HAL open science

Mesure de cohérence dans la découverte et la visualisation d'évènements médiatiques

Benjamin Renoust, Marie-Luce Viaud, Guy Melançon

To cite this version:

Benjamin Renoust, Marie-Luce Viaud, Guy Melançon. Mesure de cohérence dans la découverte et la visualisation d'évènements médiatiques. Seconde conference sur les Modèles et l'Analyse des Reseaux: Approches Mathematiques et Informatique (Marami), Oct 2011, France. pp.4. hal-00927215

HAL Id: hal-00927215

https://hal.science/hal-00927215v1

Submitted on 27 Jan 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Mesure de cohérence dans la découverte et la visualisation d'évènements médiatiques

Benjamin Renoust*, — Marie Luce Viaud* — Guy Mélançon****

** Institut National de l'Audiovisuel*
brenoust@ina.fr, mlviaud@ina.fr
*** CNRS UMR 5800 LaBRI & INRIA Bordeaux Sud-Ouest*
guy.melancon@labri.fr

RÉSUMÉ. *Nous proposons une méthode pour visualiser et analyser les évènements médiatiques à partir des sujets d'actualité des journaux télévisés de plusieurs chaînes annotées avec des descripteurs textuels. Nous présentons une interface d'exploration basée sur un modèle de graphe de similarité sémantique. Après une étape classique couplant clustering et dessin de graphe, nous avons élaboré une mesure de cohérence inspirée par les travaux de Burt et Schott [5] et offrant un retour visuel qualitatif des agrégats générés. Cette mesure de cohérence permet à l'utilisateur de contrôler et valider différents processus de filtrage et raffinage du clustering initial. La cartographie résultante met en évidence deux types d'agrégats, les agrégats thématiques et évènementiels.*
MOTS-CLÉS : *visualisation, interaction, détection d'évènements, cohérence, sémantique, clustering de documents, fouille.*

1. Introduction

L'Ina conserve exploite et met à disposition de publics divers plus de 2 million d'heures de télévision et de radio, et capte tous les jours plus de 100 chaînes télévisuelles et radiophoniques ainsi que la partie du web s'y rapportant. En relation avec ces corpus, l'Ina a lancé en 2011 un projet ANR d'Observatoire TransMedia (OTMedia, www.otmedia.fr) dont l'objectif est de mettre en place des processus, outils et méthodes pour mieux appréhender les enjeux et les mutations de la sphère médiatique. OTMedia a pour objet l'étude et la propagation des évènements médiatiques sur tous les supports de diffusion : web (blogs, site de media, agrégateurs), presse, radio, télévision et twitter. L'originalité de l'Observatoire TransMedia est de partir des besoins d'analyse exprimés par les chercheurs en SHS et les acteurs de l'information pour élaborer de nouveaux concepts et outils d'analyse adaptés aux volumes et à la diversité du paysage informationnel. Outre l'intégration de leur problématique pour la fouille, ce parti pris implique une *approche du système centrée sur l'utilisateur*, mettant un fort accent sur la *fouille visuelle interactive* : il s'agit de manipuler les points de vue, les ensembles considérés au travers, par exemple, des période de temps, des supports ou des

acteurs éditoriaux, et les paramètres d'analyse pour découvrir, analyser, naviguer mais aussi valider les processus automatiques d'analyse mis en œuvre.

Nos travaux portent sur la création d'une *cartographie interactive des évènements médiatiques* à partir des sujets d'actualité de journaux télévisés et l'élaboration d'une *mesure de cohérence sémantique*. Cette mesure est couplée à un retour visuel, elle facilite la perception, permet la validation manuelle des agrégats générés, et éventuellement le contrôle de processus de filtrage et de raffinement additionnels. De plus notre démarche nous a permis d'identifier des agrégats thématiques que nous présentons aux utilisateurs au même titre que les évènements.

2. Objectifs et Données

2.1. Objectifs

La première étape dans l'analyse de la propagation de l'information est la détection des évènements médiatiques. Selon les chercheurs en SHS, un évènement se construit à partir d'un fait se produisant dans la vie réelle privée ou publique. De l'énorme masse des évènements, le journaliste ne retient que ceux qui sont « notables »[8] et qui peuvent donc intéresser le public. Le fait devient une nouvelle lorsqu'il entre dans le circuit journalistique [11]. Un évènement médiatique est un fait rendu populaire grâce aux media. De notre point de vue, L'évènement médiatique se caractérise donc par une densité forte de « nouvelles » issues d'un même fait.

Notre objectif est de construire des cartes d'information permettant à l'utilisateur :

- De découvrir les évènements médiatiques et leur importance relative.
- De proposer des retours visuels pour permettre d'évaluer les agrégats et éventuellement d'en contrôler l'évolution.
- De proposer des marqueurs visuels ou dynamiques permettant de visualiser la proximité et la temporalité des sujets médiatiques entre eux.

2.2. Approche documentaire et données

Le travail documentaire constitue une part importante du travail d'archivage et l'Ina compte environ 180 documentalistes, qui structurent décrivent et exploitent les archives télévisuelles et radiophoniques. Les éditions principales des journaux télévisés des chaînes de plus forte audience sont décrites pour chaque sujet d'information. La notice documentaire comprend, entre autres, de 1 à 15 descripteurs sémantiques (lieux, personnes physiques, personnes morales, nom commun) issus du thésaurus de l'Ina.

Les données utilisées pour cette étude sont donc issues des journaux d'informations télévisés (JT) de TF1, France 2, France 3 (National et Outre-Mer), Arte, et M6. Chaque chaîne propose environ 10 sujets d'information par journal. La période observée s'étend de juin 2009 à octobre 2009 soit environ 8000 sujets. Ces données présentent des caractéristiques particulières :

- La proximité sémantique basée sur les descripteurs sémantiques : les descripteurs servent à structurer l'information, à la fois pour la catégoriser et la

rechercher. Aussi, les descripteurs sémantiques ont-ils des niveaux de précision diverses, partiellement reliés à leur place dans la hiérarchie du thésaurus. La documentation doit idéalement *singulariser* et *normaliser* la description de chaque document pour minimiser le bruit et le silence documentaire. L'indexation manuelle joue sur l'*enchevêtrement* des descripteurs, étant donnés leurs différents niveaux de généricité/précision, et pour satisfaire à la fois une recherche précise directe et une navigation par essence plus thématique.

- La répartition temporelle d'évènement suit souvent une distribution de type hypergéométrique [9] aux caractéristiques propres à chaque évènement. Cette forme s'explique en partie de la structuration (encore actuelle) des médias de masse traditionnels, dans laquelle les acteurs éditoriaux diffusent les évènements publiés par les agences de presse en amont et "se suivent" les uns et les autres, et la modélisation des "attentes" du public (attraction pour la "nouveauté" des sujets). Nous noterons que nous n'avons pas de connaissances a priori sur la durée et l'intensité propres à chaque évènement.

Notre objectif est de trouver et de représenter des agrégats principaux de sujets/documents *cohérents* en termes d'évènements ou thématique afin de fournir une observation globale et donner des clefs pour la perception de la structure de l'actualité informationnelle. Ainsi quelques hypothèses fortes vont nous aider à orienter notre démarche:

➔ Deux sujets médiatiques parlant d'un même évènement sont sémantiquement proches.

➔ Un agrégat de sujets médiatiques sémantiquement proches implique probablement l'existence d'un évènement médiatique.

➔ Un évènement médiatique suit une loi de continuité dans le temps.

Notre approche consiste donc dans un premier temps à regrouper les documents suivant leurs caractéristiques sémantique puis à exploiter leur caractéristique temporelle pour une identification plus fine des évènements.

2.3. Positionnement

J. Allan présente dans [1] plusieurs tâches de recherche d'information au travers du Topic Detection and Tracking (TDT) en relation avec nos objectifs. Le but du TDT est de découper le flux d'information (qui peut être multimodal) en un ensemble d'évènements individuels, de détecter les nouveaux évènements émergents, et de répartir le tout dans des classes recouvrant chacune un unique évènement [2]. Nallapati et al. [13] parlent ainsi de fils d'évènements et mettent l'accent sur les indices de temporalité qu'ils détectent à travers un apprentissage supervisé. Fung et al. [9] présente également une approche intéressante tout en se distinguant du TDT : les auteurs s'attaquent au problème de la détection d'évènements émergents sans tenter de les classifier. Patton et al. [17] s'attachent à identifier les évènements intéressants correspondant à une requête donnée au sein d'une masse de données très importante et dynamique. Plus récemment Zhao et al. [21] utilisent l'appui du graphe social comme celui des weblogs pour renforcer le

clustering classique et en fournit une intéressante visualisation. Cointet et al. [6] mettent en avant les interactions entre les acteurs de l'information en les groupant sémantiquement. Nallapati et al. [14] apporte une méthode non supervisée tenant compte de la structure sociale des blogs pour en ressortir les groupes d'acteurs et de sujets influents sans prendre en compte la temporalité. Pouliquen et al. [18] fournit bien une visualisation des sujets découverts. Aucune de ces approches ne fournit de mesure de cohérence sémantique. Peu présentent de visualisation interactive des résultats, mais surtout ne révèlent pas la structure de description mettant en commun leurs agrégats [6,9,13,14,17]. Les travaux de J.C. Lamirel [12] présentent l'intérêt de l'analyse multivue en recherche d'information par la classification et mettent en avant l'intérêt de l'utilisation de visualisations spécifiques. Herman et al. [10], ainsi que plus récemment Landesberger et al. [20] ont dressé deux états de l'art sur la visualisation et la navigation des graphes et grands graphes en fonction du type d'information que l'on doit analyser et du type d'interactions que l'on veut proposer. Nos travaux amènent une analyse non supervisée de la sémantique et de la temporalité de documents afin d'en faire ressortir les évènements médiatiques et leur structure descriptive et temporelle. Une visualisation multivues à base de graphe dans lesquelles l'utilisateur joue un rôle central dans l'identification des évènements est au cœur de nos objectifs. Nous souhaitons pour cela définir une mesure de cohérence des évènements détectés comme indicateur visuel.

3. Framework

3.1. Modèle de graphe

Nous avons choisi de modéliser ce corpus d'information par un graphe de similarités multicouches où chaque nœud représente un document (ou sujet). Les documents sont décrits selon le modèle d'espace vectoriel sémantique (Vector Space Model [19]) par des vecteurs réels positifs dont les composantes correspondent à la valeur de TF-IDF [19] de chaque descripteur. La proximité de deux vecteurs est calculée avec la similarité-cosinus [19]. Pour chaque document, nous sélectionnons ainsi ses K plus proches voisins (KNN) avec $K=30$ (empiriquement déterminé afin de maximiser le coefficient de clustering, 0.454 contre 0.005 sur un graphe aléatoire simple d'autant de nœuds et arêtes), et nous générons une *arête de voisinage* entre *documents voisins* (au sens KNN). Note : le choix du KNN nous permet de bien découper le graphe en agrégats denses, un choix de K trop élevé génère un graphe très dense si proche de la clique qu'il est presque impossible d'en identifier les agrégats topologiquement. Une arête de voisinage contient autant de couches de description que de descripteurs partagés par ses documents qu'elle relie. Le graphe ne contient pas d'arêtes réflexives (un descripteur n'ayant qu'une occurrence dans le cluster ne participe pas à l'agrégation). La répartition des degrés de ce graphe suit une loi de puissance (à partir du degré 30, dû au choix du KNN) et on y observe un comportement de petit monde qui rend pertinent la recherche d'agrégats.

3.2. Dessin de graphe et algorithme d'agrégation

Notre objectif étant de fournir une carte visuelle permettant à l'utilisateur de naviguer dans l'information, nous avons couplé les étapes de clustering et de dessin du graphe. Selon Noack [15] il existe un lien très fort entre le dessin de graphe à modèle de forces et le clustering par modularité. Nous avons choisi l'algorithme *edge-linlog* de Noack [16] pour lequel la répulsion s'exprime par une force logarithmique et l'attraction par une force linéaire. Cet algorithme regroupe ainsi fortement les densités de noeuds importantes et éclate ces zones denses dans l'espace. Il s'applique bien à nos données qui présentent un coefficient de clustering élevé. La détection des clusters s'effectue en filtrant les arêtes du graphe pour ne garder que les arêtes de petites tailles correspondant en conséquence aux arêtes intra-cluster. Nous avons choisi un seuil à partir de l'histogramme des tailles d'arêtes sur nos données : nous éliminons les arêtes de taille supérieure à 1/100 de la taille de l'arête la plus grande, soit une sélection d'environ la moitié des arêtes. Ce choix empirique est possible grâce aux caractéristiques de l'algorithme de Noack, qui présente une répulsion logarithmique et une attraction linéaire. Note :il serait intéressant de déterminer la limite du choix des longueurs d'arêtes en fonction des paramètres de l'algorithme. Nous détectons les composantes connexes de ce graphe filtré puis nous rajoutons les arrêtes intra-cluster éliminées précédemment afin de ne pas modifier les liens de proximité des documents au sein d'un même cluster. Parmi les composantes connexes, seules celles composées d'au moins une dizaine de nœuds représentent un intérêt d'étude pour nos utilisateurs, aussi ne conservons-nous que les clusters de 10 noeuds ou plus.

3.3. Tulip

Nous utilisons le framework Tulip [3] développé au LaBRI et dans l'équipe INRIA GRAVITE et proposant tous les outils communs de manipulation de graphe. En plus d'offrir une interface déjà riche en fonctionnalités, Tulip dispose d'une communauté active qui développe des plugins intégrant des algorithmes des plus utiles, allant du calcul de mesure, aux derniers dessins de graphes. Tulip permet en outre de synchroniser plusieurs vues du graphe et dispose d'interacteurs sur ces vues permettant la manipulation directe du graphe.

3.4 Observation sur les agrégats

Nous pouvons déjà observer deux cas extrêmes dans la formation des agrégats. Certains agrégats sont liés à un évènement particulier comme l'enterrement de Michael Jackson, lorsque d'autres semblent traiter de la même thématique sans montrer précisément un évènement particulier, ou bien en regroupant plusieurs sous-évènements comme les élections présidentielles en Afrique du Sud, en Algérie, et en Iran. Dans la plupart des cas, il s'agit surtout d'une composition mixte entre ces deux extrêmes. Ceci s'explique par les mesures de similarités entre sujets d'information basées sur les descripteurs qui peuvent être interprétées en terme de proximité de sujet (un même évènement traité par plusieurs chaînes) mais aussi en terme de thématique (ex : élections). On peut noter que les définitions

d'*évènement* et de *thématique* ne sont pas arrêtées et que leur spécification fait l'objet d'une étude commune avec les SHS dans le cadre du projet OTMedia. Un traitement spécifique supplémentaire est alors nécessaire pour une identification plus fine des évènements.

3.5. Graphe de description et indicateurs de cohérence sémantique

Considérons le graphe dual de notre graphe de documents: les noeuds représentent les descripteurs, et les arêtes sont valuées par le nombre de paires de *documents voisins* qui partagent deux descripteurs. Nous considérerons par la suite le graphe dual non réflexif que nous nommerons *graphe de description*. Nous générons un graphe de description pour chaque cluster du graphe de documents précédemment détecté. Les graphes de description présentent des caractéristiques assez diverses en taille et en topologie (de 6/20 à 119/572 nœuds/arêtes). Leur interprétation est riche:

- Si le graphe de description est petit en taille, dense en arêtes et avec des valeurs d'arêtes élevées, alors selon nos hypothèses le cluster présente une forte cohérence sémantique. Le cas extrême est une clique avec des valeurs d'arêtes correspondant au nombre de documents du cluster dans le graphe des documents : c'est le cas d'un fait d'actualité rapporté dans plusieurs sujets et décrit par le même ensemble de descripteurs. L'*enchevêtrement* des descripteurs entre eux est maximal puisqu'ils sont tous utilisés ensembles.

- Si un noeud du graphe de description est déconnecté, alors toutes les arêtes de voisinage portant ce descripteur dans le graphe des documents sont des arêtes monocouche : le descripteur ne présente aucun enchevêtrement avec d'autres descripteurs. De même, si le graphe dual présente plusieurs composantes connexes, alors deux descripteurs issus de deux composantes connexes distinctes ne sont jamais présents ensemble sur une arête de voisinage multicouche du graphe des documents. Les descripteurs d'une composante connexe ne présentent alors aucun enchevêtrement avec ceux des autres composantes connexes.

L'analyse de la nature de la documentation (singularité/généricité des descripteurs) et l'étude des graphes de description des clusters relatifs à nos données montrent que l'occurrence d'un descripteur et l'enchevêtrement des descripteurs entre eux au sein d'un cluster sont fortement liés à la notion cohérence sémantique des agrégats de documents observés. Nous proposons l'élaboration d'une mesure de cohérence sémantique des agrégats basée sur ces deux notions. Nous remarquons que les clusters « évènements » possèdent une valeur de cohérence plus élevée que les clusters « thématiques », ce que nous devrons vérifier avec notre mesure.

4. Mesure de cohérence d'agrégats

Nous nous sommes intéressés à la mesure d'ambiguïté proposée par Burt et Schott [5] dans le domaine des SHS. Ils cherchaient à caractériser la nature des relations entre les individus en entreprise. Pour cela, ils ont identifié plusieurs types de relations interpersonnelles, comme par exemple les relations professionnelles ou

sportives et se sont intéressés à leurs intrications. Plus les types de relations s'entremêlent, plus il devient difficile de caractériser la nature de la relation principale au sein du réseau social. Ces *intrications* attesteraient donc d'une certaine « ambiguïté » des échanges sociaux sous-jacents.

Dans notre cas d'étude, nous recherchons la cohérence d'un cluster. Celle-ci se traduit par l'enchevêtrement des descripteurs sur les arêtes du cluster. Plus les descripteurs s'entremêlent fortement, plus le cluster nous semble cohérent. On fait ainsi le parallèle avec les travaux de Burt où les types de relations correspondent à nos descripteurs et où la mesure d'ambiguïté correspond à notre mesure de cohérence.

4.1. Méthode de calcul

On note $s,t,\in T$ les descripteurs entre sommets du graphe $G=(V,E)$. On a pour chaque arête $e\in E$ un ensemble de descripteurs $\tau(e)\subset T$ portés par cette arête. On peut voir $\tau : E\rightarrow P(T)$ (l'ensemble des parties de *T*) comme une application associant à chaque arête $e\in E$ ses descripteurs. Par conséquent, $\tau^{-1}(t)$ nous donne l'ensemble des arêtes qui portent le descripteur $t\in T$. On note $n_t=|\tau^{-1}(t)|$ le nombre d'arêtes portant $t\in T$.

On calcule la matrice d'interaction des descripteurs $(c_{s,t})_{s,t\in T}$ où $c_{s,t}$ est défini comme suit. On calcule d'abord $n_{s,t}$ qui donne le nombre d'arêtes $e\in E$ telles que $(s,t)\subset\tau(e)$. On pose ensuite la probabilité qu'une arête porte le descripteur *t :* $c_{t,t}=\frac{n_{t,t}}{|E|}$, puis nous posons la probabilité conditionnelle qu'une arête porte le descripteur *s* sachant qu'elle porte le descripteur *t :* $c_{s,t}=\frac{n_{s,t}}{n_t}$. Notons que si la matrice $(n_{s,t})_{s,t\in T}$ est symétrique, ce n'est plus le cas pour la matrice $(c_{s,t})_{s,t\in T}$ puisque ses coefficients sont des probabilités conditionnelles.

La matrice $(c_{s,t})_{s,t\in T}$ ayant tous ses coefficient positifs ou nuls, nous pouvons affirmer selon le théorème de Perron-Frobenius [7] qu'elle admet une valeur propre maximale positive, avec un vecteur propre associé dont les coefficients sont aussi positifs.

Un raisonnement probabiliste mené par Burt et Schott [5] s'appuyant sur les indices d'ambiguïté des types montre que ces indices sont les coordonnées du vecteur propre associé à la valeur propre maximale λ de la matrice $(c_{s,t})_{s,t\in T}$. On peut aussi pour un cluster de documents donné, étudier la valeur propre λ en soit et la définir comme mesure de cohérence associée au cluster. Selon Perron Frobenius [7], λ satisfait la condition suivante : $\lambda\leq \max_t \sum^{s} c_{s,t}$. Dans le cas extrême où tous

les descripteurs recouvrent toutes les arêtes du graphe, la somme est maximisée à *N*. Nous pouvons ainsi considérer la mesure de cohérence normalisée : λ / N

Dans le cas particulier où le graphe de description du cluster observé est non connexe, la matrice $\left(c_{s,t}\right)_{s,t \in T}$ est diagonale par blocs. Nous déterminons ainsi une valeur de λ pour chaque sous matrice correspondant à une composante connexe du graphe de description.

4.2. Caractérisation de λ

Nous avons étudié la croissance de λ sur des graphes artificiels sur lesquels nous avons augmenté l'enchevêtrement de trois façons différentes, par ajout de liens d'un descripteur s'intriquant avec un autre, par permutation des liens d'un descripteur (sans en changer le nombre) mais en augmentant le nombre d'intrications avec un autre, par ajout successif de descripteurs se recouvrant totalement. Nous sommes partis d'un graphe à 5 sommets, où un descripteur recouvre toutes les arêtes. Comme les Figures 1 et 2 ci-contre le montrent, notre mesure de cohérence croît en fonction de ces 3 paramètres qui maximisent l'intrication entre les descripteurs.

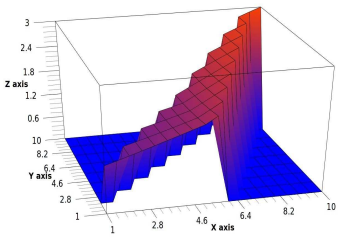

Figure 1 :Evolution de la valeur de cohérence (sur l'axe *Z*) en fonction du nombre d'arêtes total des descripteurs *B* et *C* (sur l'axe *Y*) et du nombre d'intrications entre *B* et *C* par permutation des arêtes (sur l'axe *X*).

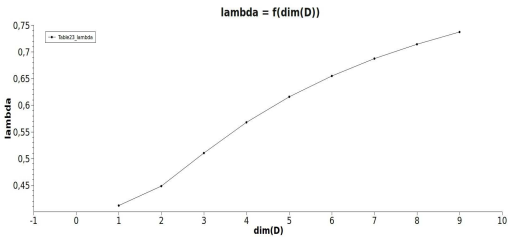

Figure 2 : Evolution de la valeur de cohérence en fonction du nombre de duplication du descripteur *D*.

4.3. Etiquetage

L'espace propre associé à notre valeur propre λ étant unidimensionnel avec son vecteur propre associé à coefficients positifs, nous proposons d'utiliser ses deux composantes principales (les deux coefficients à plus grande valeur) pour identifier les descripteurs « les plus enchevêtrés ». Nous étiquetons ainsi nos clusters à partir de deux paires de descripteurs, correspondant aux plus occurrents et aux plus enchevêtrés.

5. Visualisation

5.1. Visualisation de la mesure

Afin de rendre à l'utilisateur un retour visuel de la mesure de cohérence (Figure 3), nous colorons les clusters du graphe de documents en fonction de leur valeur d'enchevêtrement. Pour cela nous utilisons deux échelles de Brewer reconnues pour

leur lisibilité [4] qui présentent une couleur similaire pour leur minimum. Une échelle du jaune (min) au bleu (max) correspond aux clusters ayant un dual non-connexe (dont les λ ont été pondérés puis moyennés). Une échelle du jaune (min) au rouge (max) correspond aux clusters normaux.

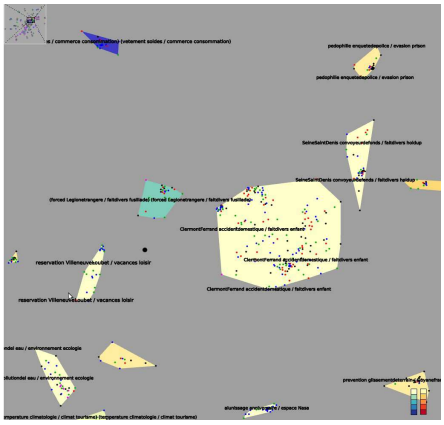

Figure 3 : Un exemple de clustering avec échelles de Brewer et coloration des nœuds par chaîne de diffusion (TF1 en bleu, France 2 en vert, France 3 en noir, Arte en rose, M6 en rouge).

5.2. Amélioration de la cohérence

Afin d'augmenter la cohérence de chaque cluster, nous offrons plusieurs solutions de filtrage et de raffinage liés à la sémantique et à la temporalité de notre information.

La première méthode consiste à filtrer une partie du bruit à travers le graphe de description en identifiant des composantes connexes de petite taille. Ce sont des artefacts liés à la méthode de création du graphe et associant des documents « éloignés de tous » à des clusters plus denses. Ces petites composantes sont souvent associées à très peu de nœuds dans le graphe des documents et leur thématique liée mais plus éloignée de celle du reste des documents. On peut ainsi supprimer ces arêtes de voisinage ainsi que supprimer les documents qui se retrouvent déconnectés. Ce filtrage permet d'augmenter légèrement la mesure de cohérence des clusters.

Le filtrage peut être aussi être manuel car l'utilisateur peut vouloir « affiner » lui même la carte avec ses connaissances. Il peut observer de manière synchronisée les documents liés aux descripteurs de son choix et ainsi supprimer les descripteurs qui lui semblent peu cohérents, les documents isolés par l'opération seront alors supprimés

5.3. Raffinage temporel

Notre observation du contenu des clusters en Section 3.4 nous amène à continuer le processus de recherche des évènements en se basant sur la caractéristique temporelle. L'hypothèse 3 nous permet de modéliser des liens temporels entre documents. Le raffinage temporel consiste ainsi à coupler la topologie et la temporalité du graphe des documents : on enrichit les liens avec la différence temporelle entre deux *documents voisins*, puis on filtre les arêtes du graphe suivant la proximité temporelle requise qui dans notre cas est fixée à 1

journée. En effet, dans les médias traditionnels, l'unité de production est la journée. Le graphe des documents se retrouve découpé en plusieurs composantes connexes correspondant aux sujets proches émis dans un intervalle d'une journée. Nous reconstruisons ensuite les liens internes à chaque composante connexe pour ne pas modifier les relations sémantiques des sujets entre eux au sein d'un même « cluster temporel ». Ce découpage permet de générer des sous clusters dans un agrégat qui contient plusieurs évènements. Ces sous-agrégats présentent chacun une très forte croissance de leur mesure de cohérence en comparaison de celle du cluster original. Les documents isolés sont conservés au sein du cluster pour permettre l'exploration.

6. Implémentation, Résultats et Discussion

6.1 Implémentation

En utilisant le framework Tulip nous fournissons à l'utilisateur une interface lui proposant une navigation dans le graphe des documents grâce aux clusters labélisés. Par cluster, l'utilisateur dispose d'un accès aux graphes de description, et aux histogrammes temporels.

Figure 4 : Démonstration de la synchronisation entre un cluster de documents (élections européennes) et son graphe de descripteurs. En rose, la sélection synchronisée (Berlin et Angela Merkel ont été sélectionnés).

Le placement des clusters identifiés est généré par le dessin de graphe (comme expliqué dans la Section 3.2) et permet à l'utilisateur de se dresser une carte mentale du panorama de l'information. Chaque cluster apparaît sur la carte comme un ensemble coloré en fonction de sa mesure de cohérence (Section 5.1). Leurs noeuds sont étiquetés avec le titre du cluster (Section 4.3).

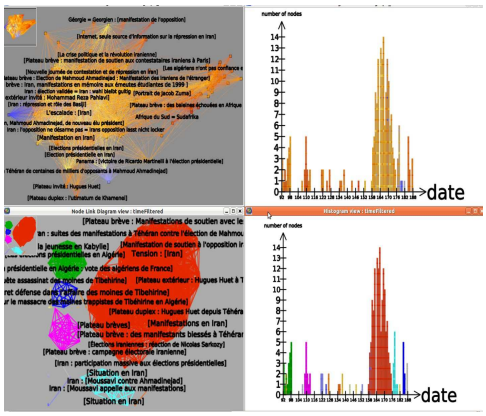

Figure 5 : Un cluster (élections présidentielles) avant et après raffinage temporel avec correspondance de couleur des sous-agrégats dans l'histogramme de temps. On observe la création d'agrégats évènementiels spécifiques : le gros cluster rouge concerne les élections en Iran et en bleu clair un écho de cet évènement, les élections en algérie en vert et l'affaire des moines de Tibehirine en bleu, et les élections en Afrique du Sud en rose.

A partir de cette carte, nous proposons une vue de chaque cluster pour en étudier le contenu. Le graphe de description est présenté avec la taille des nœuds correspondant au nombre d'occurrences de chacun des descripteurs dans les arêtesdu graphe des documents. Une vue de la répartition temporelle des documents est disponible sous forme d'histogramme. La sélection est synchronisée sur les différentes vues (Figure 4). Enfin après filtrage et raffinage temporel, une colorisation des composantes connexes permet de corréler les sous-clusters avec leur position dans l'histogramme (Figure 5).

6.2 Résultats et discussion

Afin de valider notre clustering, nous avons procédé en 3 étapes de mesures, après le premier clustering, après filtrage des nœuds isolés du graphe des descripteurs, et après raffinage temporel. On applique nos mesures sur les sous-clusters d'au moins 5 nœuds. Nous avons aussi étendu la définition de densité de graphe au modèle multicouche en la calculant comme suit : $\frac{2*\sum n_i}{N*(V*(V-1))}$ où n_i représente le nombre d'arêtes contenant le descripteur *i*, *N* étant le nombre de descripteurs présents sur le cluster et *V* le nombre de nœuds du cluster.

Après la première étape, nous générons 114 clusters, en conservant 77.5% des nœuds et 59.3% des arêtes du graphe original. La taille moyenne d'un cluster est de 53.9 nœuds pour 856.7 arêtes. La densité moyenne est de 0.133 pour une cohérence moyenne de 0.193.

Après filtrage des nœuds isolés, nous conservons toujours 114 clusters, avec 77.4% et 59.3% des nœuds et des arêtes. Un cluster a en moyenne 53.3 nœuds et 848.9 arêtes. La densité moyenne passe à 0.140 et la cohérence à 0.208.

Après raffinage temporel, en ne considérant que les clusters d'au moins 5 nœuds, nous conservons 46.0% et 20.1% des nœuds et arêtes d'origine et obtenons 254 clusters de 14.2 nœuds pour 128.6 arêtes en moyenne. La densité moyenne y est de 0.384 et la cohérence de 0.430. 31.4% des nœuds d'origine restent au sein des premiers clusters obtenus avant le raffinage.

Les gains en termes de densité et de cohérence sont excellents et valident notre procédure. Nous avons pu malgré tout constater que la mesure de cohérence reste faible pour les clusters contenant soit beaucoup de nœuds, soit beaucoup de descripteurs, même si ces clusters présentent pour l'observateur une bonne cohérence et peu de bruit. Nous avons observé deux cas extrêmes entre lesquels se placent nos clusters, ceux qui sont fondamentalement thématiques, et ceux qui sont ciblés sur un évènement particulier. Lorsqu'il contient beaucoup de nœuds ou de descripteurs, un cluster se trouve souvent à mis chemin entre la thématique et l'évènement. Le raffinement permet l'expression de ces clusters « mixtes » (Figure 5).

Nous n'avons pas encore effectué de tests utilisateurs car l'ergonomie globale du système doit être optimisée, mais nos observations nous permettent déjà de constater que la mesure de cohérence calculée nous permet d'identifier visuellement

les sujets d'actualités les plus saillants dans notre corpus (agrégats de taille moyenne très colorés comme le tour de France), et les thématiques principales (gros agrégats de couleur pale comme les élections européennes).

7. Conclusion et perspectives

Nous avons présenté dans ce papier une méthode pour la détection d'évènements médiatiques à partir de sujets de journaux télévisés annotés par des descripteurs textuels. Nous avons proposé une implémentation qui permet non seulement de réaliser un clustering mais qui place l'utilisateur au centre du système en lui proposant un retour visuel de la qualité des agrégats basé sur une mesure de cohérence sémantique et des outils de raffinement sémantique et temporel. La carte ainsi générée nous a permis de vérifier les deux tendances d'agrégation présentes dans nos données, la thématique et l'évènementiel (Figure 6).

Le cadre du projet OTMedia va nous amener aussi à tester très prochainement notre méthode sur des données issues du TAL. Nous souhaitons aussi appliquer notre méthode à d'autres corpus tels que la base Pascal de l'Inist-CNRS, afin d'évaluer nos résultats au regard d'autres travaux. Le passage à l'échelle de la mesure de cohérence devra être testé (OTMedia capte plus de 1000 sources d'information par jour). Du point de vue de la visualisation, de gros efforts sont encore nécessaires pour permettre la fouille visuelle dynamique dont rêvent nos utilisateurs, tant au niveau ergonomique qu'en niveau visuel. Des tests utilisateurs seront conduits tout au long du projet. Enfin, une question reste ouverte : Comment réintroduire les sujets filtrés par les différents traitements et trouver un angle d'observation particulier pour ces documents?

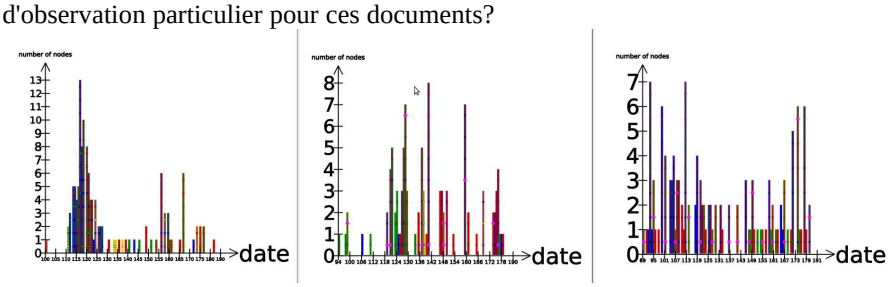

Figure 6 : 3 histogrammes (colorés par chaîne) de temps de clusters démontrant à gauche les caractéristiques temporelles d'un évènement (apparition et premières mesures pour la grippe A), à droite d'une thématique (sorties cinéma), et au milieu, mixte, d'évènements dans une thématique (intempéries, tempêtes, orages).

Bibliographie

[1] Allan J., "Introduction to topic detection and tracking", *Topic Detection and Tracking* Kluwer Academic Publishers, 2002

[2] Allan J., "Topic detection and tracking: event-based information organization," The Kluwer International *Series On Information Retrieval*, 2002

[3] Auber D., Mary P.,Mathiaut M., Dubois J., Lambert A., Archambault D., Bourqui R., Pinaud B., Delest M. and Melançon G., "Tulip: a Scalable Graph Visualization Framework", *Extraction*, 2010

[4] Brewer C. and Harrower M., Colorbrewer: Color advice for maps. *http://colorbrewer2.org*, 2009
[5] Burt R.S. and Schott T., “Relation contents in multiple networks”, *Social Science Research* Elsevier, 1985
[6] Cointet J.P., Faure E., Roth C.. “Intertemporal topic correlations in online media”. *Proceedings of the 1st ICWSM International Conference on Weblogs & Social Media*, Boulder, Col., E.-U., 2007
[7] Ding, J. and A. Zhou (2009). Nonnegative Matrices, Positive Operators and Applications. Chap. 1 and 2 Singapore, World Scientific.
[8] Esquenazi, J.P., “L'écriture de l'actualité : pour une sociologie du discours médiatique”, 2002, Presses universitaires de Grenoble
[9] Fung G.P.C.,Yu J.X., Yu P.S., and Lu H., “Parameter free bursty events detection in text streams”, *Proceedings of the 31st international conference on Very Large Data Bases*, 2005
[10] Herman I. and Melançon G. “Graph visualization and navigation in information visualization: A survey”, *IEEE Transactions on Visualization and Computer Graphics*, 2000
[11] Hermelin C., “Apprendre avec l’actualité. Théorie et pédagogie de l’événement”. Paris, Retz, 1993, p. 32.
[12] Lamirel J.C., “Vers une approche systémique et multivues pour l'analyse de données et la recherche d'information: un nouveau paradigme”,HAL: *HDR*, 2010
[13] Nallapati R., Feng A., Peng F., and Allan J., "Event threading within news topics," *Conference on Information and Knowledge Management*, 2004.
[14] Nallapati R. and Cohen W., "Link-PLSA-LDA: A new unsupervised model for topics and influence of blogs," *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, Association for the Advancement of Artificial Intelligence, 2008.
[15] Noack A., “Energy models for graph clustering”, *Journal of Graph Algorithms and Applications*, 2007
[16] Noack A., “Modularity clustering is force-directed layout”, *Physical Review* E, 2009
[17] Patton R.M. and Potok T.E., "Discovering event evidence amid massive, dynamic datasets," *Genetic And Evolutionary Computation Conference*, 2007.
[18] Pouliquen B., Steinberger R., and Deguernel O., "Story tracking: linking similar news over time and across languages," Coling: *Proceedings of the workshop Multi-source Multilingual Information Extraction and Summarization*, Manchester, 2008
[19] Salton G. and McGill M., “Introduction to Modern Information Retrieval”. *McGraw-Hill*, 1983.
[20] Von Landesberger T., Kuijper A., Schreck T., Kohlhammer J., Van Wijk JJ, Fekete JD and Fellner D., “Visual analysis of large graphs“, Proceedings of Euro-Graphics: State of the Art Report, 2010
[21] Zhao Q., and Mitra P. “Event Detection and Visualization for Social Text Streams“, International Conference on Weblogs and Social Media. (ICWSM’07), Bouler, CO, 2007