Modèles domain-topic avec apprentissage par transfert

Luis Daniel Medina Zuluaga*

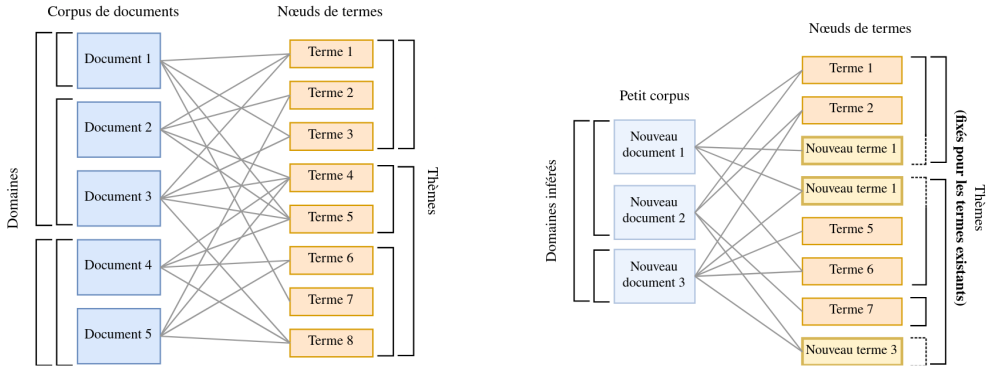
*LISIS, Université Gustave Eiffel, Champs-sur-Marne, France
luis-daniel.medina-zuluaga@univ-eiffel.fr

Le *topic modeling* est un outil largement utilisé pour extraire des informations d'un ensemble de documents, en découvrant les thèmes abstraites qui les composent. Ces thèmes consistent en des groupes de termes inférés du contenu textuel des documents. Les documents peuvent alors être caractérisés en fonction de la prévalence des termes associés à chaque thème. Une nouvelle approche de topic modeling basée sur la détection des communautés dans les réseaux a été récemment introduite par Gerlach et al. (2018). Dans cette approche, les documents sont modélisés comme un graphe document-terme biparti, ensuite un modèle à blocs stochastiques imbriqués est ajusté à ce graphe, produisant une clusterisation hiérarchique des documents et des termes. Abdo et al. (2021) utilisent la même approche pour développer les modèles *domain-topic*, en introduisant un ensemble de mesures et d'interfaces qui facilitent l'étude des hiérarchies de documents et de termes, et montrent comment regrouper les métadonnées associées en fonction des clusters de documents obtenus. Ils appliquent la méthodologie pour étudier l'évolution thématique du domaine de l'oncologie, en s'appuyant sur les clusters thématiques, temporels et géographiques inférés.

Pour produire un modèle domain-topic, le corpus de documents est représenté sous forme d'un graphe d'incidence, où chaque document est relié à ses termes. Le clustering hSBM, tel qu'introduit par Peixoto (2019), est appliqué à ce graphe, avec la restriction que les documents et les termes ne peuvent pas être regroupés. Cela donne une hiérarchie de groupes de termes, les thèmes, et une hiérarchie de groupes de documents, les domaines.

Bien qu'utile dans de nombreux cas, l'approche repose sur les informations disponibles dans un corpus pour identifier des patrons d'utilisation des mots. Cela pose un problème lorsqu'il s'agit de traiter des petits corpus ou lorsqu'il existe des connaissances préalables sur leur organisation thématique. Nous proposons ici une stratégie de réutilisation d'une structure thématique existante, soit déduite d'un ensemble plus large de documents, soit fournie de l'extérieur, pour informer l'inférence du modèle domain-topic sur un ensemble plus restreint de documents, thématiquement proches de l'ensemble plus large. Cette stratégie est illustrée dans la figure 1.

Dans nos expérimentations, nous utilisons plusieurs datasets différents : des articles Plos ONE de 2011 ; des articles de presse de *Reuters financial newswire* (Reuters-21578, Distribution 1.0) et un ensemble d'articles de recherche sur les pesticides provenant de Web of Science. Nous ajustons des modèles domain-topic à ces datasets. Ensuite, nous prenons des sous-ensembles de documents de chaque dataset comme exemples de datasets plus petits. Pour chacun de ces sous-ensembles, nous produisons un clustering non informé, et un clustering informé par la hiérarchie de topiques du dataset entier. Nous comparons les partitions obtenues, à l'aide de mesures de variation de l'information et d'information mutuelle, aux affectations



(a) Modèle domain-topic (b) Réutilisation d'une hiérarchie de thèmes

FIG. 1: La structure thématique d'un modèle existant (a) peut être utilisée pour améliorer l'adaptation des modèles domain-topic sur d'autres ensembles de documents (b). Nous construisons le graphe pour le petit corpus (b), nous identifions ensuite les termes qui appartiennent au même topic dans la structure thématique préexistante (a), et les regroupons en conséquence. Nous fixons ces regroupements afin d'assurer qu'ils ne se divisent pas, ne fusionnent pas et ne forment pas de nouveaux topics. Des hiérarchies de documents et de termes se forment, tout en préservant la structure thématique préalable (b). Les termes non présents dans la hiérarchie de topics peuvent soit former des nouveaux topics, soit rejoindre les existants, tandis que les documents sont regroupés librement.

originales des documents dans le modèle domain-topic préalable. Les résultats montrent l'effet souhaité de transfert : pour chaque sous-ensemble de documents, la variation d'information est plus faible et l'information mutuelle est plus élevée pour le clustering informé, ce qui indique que les partitions obtenues avec la stratégie proposée ressemblent davantage aux affectations des documents dans le modèle domain-topic initial.

Une autre application de cette approche consiste à informer le clustering non pas avec une hiérarchie de termes calculée, mais avec un regroupement de termes fourni par des experts du domaine, souvent appelé *dictionnaire* dans la recherche qualitative. Par exemple, pour emprunter des associations à des ontologies officielles ou pour affiner manuellement un modèle déjà ajusté, ce qui permettrait d'incorporer des connaissances qualitatives dans le modèle.

Références

Abdo, A. H., J.-P. Cointet, P. Bourret, et A. Cambrosio (2021). Domain-topic models with chained dimensions: Charting an emergent domain of a major oncology conference. *Journal of the Association for Information Science and Technology 73*(7), 992–1011.

Gerlach, M., T. P. Peixoto, et E. G. Altmann (2018). A network approach to topic models. *Science Advances 4*(7), eaaq1360.

Peixoto, T. P. (2019). Bayesian stochastic blockmodeling. In P. Doreian, V. Batagelj, et A. Ferligoj (Eds.), *Advances in Network Clustering and Blockmodeling*, pp. 289–332. Wiley.