Extraction d'expressions et mise en réseau d'un corpus

Matthieu Quantin[*,**], Benjamin Hervy[*], Florent Laroche[*]

[*]École Centrale de Nantes -LS2N UMR_6004
prenom.nom@ls2n.fr
[**]Université de Nantes, Centre François Viète EA_1161
prenom.nom@univ-nantes.fr

1 Introduction

La méthode proposée s'intéresse à l'analyse de corpus de textes en histoire. L'historien est majoritairement confronté à des textes non-structurés au sens informatique du terme, organisés en corpus. Le corpus est un contenu précis, limité, résultant d'une sélection. Il est connu qualitativement (contextualisé voire lu ou analysé). L'analyse quantitative associée aux connaissances de l'historien se révèle un puissant outil heuristique : elle permet la mise en évidence de nouvelles hypothèses de recherche ou en confirme d'autres déjà établies.

Pour cela il faut faire l'hypothèse d'une analyse centrée sur le contenu du corpus sans orientation *a priori*. Tandis que la pratique montre une utilité pour ce type d'analyse auprès des historiens, l'état de l'art montre l'absence d'intérêt dans le domaine de la gestion des connaissances. La plupart des travaux à l'échelle du corpus intègrent des données extérieures (apprentissage, reconnaissance d'entité nommées) ou des données structurées.

2 Méthode

Le processus se divise en 4 étapes successives, transformant un corpus de documents texte (txt, odt, doc, pdf, ...) en un graphe multiple pondéré de document liés par co-occurrences d'expression-clés.

Gestion du corpus Pour simplifier les calculs le texte du corpus peut être réduit en lemmes avec TreeTagger (Schmid, 1994) et purgé d'une liste de mots vides. Selon le volume et l'hétérogénéité du corpus, il est utile de le subdiviser. Le volume implique des calculs plus long (maximum $5 \cdot 10^6$ mots), une forte hétérogénéité entraîne des résultats moins précis ou du bruit. Nous utilisons une factorisation de matrice non-négative (Berry et Browne, 2005).

Extraction L'implémentation d'un algorithme type C-value inspiré d'ANA (Enguehard, 1993) extrait des expressions complexes sans apprentissage ni données extérieures, grâce à l'identification de motifs autours pivots (de, du, en, au). Cette étape fonctionne en français et anglais. Un post-traitement propose de classifier certaines expressions grâce aux portails et catégories issus de requêtes Wikipedia (Milne et Witten, 2008).

Modération L'utilisateur peut modérer les résultats (liste d'expression et catégories), des indicateurs l'assistent dans l'évaluation de la pertinence des expressions extraites : mesure d'homogénéité inspirée de MED (Bu et al., 2010), présence de verbe, sur-représentation (*weirdness ratio*).

Création et pondération des liens Il s'agit de créer un graphe avec les documents pour nœuds et un lien pondéré entre chaque paire de documents où une co-occurrence d'expression est observée. La pondération inspirée de TF-IDF (Salton, 1983), associé à une sigmoïde (fonction de Gompertz) permet un effet de seuil paramétrable, filtrant les termes les plus génériques. Ce procédé a l'avantage d'être précis et l'inconvénient de créer beaucoup de liens. En réponse, nous créons un *lien unique*, somme de tous les liens de co-occurrence, permettant d'évaluer combien deux documents sont liés dans l'absolu.

3 Conclusion

La combinaison du graphe multiple (liens de co-occurrences) avec celui de *liens uniques* ainsi que la définition de seuils sur la pondération du lien (logique floue) permet de construire des vues du graphe en adaptant le niveau de détails.

Ainsi l'historien peut visualiser un premier graphe léger, puis focaliser l'analyse sur certains points : lien ou cluster inattendu, lien très significatif... En augmentant le niveau de détail (liens de co-occurrences entre certains nœuds, du mieux au moins bien noté) il peut confirmer une hypothèse ou poser de nouvelles questions.

La mise en évidence de signaux faibles (co-occurrences d'expressions complexes) permet une analyse fine s'adressant à un spécialiste. Des inférences "bas-niveau" (calcul statistiques) pondèrent les liens de co-occurrences, laissant à l'historien les inférences "haut niveau" : l'interprétation.

Références

Berry, M. et M. Browne (2005). Email surveillance using non-negative matrix factorization. *Computational & Mathematical Organization Theory 11*(3), 249–264.

Bu, F., X. Y. Zhu, et M. Li (2010). A new multiword expression metric and its applications. *Journal of Computer Science and Technology 26*(1), 3–13.

Enguehard, C. (1993). Acquisition de terminologie à partir de gros corpus. *Informatique & Langue Naturelle*, p.373–384.

Milne, D. et I. H. Witten (2008). Learning to link with Wikipedia. In *Proceeding of the 17th ACM Conference on Information and Knowledge Management (CIKM '08)*, Napa Valley, USA (CA), pp. 509–518. ACM.

Salton, G. (1983). *Introduction to Modern Information Retrieval*. New York, NY, USA : McGraw-Hill, Inc.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing* (4), 44–49.