

Bibliométrie Indirecte par Analyse de Réseaux Complexes

J. Raimbault^{1,2}

¹Géographie-cités (UMR 8504 CNRS)

²LVMT (UMR-T 9403 IFSTTAR)

Séminaire *Cartha-Géo-Prisme*

Mercredi 17 février 2016

Données massives ?

Caractéristiques nécessaires possibles

- **Taille relative** : algorithmes et/ou structures de stockage non-conventionnels
- **Temps réel** : traitement d'un flux de données en temps réel
- **Hétérogénéité** : différentes sources, types, nature

Cas d'étude

Revue scientifique *Cybergeos* : analyse bibliométrique par approches variées

→ Enjeu par rapport au positionnement de la revue : interdisciplinarité ;
contre une bibliométrie quantitative pure en SHS

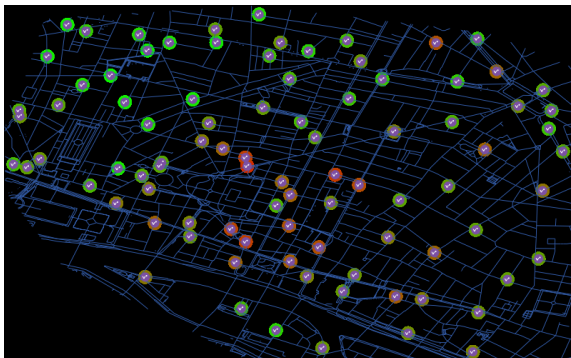
→ Elaboration d'une approche par *Hyperréseau* : croisement du réseau de
citations au réseau sémantique.

→ Données difficiles d'accès : base à construire

Collecte des données

Crawling de données semi-ouvertes : exemples en géographie

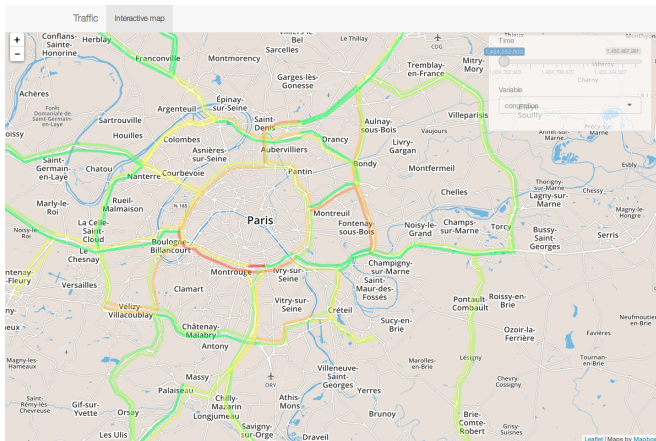
Données de mobilité : statuts des stations Vlib en temps réel (API)[?]



Collecte des données

Exemples en géographie (suite)

Traffic routier : collecte de *sytdin* (pas d'API : *scrapping* nécessaire)



Collection des données

Données des articles : structuration, nettoyage et consolidation (sources différentes)

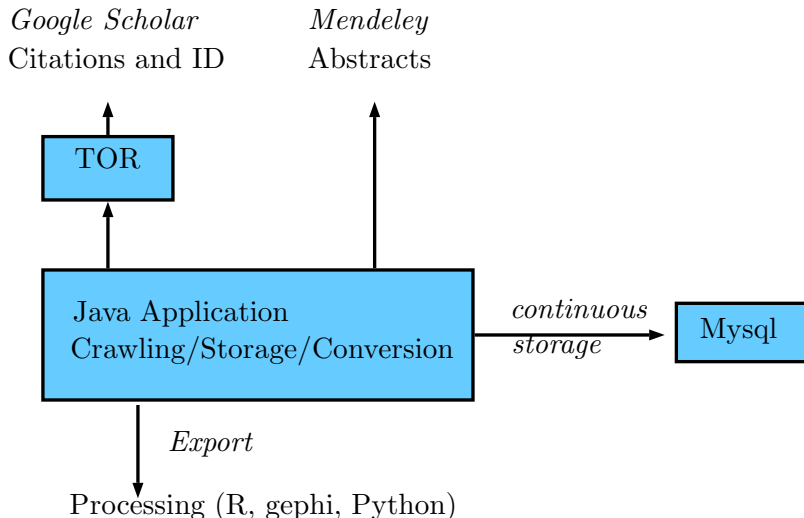
Données de citation : revue non référencée par base “classiques” (de plus non libres !)

→ *crawling* de google scholar par utilisation de l'option “*cité par*” [Noruzi, 2005]

Données textuelles : besoin des résumés pour l'ensemble des références liées

→ utilisation de l'API Mendeley [Mendeley, 2015] (gratuite mais non ouverte).

Architecture de collecte des données

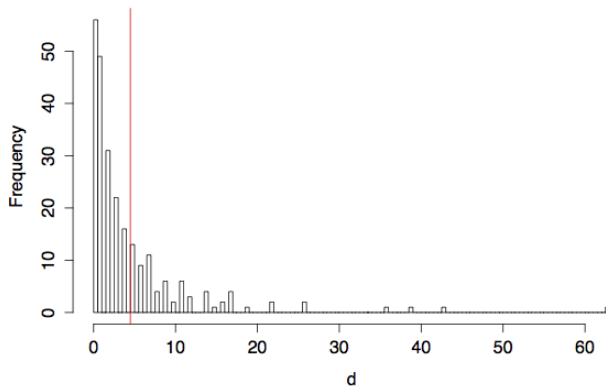


Caractéristiques du réseau

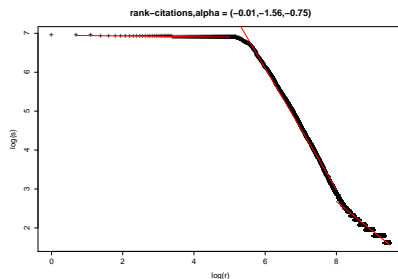
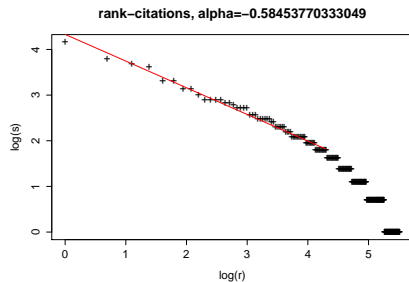
→ Après raffinement, $\simeq 947$ références de cybergeo exploitables, sur $\simeq 1200$; certaines inexistantes, d'autres mal enregistrées sur scholar

→ 418670 Noeuds et 570352 Liens ; Diamètre : 9 ; Densité : $3.25E-6$; degré moyen : 2.724284

Degree distribution, mean (impact factor) = 3.18



Degrés : Loi rang-taille

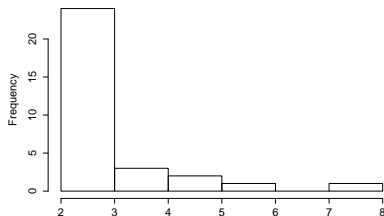


Gauche : Cybergéo ; Droite : Ensemble du réseau

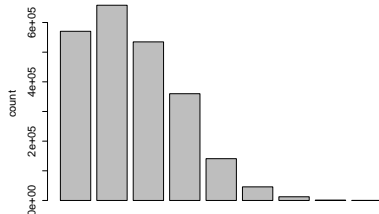
Clustering

Composante géante : plus de 99% des noeuds.

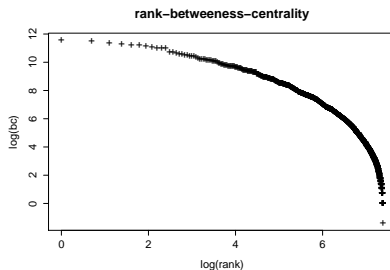
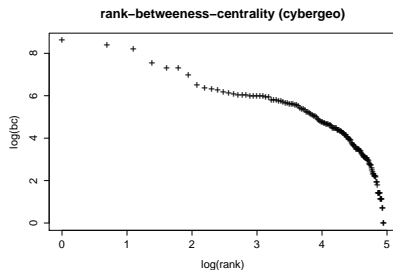
Weak clusters size without giant component



path length distribution



Centralité



Centralités faibles (rq : impossibilité des clusters forts pour des citations car causalité temporelle). Gauche : Cybergéo ; Droite : Ensemble du réseau

Cliques

viz cliques

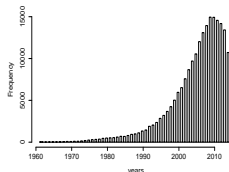
Réseau sémantique

Réseau sémantique. Collection des résumés/années/auteurs/mots-clés pour les 400000 références via l'API Mendeley → ~ 215000 références avec données complètes.

Statistiques

Langues : anglais 206607, français 4109, espagnol 2029, allemand 892, portugais 891, néerlandais 124, autres 182

Repartitions par années :



Extraction des mots-clés brute

Text-mining en python avec nltk [Bird, 2006], méthode adaptée de [Chavalarias a

- Detection de la langue par *stop-words* (mots vides de sens)
- Parsing et tokenizing (isolation des mots) /pos-tagging (fonction des mots) /stemming (extraction de la racine) effectués différemment selon la langue :
 - Anglais : pos-tagger intégré à nltk, combiné à un PorterStemmer
 - Français ou autre : utilisation de TreeTagger [Schmid, 1994]
- Selection des n-grams potentiels (avec $1 \leq n \leq 4$) : anglais $\cap \{NN \cup VBG \cup JJ\}$; franais $\cap \{NOM \cup ADJ\}$
- Insertion en base pour extraction quasi-instantane plus tard (10j \rightarrow 5min)

Estimation de la pertinence par bootstrap

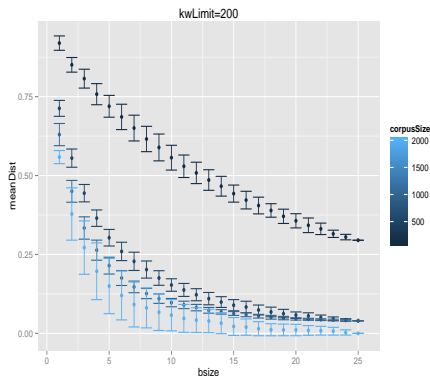
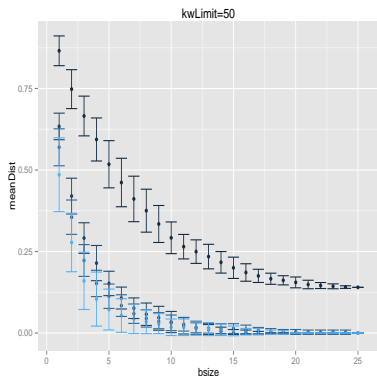
Estimation exacte de la pertinence via la repartition statistique des co-occurrences (score de χ^2) : *termhood* définie comme

$$t_i = \sum_{j \neq i} \frac{(M_{ij} - \sum_k M_{ik} \sum_k M_{jk})^2}{\sum_k M_{ik} \sum_k M_{jk}}$$

en $\Theta(\sum_i N_i^2)$ (N_i taille des résumés) : impossible sur un corpus où $\sum_i N_i^2 \simeq N < N_i >^2 \simeq 4 \cdot 10^6$

→ Estimation par *bootstrap* sur des échantillons du corpus : moyenne de la *termhood* sur B échantillons de taille C , avec nombre de mots clés K_L

Bootstrap : convergence de l'estimateur



Construction du réseau sémantique

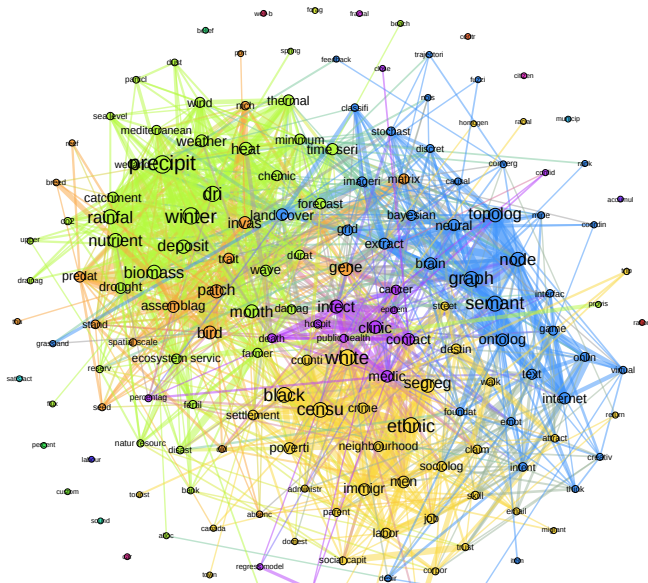
Noeuds : Mots-clés avec la plus grande pertinence cumulée

Liens : Co-occurrences pondérées

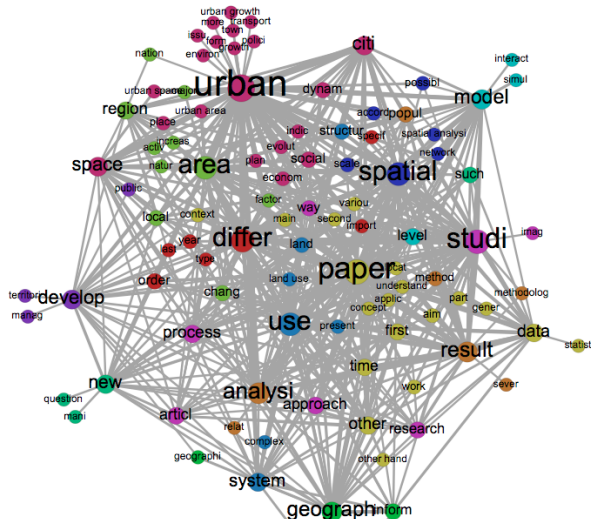
Filtrage des liens en dessous d'un seuil ; ajustement manuel de mots parasites

Détection de communautés par maximisation de modularité [Blondel et al., 2008]
après suppression des *hubs* (model, space, structure, process)

Réseau sémantique : corpus complet



Réseau sémantique : corpus cybergeo



Application : degré d'interdisciplinarité

Un article peut être associé aux communautés sémantiques par ses mots clés : probas p_i pour chaque communauté.

Mesure d'interdisciplinarité :

$$o = 1 - \sum p_i^2$$

Perspectives

TBW

References I



Bird, S. (2006).

Nltk: the natural language toolkit.

In [Proceedings of the COLING/ACL on Interactive presentation sessions](#), pages 69–72. Association for Computational Linguistics.



Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008).

Fast unfolding of communities in large networks.

[Journal of statistical mechanics: theory and experiment](#), 2008(10):P10008.



Chavalarias, D. and Cointet, J.-P. (2013).

Phylomemetic patterns in science evolution—the rise and fall of scientific fields.

[Plos One](#), 8(2):e54847.

References II



Mendeley (2015).

Mendeley reference manager.

<http://www.mendeley.com/>.



Noruzi, A. (2005).

Google scholar: The new generation of citation indexes.

Libri, 55(4):170–180.



Schmid, H. (1994).

Probabilistic part-of-speech tagging using decision trees.

In Proceedings of the international conference on new methods in language processing, volume 12, pages 44–49. Citeseer.