TP4 : Analyse complète autour de données de la dengue

Yoann Pitarch

1 Objectifs

Ce dernier TP va être l'occasion pour vous d'effectuer une analyse complète autour de données associées à la dengue. Dans un premier temps, vous répèterez les étapes d'acquisition de données, de premiers comptages, de génération de dictionnaires de mots vides et de synonymes et de comptage de co-occurrences. Ces étapes sont obligatoires comme prélude à une bonne analyse de données avec Tetralogie. Mais vous découvrirez également quelques analyses poussées telles que l'Analyse Factorielle des Correspondances (AFC) ou le Clustering Ascendant Hiérarchique (CAH). Nous verrons également comment combiner le résultat de ces méthodes pour affiner l'analyse.

2 Instructions

1. Dans la base PubMed, faites une recherche associée à la dengue.
2. Exportez les résultats au format Medline et ouvrez ce jeu de données dans Tetralogie.
3. Effectuez un premier comptage sur les champs `pays` (PA), `email`, `auteurs`, `journaux`, `MESH` et `années`. Quel est le journal le plus important dans ce domaine ? Considérez maintenant les pays de publication. Que pouvez vous en dire ?
4. Construisez un dictionnaire des synonymes pour diviser les années de publication en 4 périodes (`<1990}`, `1990-1999`, `2000-2010` et `>2010`).
5. Appliquer un filtre positif, confectionné par vos soins à l'aide d'une commande UNIX, sur le champs `email` pour n'avoir réellement que des emails comme valeurs de cet attribut.
6. Croisez les dimensions `pays` et `années` et visualiser le résultat sur une carte. **Attention,** lors du croisement de variables, pour que les synonymes du et les filtres soient pris en compte, effectuez les traitements dans cet ordre : (1) appliquez le fichier de synonymes (en le précisant sur la ligne correspondante à la dimension `Date`), (2) créez un filtre négatif à partir du fichier `PA-DP.var` et (3) relancez un croisement en précisant à la fois le filtre négatif et le dictionnaire des synonymes. Pouvez-vous confirmer ce que vous avez observé lors de la question 3 ?
7. Il est possible de classer les pays selon leur profil d'évolution en cliquant sur `Classer` dans l'interface de visualisation de l'information géostratégique. Effectuez ce traitement et commentez-le.

8. Sur le croisement `Pays/Date de publication`, lancez un CAH. En faisant un clic droit dans la fenêtre de visualisation du clustering, il est possible d'exporter le résultat de ce clustering (clic droit sur `export_c`). Effectuez, cet export pour visualiser sur la carte le résultat du clustering.
9. Pour analyser l'évolution d'un attribut, il est également possible d'effectuer une Analyse Factorielle des Correspondances (AFC) sur l'attribut en question et l'attribut temporel. Nous illustrerons cela en choisissant l'attribut `Journaux`. Appliquer une AFC sur le croisement `Journaux/Date` et visualiser en 4D le fichier résultat stocké dans `AFC_V_1`. Effectuez quelques rotations de la visualisation (clic droit, choisissez l'axe de rotation puis cliquez sur la visualisation).
10. Sans fermer la visualisation précédente, ouvrez une vue 4D de l'autre fichier généré (`AFC_I_V_1`). Vous pouvez exporter la rotation de la première visualisation via le menu contextuel `Rotations->>`. Récupérez quelques journaux qui ont essentiellement publiés des articles récemment sur cette thématique. Placez cette liste dans un fichier pour qu'il puisse éventuellement servir de filtre.
11. Nous allons maintenant illustrer comment les méthodes d'analyse peuvent coopérer.Pour cela nous allons appliquer un CAH sur le résultat d'une AFC. Cela permettra de créer des clusters de journaux présentant des profils d'évolution dans le temps similaire. Lancez donc un CAH sur `AFC_I_V_1`, visualisez la hiérarchie ainsi qu'une vue 4D de `AFC_I_V_1`. A partir de la hiérarchie, exportez le résultat du clustering pour faire apparaître 5 classes différentes. En analysant cette vue 4D, existe-t-il une classe particulièrement marquée ?
12. Maintenant que nous avons une bonne vue des journaux du domaine, nous allons nous intéresser aux auteurs et analyser le réseau social construit sur la relation de co-écriture d'articles. Pour permettre une visualisation claire du réseau, nous faisons le choix de ne considérer que les 150 auteurs les plus importants, i.e., les plus prolifiques. Construisez un filtre positif pour ne sélectionner que ceux-ci.
13. Effectuez le croisement `Auteur/Auteur`, visualisez la matrice et faites une première analyse en vous aidant des outils associés à cette visualisation matricielle.
14. Testez la visualisation circulaire et l'animation `Forces semi-paramétrées` semi-paramétrées. Modifiez les paramètres de visualisation pour faire ressortir distinctement les communautés.
15. Lancez un clustering stochastique pour faire clairement ressortir ces communautés.
16. Nous allons maintenant ajouter la dimension temporelle à cette analyse du réseau social. En effet, cette dimension est bien souvent essentielle pour comprendre réellement un réseau, comment sa structure évolue pour anticiper les changements majeurs. Effectuez un croisement `Auteur/Auteur/Date` (4 périodes). Le champ `Référence` permet de spécifier la dimension temporelle (il faut écrire la version longue du nom de l'attribut).
17. Visualisez le graphe et ses différentes instances (pas de temps). Que pouvez-vous dire ?