Wordfish

Depuis quelques années, les outils d'analyse de contenu assisté par ordinateur sont devenus largement disponibles pour les chercheurs afin d'estimer le positionnement des acteurs à partir de textes politiques.[1] Les méthodes fonctionnant avec un dictionnaire représentent une approche. ***Wordscore*** en est une autre(Laver, Benoit et Garry 2003). Quoique prometteuse, car elle constitue la première vraie technique entièrement informatisée, l'utilisation de ***Wordscore*** est sujette à plusieurs limites lorsque vient le temps de produire des séries chronologiques. On retrouve, par exemple, le changement constant du vocabulaire politique au fil du temps et la nécessité pour le chercheur d'identifier correctement les textes de références(Slapin et Proksch 2008).

Empruntant une voie radicalement différente basée sur une modélisation statistique sophistiquée assistée par ordinateur, la méthode ***Wordfish*** permet d'éliminer ces limites. ***Wordfish*** peut produire des estimations tirées de séries chronologiques, ne nécessite pas l'emploi de textes de référence, utilise chaque mot dans chaque document et estime l'importance de chacun de ces mots.

Inspiré de la linguistique, ***Wordfish*** utilise une inférence bayésienne naïve pour représenter le processus par lequel les mots sont générés dans un texte. Un texte est représenté comme un vecteur de fréquences de mots (occurrence) et les mots individuels sont présumés être distribués aléatoirement. La probabilité que chaque mot se retrouve dans le texte est indépendante de la position des autres mots dans le texte. Bien qu'empiriquement faux, le modèle bayésien naïf performe souvent bien dans les tâches de classification (McCallum et Nigam 1998). ***Wordfish*** présume que la fréquence des mots est générée par un processus Poisson. Slapin et Proksch ont choisi cette distribution particulière à cause de la simplicité de son estimation. Un seul paramètre, λ, est à la fois la moyenne et la variance. La forme fonctionnelle du modèle va comme suit :

$$\lambda_{ijt} = \exp(\alpha_{it} + \psi_j + \beta_j * \omega_{it})$$

λ_{jit} est le compte du mot ***j*** dans le manifeste du parti ***i*** au temps ***t***, α est un ensemble d'effets parti-élection fixes, ψ est un ensemble d'effets mots fixes, a un estimé du poids spécifique

[1] Les méthodes CMP, dictionnaire et *Wordscore* sont décrites sur le site internet Poltext.

du mot *j* pour distinguer entre les positions des partis politiques et ω, l'estimé de la position du parti *i* durant l'année électorale *t*. Chaque plateforme est traitée comme une position distincte d'un parti et toutes les positions sont estimées simultanément. Cela signifie qu'il n'y a pas de contraintes temporelles sur le positionnement de la plateforme du parti lors de l'élection *t*. Donc, si un parti utilise des mots avec une fréquence relative identique au fil du temps, sa position va demeurer la même. Le mouvement des partis est causé par le changement dans la fréquence des mots, non par un changement de signification de ceux-ci.

La prochaine étape est l'estimation du côté droit de l'équation de régression en utilisant l'algorithme EM, pour ***expectation maximization*** (voir Dempster, Laird et Rubin 1977). L'algorithme EM est un processus itératif pour calculer les estimés de vraisemblance maximale pour les variables latentes. D'abord, les valeurs initiales sont calculées. Ensuite, les paramètres partisans (α et ω) sont estimés. Troisièmement, les paramètres des mots (ψ et β) sont estimés. Quatrièmement, le log-vraisemblance est calculé (somme des vraisemblances des mots pris individuellement). Finalement, les étapes 2 à 4 sont répétées jusqu'à ce qu'il y ait convergence.

Pour calculer un intervalle de confiance à 95 %, une ***bootstrap*** paramétrique est nécessaire. La procédure est une méthode de rééchantillonage assistée par ordinateur. En employant l'algorithme EM, λ_{jit} est calculé pour chaque cellule dans la base de données qui contient les fréquences des mots. 500 nouvelles bases sont ainsi générées en pigeant aléatoirement dans une distribution Poisson avec le paramètre λ_{jit} pour chaque cellule. En utilisant les estimations de la vraisemblance maximale comme valeur initiale, l'algorithme est de nouveau employé sur chacune de ces nouvelles bases de données et 500 nouvelles positions sont estimées pour les partis politiques. Avec les quantiles 0,025 et 0,0975, un intervalle de confiance peut ainsi être estimé. Avec une ***bootstrap*** paramétrique, l'intervalle de confiance diminue au fur et à mesure que le nombre de mots augmente.

C'est l'architecture mathématique de ***Wordfish***. Alors que ***Wordscore*** fonctionne avec ***STATA***, ***Wordfish*** est développé en tant qu'application informatique programmée avec le langage statistique ***R***.[2] Le programme nécessite une matrice de la fréquence des mots qui constitue l'input. De telles matrices

[2] Pour plus d'information, voir le site The R Project for Statistical Computing.

peuvent être obtenues avec des programmes conviviaux tels que *Yoshikoder* et *jfreq*.[3] Le programme utilise la base de données des fréquences de mots, génère les valeurs de départ et emploie par la suite l'algorithme. L'opération produit des estimations du positionnement des partis politiques ainsi que des valeurs pour la pondération des mots et l'effet fixé des mots. En faisant une *bootstrap* paramétrique, il est possible de générer des intervalles de confiance. Cependant, une telle procédure peut prendre plusieurs jours à compléter.

À l'instar de *Wordscore*, *Wordfish* suppose que la signification des mots demeure la même au fil du temps. Le changement du positionnement politique se traduit par une modification de la fréquence relative des mots et non par un changement de signification de ceux-ci. *Wordfish* suppose aussi que la dimension politique extraite des textes politiques reflète le contenu de ces textes. Si les chercheurs veulent des estimations du positionnement des partis politiques à propos du thème de l'éducation, ils devraient sélectionner des textes dont le sujet est spécifiquement l'éducation. Cela signifie identifier de tels textes. Dans le cas de pays ou de langues étrangers, cette étape peut être problématique. Finalement, *Worfish*, tout comme *Wordscore* produisent des estimations sur une seule dimension politique à la fois. Pour analyser plus d'une dimension, il est recommandé de définir ex ante les dimensions à étudier et utiliser des documents qui traitent de sujet ayant rapport au contenu de ces dimensions.

Références

Dempster, A. P., N. M. Laird et D. B. Rubin (1977). « Maximum likelihood from incomplete data via the EM algorithm. » Journal of the Royal Statistical Society, Series B (Methodological) **39**(1): 1-38.

Laver, M., K. Benoit et J. Garry (2003). « Extracting Policy Positions from Political Texts Using Words as Data. » American Political Science Review **97**(2): 311-331.

McCallum, A. et K. Nigam (1998). A Comparison of Event Models for Naive Bayes Text Classification. AAAI-98 Workshop on Learning for Text Categorization.

Slapin, J. B. et S.-O. Proksch (2008). « A Scaling Model for Estimating Time-Series Party Positions from Texts. » American Journal of Political Science **52**(3): 705-722.

[3] Les deux programmes sont disponibles au https://people.iq.harvard.edu/~wlowe/Software.html.