Extraction et chaînage supervisés de connaissances d'un corpus d'entretiens en histoire des sciences

Benjamin Hervy, Matthieu Quantin, Pierre Teissier

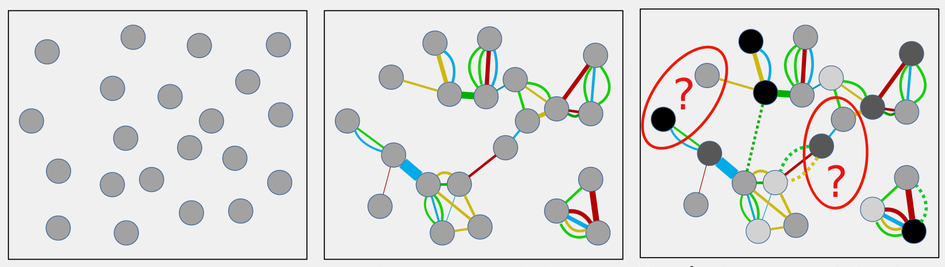
Extraction et chaînage supervisés de connaissances d'un corpus d'entretiens en histoire des sciences

Benjamin Hervy, Matthieu Quantin, Pierre Teissier
benjamin.hervy@univ-angers.fr, matthieu.quantin@ec-nantes.fr, pierre.teissier@univ-nantes.fr

Objectifs

Pour un groupe spécialiste d'un corpus de textes en histoire des sciences :
- confirmer ou infirmer des connaissances qualitatives existantes
- faire émerger de nouvelles pistes de recherche

Pour cela, nous cherchons à calculer des proximités entre les textes d'un corpus :
(1) sans modèle pré-défini; (2) sans vocabulaire de référence; (3) sans corpus d'entraînement.



Problématique

Comment mettre en évidence des chaînes de connaissance entre les textes d'un corpus tout en préservant la richesse terminologique non définie au préalable ?
Approche : créer un graphe multiple pondéré de co-occurrences d'expressions entre documents.

Méthode

Les données des sciences de l'homme forment souvent des corpus de textes, qui sont hétérogènes par leurs forme et contenus; spécifiques par leurs terminologie et signification. Nous présentons une méthode supervisée générant un réseau de documents liés par leurs proximités de contenus. Il s'agit d'un graphe multiple flou, basé sur l'extraction de *n-grams* à taille variable. L'extraction se base sur l'algorithme ANA pour construire des *MultiWord Expressions* (MWE) à partir des termes du corpus sans entraînement ni pré-traitement.
Un arc pondéré est généré par chaque co-occurrence de MWE entre deux documents. Le calcul du poids d'un arc est *poids* x *proximité* définis comme suit :

1. poids des MWE (adaptation d'**idf** avec cosinus pour créer un effet de seuil) :
   $poids = \left(\cos \frac{2-in\_docs}{nb\_docs}\right)^{factor_1}$ Ici, $factor_1 = 100$ empiriquement. $nb\_docs = 37$. in_docs : nombre de documents où la MWE apparaît.
2. proximité entre deux documents sur une MWE (adaptation de **tf** : le minimum d'occurrences favorise l'équi-répartition des termes entre deux documents) :
   $proximite = log_{10}((O^{t_i}_{p_1} + O^{t_i}_{p_2}) \times min(O^{t_i}_{p_1}, O^{t_i}_{p_2})^{factor_2})$
   Ici, $factor_2 = 3$ et $O^{t_i}_{p_j}$: nombre d'occurrences du MWE t_i dans le document p_j.

Contexte

Ce travail s'intègre dans une démarche de co-construction d'**outils heuristiques** avec l'historien pour l'écriture de l'**Histoire des sciences et des techniques**. La rétro-conception, documentation et valorisation d'objets du patrimoine industriel nous confrontent systématiquement à l'**extraction et la gestion de connaissances issues de textes** (archives, récits).

Corpus

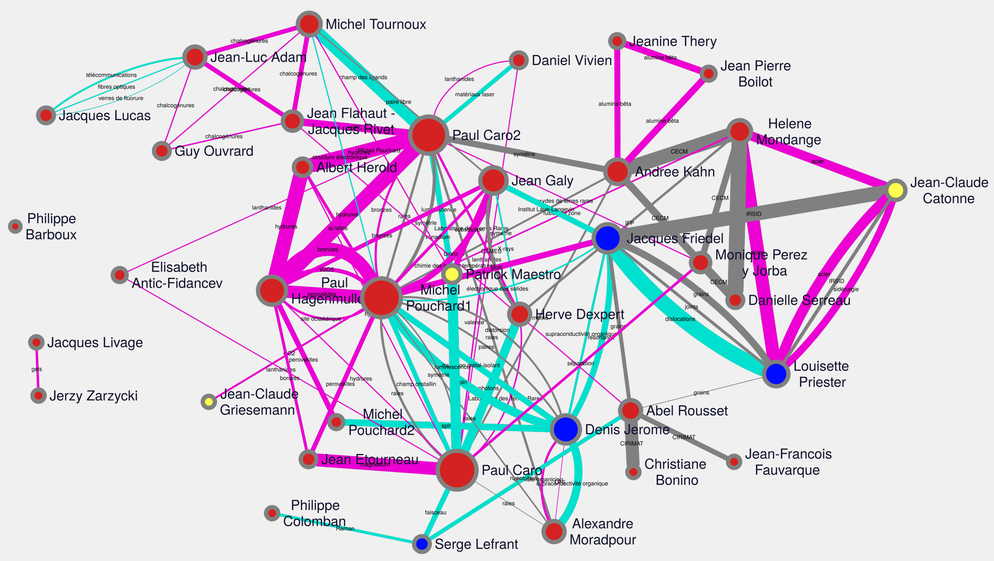
Le corpus est formé par la retranscription de **37 entretiens de chercheurs** racontant leur carrière parfois depuis les années 1940. Le corpus global contient **293k mots**. Il entrecroise des questions techniques (recherche), des énoncés relationnels et affectifs (interpersonnels) et des positionnements identitaires (discipline, génération, genre, etc.). Le corpus a été un objet d'étude (analyse "manuelle") en histoire des sciences depuis 2006.
→ ***Saisir les structures et les dynamiques d'une communauté scientifique***

Résultats

Graphe produit : 37 noeuds, 93925 arêtes.
Exemples de termes extraits : "Microscopie électronique à transmission à haute résolution", "École nationale supérieure de chimie de Paris", "Émission thermoionique", "résonance magnétique nucléaire".

Mot-clé	Pond.	occ. A	occ. B	occ corpus
bronzes	0.58	13	33	59
perovskites	0.5632	4	18	27
Oxyde de vanadium	0.38	2	7	9
site octaédrique	0.35	2	4	6
sites tétraédriques	0.31	2	2	4
cobalt	0.31	2	13	19
octaèdres	0.23	2	10	19
cations	0.19	6	7	27
John Goodenough	0.18	5	11	26
transition métal-isolant	0.15	1	6	11
Configuration électronique	0.15	1	6	10

Exemples de termes extraits par ANA, de répartition des occurrences de ces termes et de pondération du lien de co-occurrence créé.
Liste tronquée des liens entre *M. Pouchard1* et *P. Hagenmüller*.



Graphe (vue) obtenu par requête. Les liens en bleus sont relatifs à la physique, les rouges à la chimie, les gris sont de science indifférenciée. Les nœuds bleus sont des physiciens, les rouges des chimistes, les jaunes des industriels. Le diamètre d'un nœud correspond à son degré de connectivité.

Conclusion

Les résultats de la méthode numérique trouvent une répercussion dans l'état des connaissances de l'historien. Au delà d'illustrer des connaissances existantes, un "dialogue" heuristique s'instaure entre historien et analyse numérique. Premièrement, l'**interprétation de relations "surprenantes"** permet de braquer le regard sur un angle mort ou d'ouvrir une voie non explorée. Deuxièmement, le tracé de représentations numériques initié par des questionnements historiques met en évidence des réseaux de nœuds et des clusters inédits, qui peuvent **renouveler les interprétations historiques** ou, au contraire, s'avérer dénuées de sens.

Références

Enguehard, C. and Pantera, L. (1995). Automatic natural acquisition of a terminology. *Journal of quantitative linguistics 2(1)*.
Teissier, P. (2014). Une histoire de la chimie du solide. Synthèses, formes, identités. *Paris, Hermann*.

EGC 2017 - 24/01/2017 - Grenoble