées par la souris. Cela empêche une interprétation exhaustive de la séquence de termes dans son ensemble. Nous travaillons actuellement à définir un étiquetage alternatif adapté à l'analyse de nombreuses séquences de termes.

CONCLUSION

Dans cet article, nous avons décrit un outil de visualisation analytique aidant les journalistes analytiques à explorer de grands corpus. Notre méthode hybride de biclustering permet une analyse multi-resolution. Pour dresser une vue d'ensemble des sujets, nous avons conçu une carte pondérée des sujets. Nous avons proposé une nouvelle approche pour explorer de nombreux biclusters *Termes*×*Documents* non-disjoints sur la base d'une hiérarchie des termes. Une évaluation qualitative montre que cette hiérarchie, coordonnée avec un outil de comparaison, permet à l'analyste d'explorer un nombre important de *variantes de sujet* et de rechercher ainsi des points de vue et angles inédits corroborant les faits et récits contenus dans les documents.

REMERCIEMENTS

Nous remercions Warda Mohamed pour sa contribution à la conception et à l'évaluation du système. Nous remercions également les relecteurs anonymes pour leur contribution à l'amélioration de cet article.

BIBLIOGRAPHIE

1. M. Ailem, F. Role, and M. Nadif. 2015. Co-clustering Document-term Matrices by Direct Maximization of Graph Modularity. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM '15)*. ACM, 1807–1810.
2. E. Alexander and M. Gleicher. 2015. Task-Driven Comparison of Topic Models. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (Aug. 2015), 320–329.
3. E. Alexander, J. Kohlmann, R. Valenza, M. Witmore, and M. Gleicher. 2014. Serendip : Topic model-driven visual exploration of text corpora. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 173–182.
4. S. Barkow, S. Bleuler, A. Prelić, P. Zimmermann, and E. Zitzler. 2006. BicAT : a biclustering analysis toolbox. *Bioinformatics* 22, 10 (May 2006), 1282–1283.
5. D.M. Blei, A.Y. Ng, and M.I. Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3 (Jan. 2003), 993–1022.
6. M. Bostock. 2016. D3-Sunburst. (2016). http://bl.ocks.org/mbostock/4063423(accédé en août 2016).
7. M. Brehmer, S. Ingram, J. Stray, and T. Munzner. 2014. Overview : The Design, Adoption, and Analysis of a Visual Document Mining Tool for Investigative Journalists. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (Dec. 2014), 2271–2280.
8. J. Choo, C. Lee, C.K. Reddy, and H. Park. 2013. UTOPIAN : User-Driven Topic Modeling Based on Interactive Nonnegative Matrix Factorization. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (Dec. 2013), 1992–2001.
9. C. Collins, F.B. Viegas, and M. Wattenberg. 2009. Parallel Tag Clouds to explore and analyze faceted text corpora. In *2009 IEEE Symposium on Visual Analytics Science and Technology (VAST)*. 91–98.
10. W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong. 2011. TextFlow : Towards Better Understanding of Evolving Topics in Text. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (Dec. 2011), 2412–2421.
11. J. Davies. 2016. D3-cloud. (2016). https://github.com/jasondavies/d3-cloud(accédé en août 2016).
12. I.S. Dhillon. 2001. Co-clustering Documents and Words Using Bipartite Spectral Graph Partitioning. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '01)*. ACM, 269–274.
13. A. Don, E. Zheleva, M. Gregory, S. Tarkan, L. Auvil, T. Clement, B. Shneiderman, and C. Plaisant. 2007. Discovering Interesting Usage Patterns in Text Collections : Integrating Text Mining with Visualization. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management (CIKM '07)*. ACM, 213–222.
14. W. Dou, X. Wang, R. Chang, and W. Ribarsky. 2011. ParallelTopics : A probabilistic approach to exploring document collections. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 231–240.
15. W. Dou, X. Wang, D. Skau, W. Ribarsky, and M.X. Zhou. 2012. LeadLine : Interactive visual analysis of text data through event identification and exploration. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 93–102.
16. W. Dou, L. Yu, X. Wang, Z. Ma, and W. Ribarsky. 2013. HierarchicalTopics : Visually Exploring Large Text Collections Using Topic Hierarchies. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (Dec. 2013), 2002–2011.
17. P. Fiaux, M. Sun, L. Bradel, C. North, N. Ramakrishnan, and A. Endert. 2013. Bixplorer : Visual Analytics with Biclusters. *Computer* 46, 8 (Aug. 2013), 90–94.
18. P. Gambette and J. Véronis. 2010. Visualising a Text with a Tree Cloud. In *Classification as a Tool for Research*. Springer Berlin Heidelberg, 561–569.
19. M. Ghoniem, M. Cornil, B. Broeksema, M. Stefas, and B. Otjacques. 2015. Weighted maps : treemap visualization of geolocated quantitative data. In *Proc. SPIE, Visualization and Data Analysis 2015*. 93970G–93970G–15.
20. M. Ghoniem, J.D. Fekete, and P. Castagliola. 2005. On the Readability of Graphs Using Node-Link and Matrix-Based Representations : A Controlled Experiment and Statistical Analysis. *Information Visualization* 4, 2 (June 2005), 114–135.
21. G. Govaert and M. Nadif. 2013. *Co-Clustering : Models, Algorithms and Applications*. Wiley.
22. T.L. Griffiths, M.I. Jordan, J.B. Tenenbaum, and D.M. Blei. 2004. Hierarchical Topic Models and the Nested Chinese Restaurant Process. In *Advances in Neural Information Processing Systems 16*. MIT Press, 17–24.
23. J. Han, J. Pei, Y. Yin, and R. Mao. 2004. Mining Frequent Patterns without Candidate Generation : A Frequent-Pattern Tree Approach. *Data Mining and Knowledge Discovery* 8, 1 (Jan. 2004), 53–87.
24. B. Hanczar and M. Nadif. 2011. Using the bagging approach for biclustering of gene expression data. *Neurocomputing* 74, 10 (May 2011), 1595–1605.
25. S. Havre. 2002. ThemeRiver : visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics* 8, 1 (Jan. 2002), 9–20.
26. J. Heinrich, R. Seifert, M. Burch, and D. Weiskopf. 2011. BiCluster Viewer : A Visualization Tool for Analyzing Gene Expression Data. In *ISVC'11 : Proceedings of the 7th international conference on Advances in visual computing*. Springer Berlin Heidelberg, 641–652.

27. B. Lee, N.H. Riche, A.K. Karlson, and S. Carpendale. 2010. SparkClouds : Visualizing Trends in Tag Clouds. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (Nov. 2010), 1182–1189.
28. D.D. Lee and H.S. Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 6755 (Oct. 1999), 788–791.
29. M. Lee Hunter, N. Hanson, S. Rana, L. Sengers, D. Sullivan, and P. Thordsen. 2009. L'enquête par hypothèse : manuel du journaliste d'investigation. (2009). http://markleehunter.free.fr/documents/SBI_french.pdf(accédé en août 2016).
30. S. Liu, X. Wang, J. Chen, J. Zhu, and B. Guo. 2014. TopicPanorama : A full picture of relevant topics. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 183–192.
31. S.C. Madeira and A.L. Oliveira. 2004. Biclustering Algorithms for Biological Data Analysis : A Survey. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 1, 1 (Jan. 2004), 24–45.
32. T. Munzner. 2009. A Nested Model for Visualization Design and Validation. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (Nov. 2009), 921–928.
33. A. Prelić, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler. 2006. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics* 22, 9 (May 2006), 1122–1129.
34. R. Santamaría, R. Therón, and L. Quintales. 2008. A visual analytics approach for understanding biclustering results from microarray data. *BMC Bioinformatics* 9, 1 (May 2008), 247.
35. M.M. Shafiei and E.E. Milios. 2006. Latent Dirichlet Co-Clustering. In *Sixth International Conference on Data Mining, 2006. ICDM '06*. 542–551.
36. B. Shneiderman. 1996. The eyes have it : a task by data type taxonomy for information visualizations. In *IEEE Symposium on Visual Languages, 1996. Proceedings*. 336–343.
37. K. Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28, 1 (1972), 11–21.
38. J. Stasko, R. Catrambone, M. Guzdial, and K. McDonald. 2000. An evaluation of space-filling information visualizations for depicting hierarchical structures. *International Journal of Human-Computer Studies* 53, 5 (Nov. 2000), 663–694.
39. J. Stasko, C. Görg, and Z. Liu. 2008. Jigsaw : Supporting Investigative Analysis through Interactive Visualization. *Information Visualization* 7, 2 (June 2008), 118–132.
40. M. Streit, S. Gratzl, M. Gillhofer, A. Mayr, A. Mitterecker, and S. Hochreiter. 2014. Furby : fuzzy force-directed bicluster visualization. *BMC Bioinformatics* 15, Suppl 6 (May 2014), S4.
41. M. Sun, P. Mi, C. North, and N. Ramakrishnan. 2015. BiSet : Semantic Edge Bundling with Biclusters for Sensemaking. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (Aug. 2015), 310–319.
42. M. Sun, C. North, and N. Ramakrishnan. 2014. A Five-Level Design Framework for Bicluster Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (Dec. 2014), 1713–1722.
43. Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei. 2006. Hierarchical Dirichlet Processes. *J. Amer. Statist. Assoc.* 101, 476 (Dec. 2006), 1566–1581.
44. F. B. Viegas, M. Wattenberg, and J. Feinberg. 2009. Participatory Visualization with Wordle. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (Nov. 2009), 1137–1144.
45. M. Wattenberg and F.B. Viegas. 2008. The Word Tree, an Interactive Visual Concordance. *IEEE Transactions on Visualization and Computer Graphics* 14, 6 (Nov. 2008), 1221–1228.
46. F. Wei, S. Liu, Y. Song, S. Pan, M.X. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang. 2010. TIARA : a visual exploratory text analytic system. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '10)*. ACM, 153–162.
47. J.A. Wise, J.J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. 1995. Visualizing the non-visual : spatial analysis and interaction with information from text documents. In *Information Visualization, 1995. Proceedings.* 51–58.