ina SUP

Pôle européen
des sciences et métiers
de l'image et du son

Détection, visualisation et validation d'évènements médiatiques

INA/LaBRI - Benjamin Renoust – EGC – 20/06/11

Interface d'exploration et de manipulation pour la détection d'évènements médiatiques

Contrainte :

→ Mettre l'utilisateur au centre du système

Objectifs :

→ Produire des cartes visuelles

→ Retour sur la qualité des évènements proposés par le système

→ Validation ou invalidation des résultats

→ Corrélation avec des vues complémentaires

→ Manipulation des critères de visualisation

→ Possibilité de filtrage et de raffinement

Cadre du projet

Observatoire TransMedia (OTMedia[1]) (projet ANR)

Objectif du projet : Analyse de la propagation des évènements médiatiques sur la TV, la radio, la presse, le web et twitter

Tâches : capture, analyse linguistique, analyse audio, analyse d'image, moteur de recherche et indexation, fouille de données, visualisation, analyse en sciences sociales

Sources : + de 300 acteurs nationaux

dépêches AFP, 15 chaînes TV, 10 en radio, 15 journaux, 250 blogs et sites web (sites institutionnels, politiques, commentateurs...) et ~20000 fils twitter

Partenaires : AFP, CIM-Paris 3, Ina, INRIA, LIA, Syllabs

[1] http://www.otmedia.fr

Visualisation

Objectifs : apporter des visualisations interactives à différents niveaux pour faciliter l'analyse SHS du paysage informationnel

Besoins : agrégation de sujets d'information pour **identification** et **quantification** des évènements médiatiques et de leurs caractéristiques (sources impliquées, sémantique, répartition temporelle, dynamique de diffusion...)

→ modélisation du problème par un **graphe de similarité** entre documents (multi-couches car plusieurs types de liens sont possibles, descripteurs visuels, descripteurs textuels, liens de reprises...)

→ construction du graphe via les **KNN** extraits d'une **similarité cosinus** entre les vecteurs de description des documents

Caractéristiques : petit monde, scale free, grande taille (suivant l'échantillon)

Données

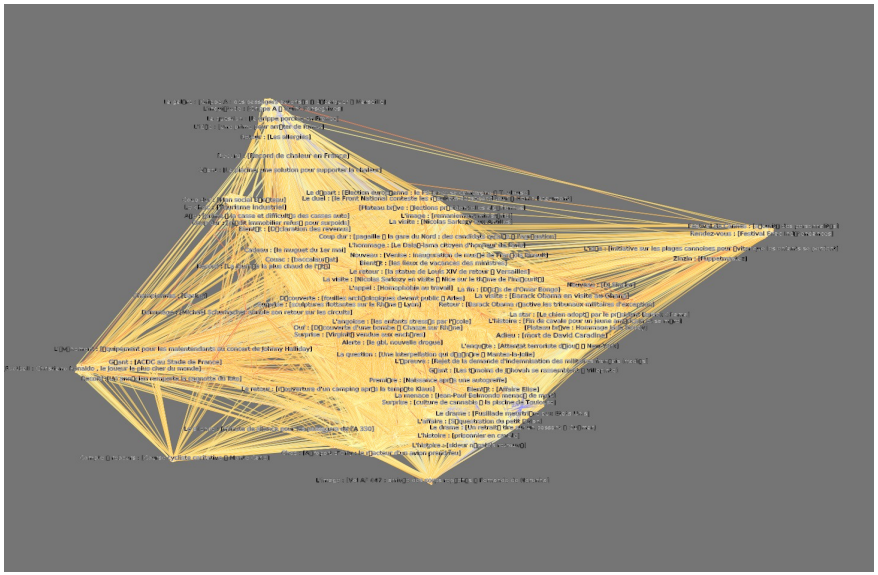
Étude des JT de M6, sur 6 mois (V~1600 E~27000)

→ chaque JT est divisé en sujets (1 sujet = 1 document, ~10 sujets/JT)

→ chaque sujet est annoté par une liste de descripteurs textuels (2 à 10 mots clefs définis par les documentalistes Ina, ou mots saillants extraits par TAL)

1 nœud = 1 document + descripteurs
1 lien = liste de descripteurs partagés par 2 documents

→ des documents « proches » dans le modèle sont « sémantiquement proches » car partagent des descripteurs sémantiques



Événements médiatiques

Selon les SHS [2] :

- un **événement** est un fait survenu dans la réalité
- un **sujet médiatique** propose un « cadrage narratif » d'un évènement
- un **évènement médiatique** génère une densité forte de **sujets médiatiques** se rapportant à un même évènement

Hypothèse 1 : deux sujets parlant d'un même évènement sont proches sémantiquement

Hypothèse 2 : un agrégat de documents implique peut être l'existence d'un événement médiatique

Notre objectif est d'éliminer **les densités ne correspondant pas à des évènements**:

→ Certains regroupements ne sont que thématiques
→ Polysémie et mots valises introduisent du bruit
→ Une description trop fine ou large peut agréger ou isoler un document

[2] L'écriture de l'actualité: pour une sociologie du discours médiatique, Jean-Pierre Esquenazi 2002

Visualisation des documents

Besoins :

→ Clustering des documents pour en extraire les évènements

→ Rendu visuel de la carte médiatique des documents

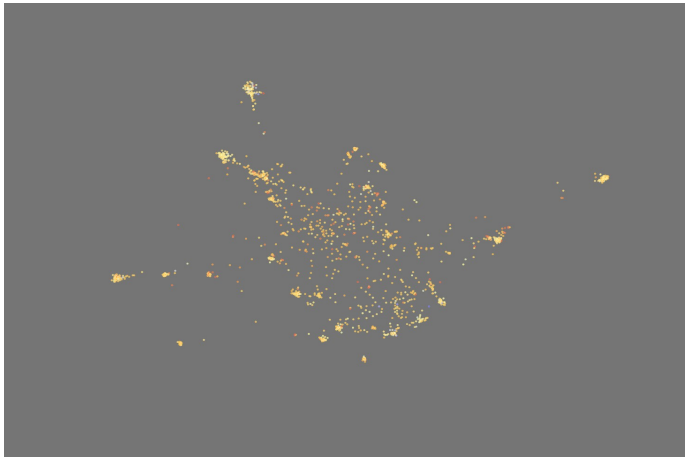
Layout à base de forces et clustering (diminue les distances intra-cluster et augmente les distances inter-clusters) :

Edge Linlog layout d'Andreas Noack[3] qui se rapproche d'un Modularity clustering[4]

Paramètres par défaut :

Répulsion logarithmique

Attraction linéraire



[3] Energy Models for Graph Clustering, Andreas Noack, 2007

[4] Modularity Clustering is Force-Directed Layout, Andreas Noack, 2008

Détection des évènements

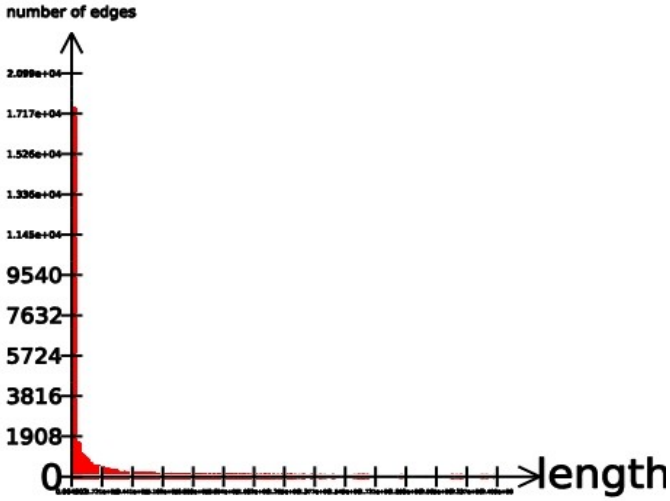
Après calcul du layout, **filtrage** de l'histogramme des longueurs d'arêtes par sélection du 1er % des valeurs de longueurs

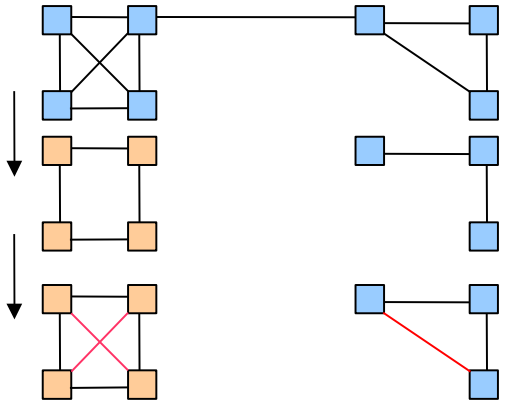
Détection des **composantes connexes** qui correspondent à des clusters

80% des documents appartiennent à un cluster

Ajout des arêtes précédemment filtrées, internes aux clusters

40% des arêtes sont sélectionnées

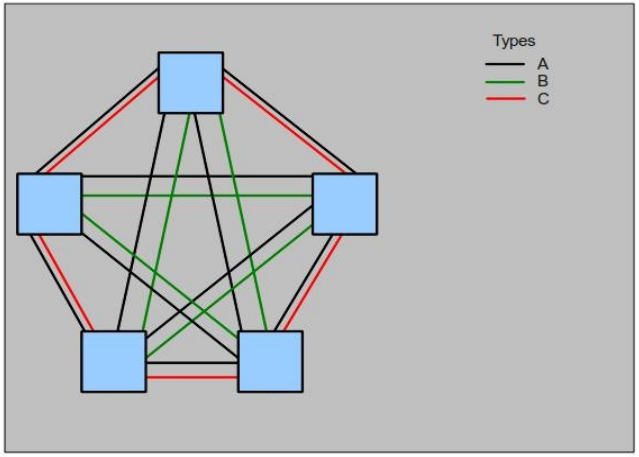


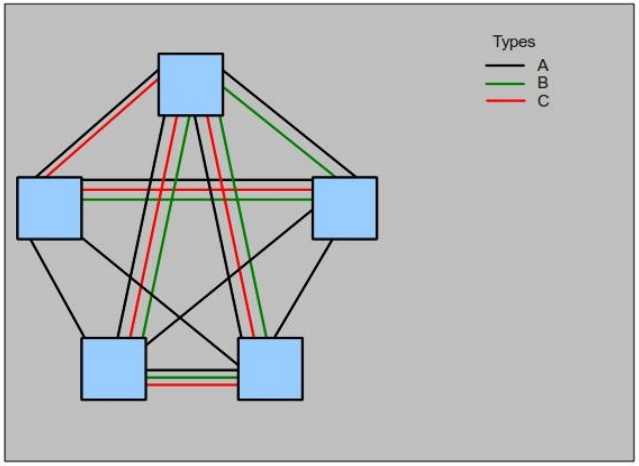


Validation des évènements

Certains clusters fédèrent le bruit
Certains documents hors sujet sont ajoutés à un cluster
Certains descripteurs rassemblent plusieurs évènements

Objectif : retour utilisateur sur la qualité des évènements calculés
On veut mesurer la « cohérence » d'un cluster
Hypothèse : **taux d'intrication** des descripteurs dans un cluster est lié à sa cohérence





Mesure d'intrication

Burt et Schott [5] s'intéressent aux réseaux sociaux multiples. Ils cherchent à caractériser les groupes de personnes et la nature de leur relations :

→ **Ambiguïté** des types de relation dans un réseau social

→ **Fréquence de co-occurrence** de ces types entre deux individus dans le réseau

Plus des types de relations apparaissent ensemble sur les liens du réseau, plus ces types sont considérés comme ambigus

Dans notre contexte :

Ambiguïté des relations = Intrication des descripteurs

Types de relation = descripteurs

Plus des descripteurs apparaissent ensemble dans un cluster, plus le cluster a de chances d'être cohérent

/! mesure relative au cluster observé

(Corrèze en politique et Corrèze en géographie)

[4] Relation Contents in Multiple Networks, Ronald Burt et Thomas Schott, 1985

Mesure d'intrication intra cluster

Construction de la matrice de fréquence des co-occurrences:

$$C=\begin{bmatrix} cii & cij & \dots \\ cji & \dots & \dots \\ \dots & \dots & \dots \end{bmatrix}$$

avec **cii** = **ni/N**, **ni** nombre d'occurrences du type **i**

N nombre de liens total

cij = **nij/ni**, **nij** nombre d'occurrences du type **i** sachant le type **j**

Évaluation de la mesure globale d'intrication **λ** (celle du type le plus intriqué) :

$\lambda = \max(\text{eigenval}(C))$

Le vecteur propre **v** associé à **λ** décrit la participation de chaque type dans son intrication avec **λ**

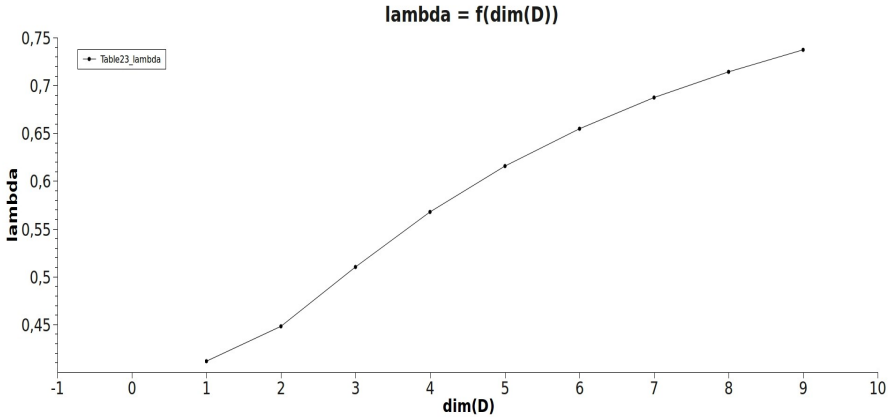
Utilisation de **v** pour l'étiquetage du cluster

Mesure d'intrication intra cluster

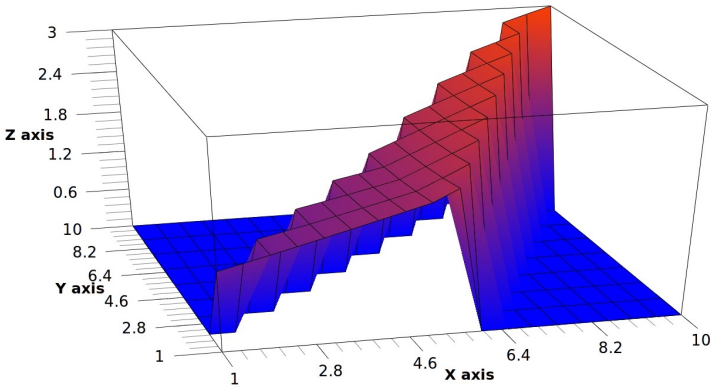
Caractérisation de la mesure d'intrication :

Positive et maximisée à **N**, nombre de types (Théorème de Perron-Frobenius [6])

Croissante sur l'augmentation du nombre d'intrications (par ajout de liens intriqués supplémentaires, par ajout de dimensions = types, et par permutations des liens)



Lambda = f(dim(D))



Lambda = f(E(b,c),perm(b,c))

[6] http://en.wikipedia.org/wiki/Perron%E2%80%93Frobenius_theorem

Graphe dual

Cas de C diagonale par bloc → cluster *multiple*
Traitement de **A** et **B** séparé

$$C = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$$

C peut être considérée comme définissant le graphe d'interaction des types avec

nœuds : types de liens

arêtes : 2 types partagent un lien dans le graphe initial

Peut avoir plusieurs composantes connexes → cluster *multiple*

Filtrage des types hors sujet ou générant le bruit
Proposition de raffinement du clustering

Visualisation et Interfaces

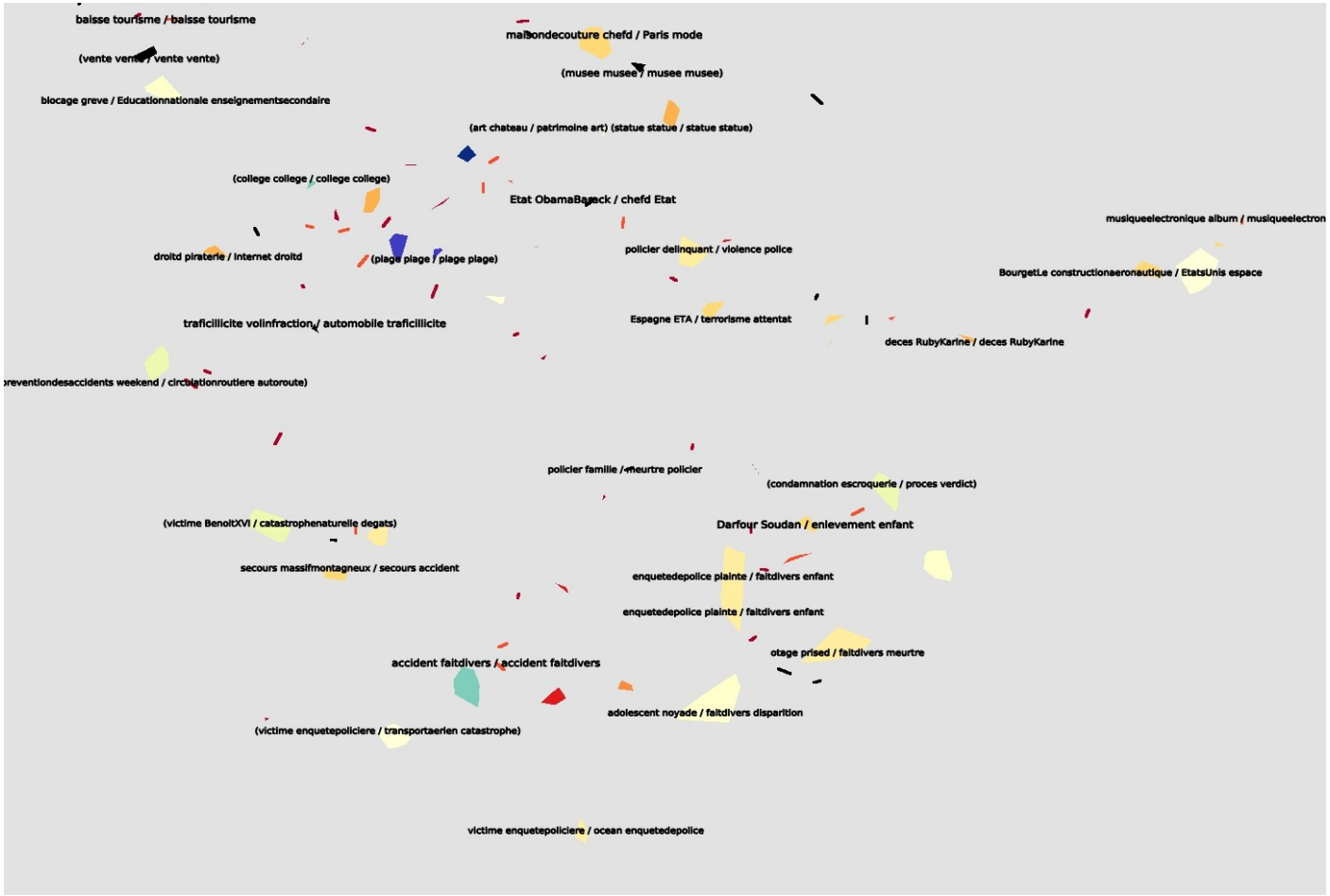
Clustering par layout de force

Taux d'intrication du plus clair au plus foncé

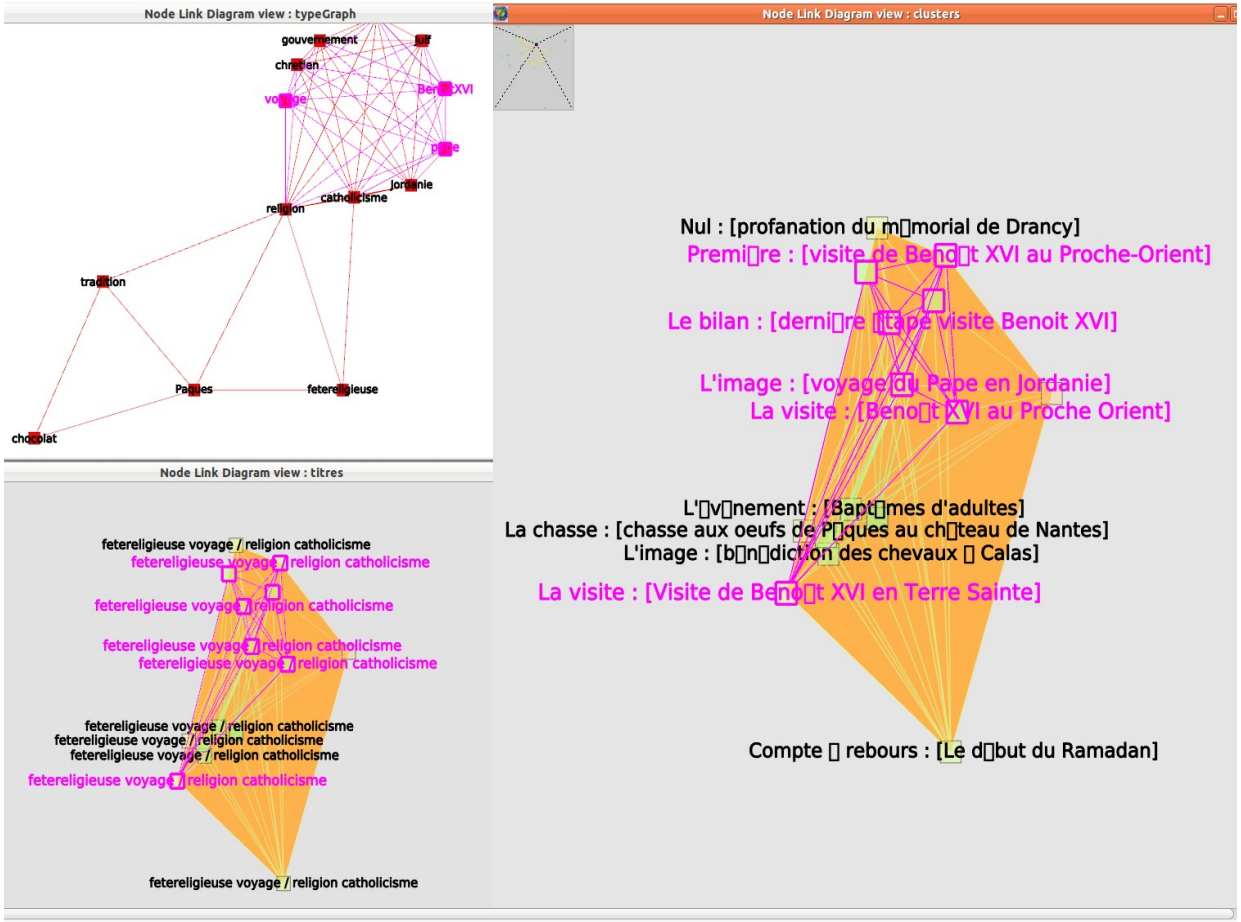
Echelles de Brewer pour les Clusters :

orange/rouge pour les clusters Simples

vert/bleu pour les clusters multiples



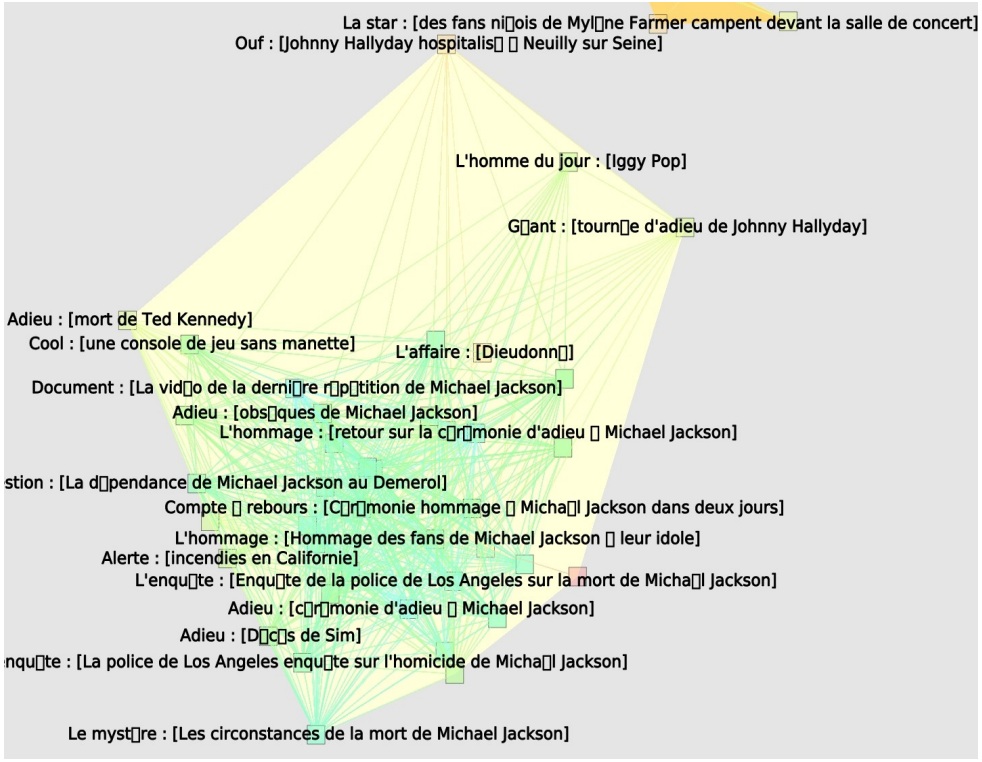
Visualisation et Interfaces

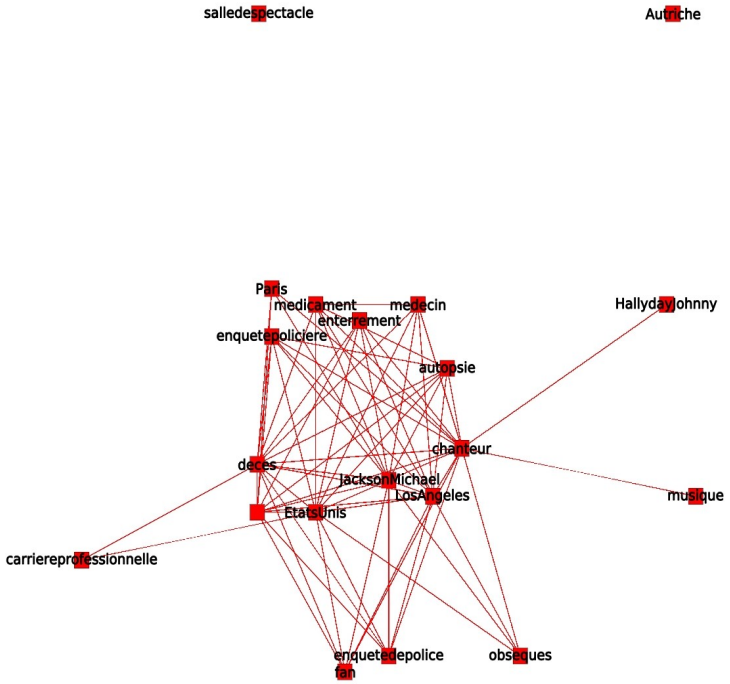


Cluster simple étiqueté « *fête religieuse, voyage, religion, catholicisme* » concernant le voyage du pape Benoit XVI

En rose, sélection **synchronisée** entre les descripteurs *voyage*, *pape* et *Benoît XVI* et les documents liés à ces descripteurs

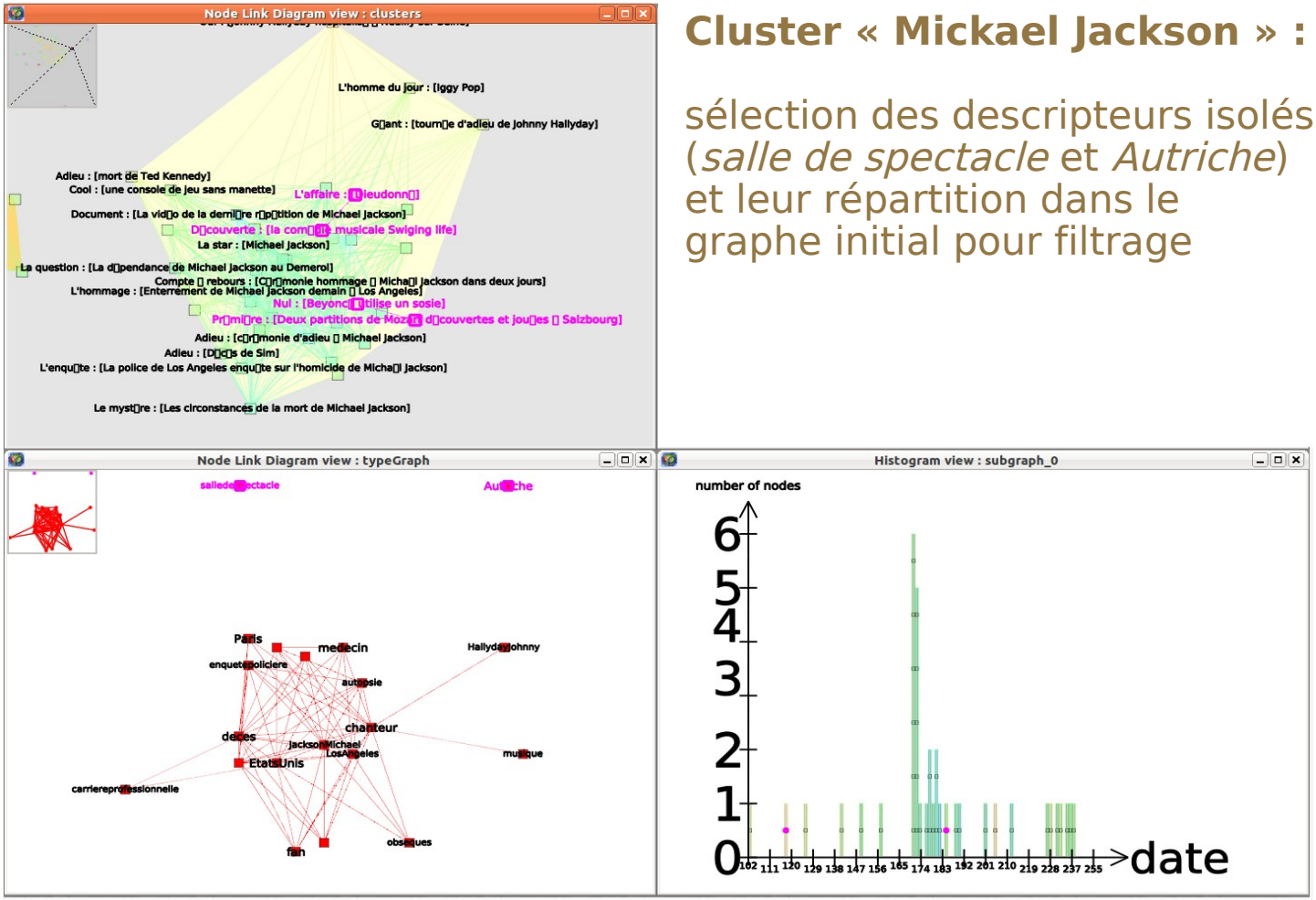
Visualisation et Interfaces





Le cluster multiple « *Michael Jackson* » et son dual **non connexe**

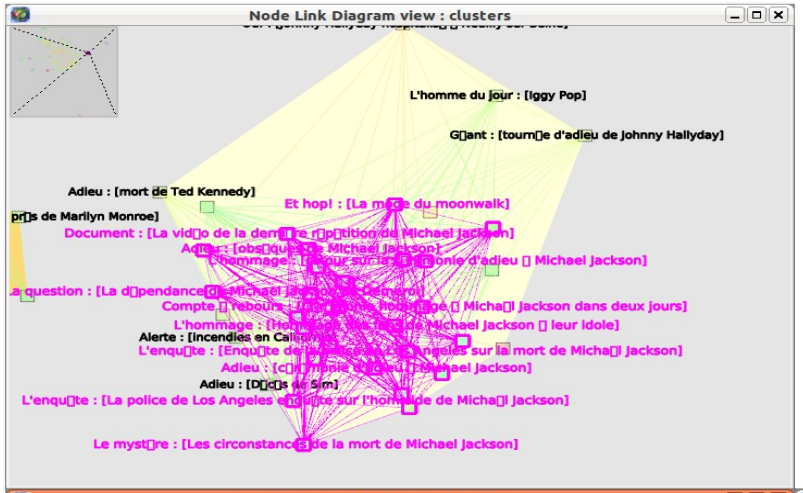
Visualisation et Interfaces

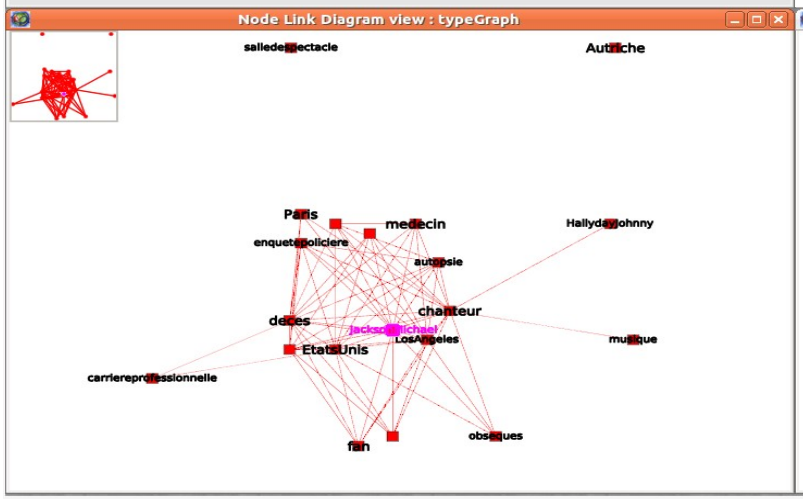


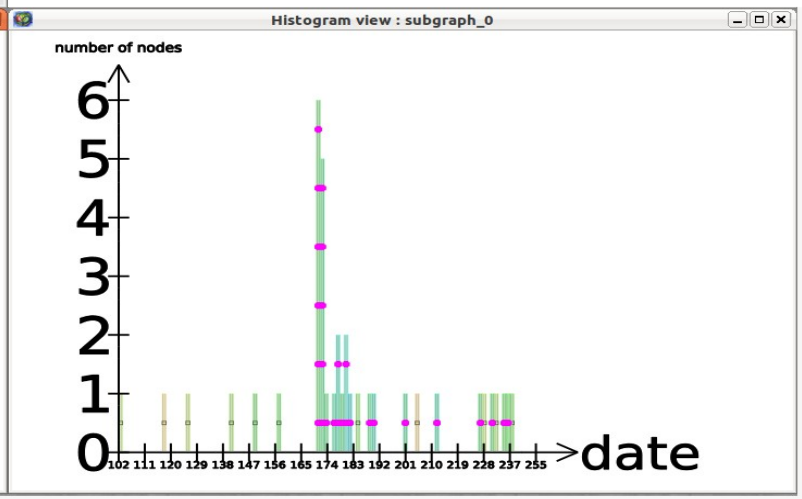
Cluster « Mickael Jackson » :

sélection des descripteurs isolés (*salle de spectacle* et *Autriche*) et leur répartition dans le graphe initial pour filtrage

Visualisation et Interfaces



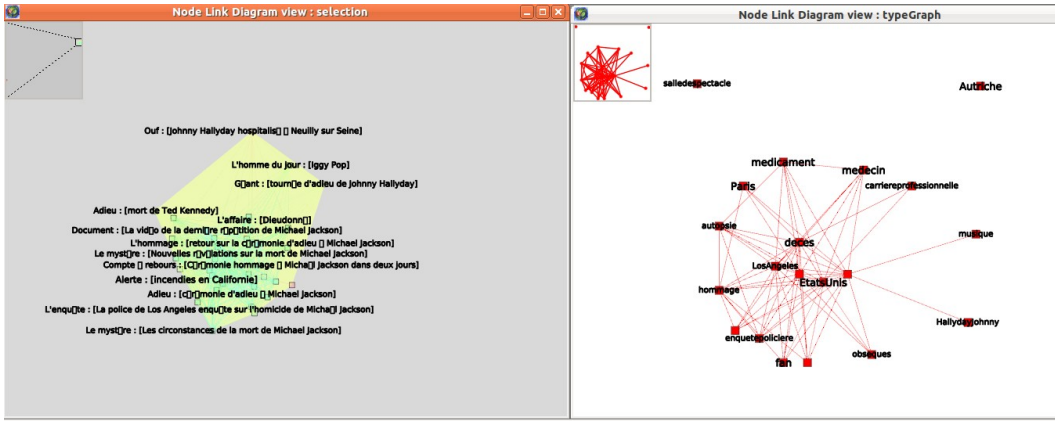


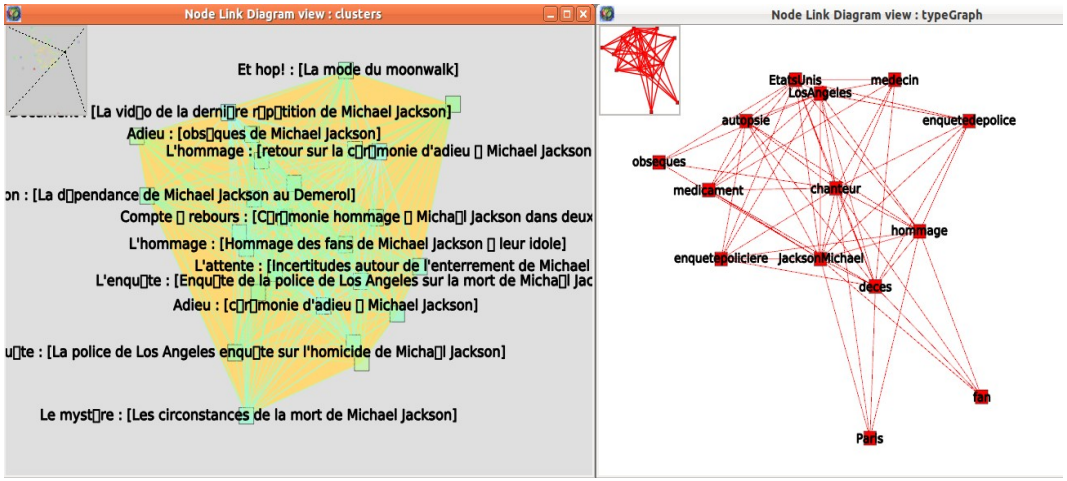


Cluster « Mickael Jackson » :

sélection des nœuds liés au descripteur « *Mickael Jackson* » et leur répartition dans le temps pour validation

Visualisation et Interfaces





Cluster « Mickael Jackson » :

En haut : avant filtrage, intrication = 0.109

En bas : après filtrage, intrication = 0.328

Résultats et Limitations

les clusters larges ont tendance à avoir un faible taux d'intrication et à amasser le bruit

→ raffinage nécessaire

lien temporel non pris en compte dans la notion d'évènement

→ à la création du graphe et dans le layout

appartenance d'un nœud à plusieurs clusters

→ cas isolés, à la discrétion de l'utilisateur

Travaux en cours et futurs

Validation de la mesure :

→ amélioration de l'interface pour proposer des recherches par clusters avec un second tour de génération de clusters après inclusion ou exclusion de documents

→ gestion des classes d'équivalence de types dans un graphe

→ ajout de nouvelles sources d'information pour augmenter la redondance

→ création d'une vérité-terrain pour contrôler les résultats

Améliorations :

→ gestion de la dimension temporelle, pondération du layout par le temps, détection de clusters dynamiques (T. Aynaud et JL. Guillaume 2010 [7])

→ passage à l'échelle

Autres applications :

→ identification des sources et des reprises pour le traçage de l'information

→ application aux similarités d'images avec les regroupements de zones d'images issues de descripteurs SIFT

[7] Détection de communautés à long terme dans les graphes dynamiques, Thomas Aynaud et Jean-Loup Guillaume, 2010

Références

Réalisations faites en C++ et python en utilisant le framework Tulip[8]

[1] http://www.otmedia.fr

[2]L'écriture de l'actualité: pour une sociologie du discours médiatique, Jean-Pierre Esquenazi 2002

[3] Energy Models for Graph Clustering, Andreas Noack, 2007

[4] Modularity Clustering is Force-Directed Layout, Andreas Noack, 2008

[5] Relation Contents in Multiple Networks, Ronald Burt et Thomas Schott, 1985

[6] http://en.wikipedia.org/wiki/Perron%E2%80%93Frobenius_theorem

[7] Détection de communautés à long terme dans les graphes Dynamiques, Thomas Aynaud et Jean-Loup Guillaume, 2010

[8] http://tulip.labri.fr