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).