Analyse non supervisée de corpus de documents pour la recherche de sujets : une nouvelle approche basée sur le clustering et la maximisation des traits

Jean-Charles Lamirel

Analyse non supervisée de corpus de documents pour la recherche de sujets : une nouvelle approche basée sur le clustering et la maximisation des traits

Jean-Charles Lamirel[1]
j.lamiel@unistra.fr

[1] IUT Robert Schuman, Université de Strasbourg
Equipe Synalp, LORIA, Université de Lorraine

Thèmes – *Informatique - Mathématiques – Statistiques - Langues*

Résumé – *Le présent article s'intéresse à l'extraction non supervisée des sujets portés par les collections de documents qui est un thème du front de recherche. Il présente une nouvelle méthode d'extraction de sujets qui combine le clustering avec la métrique de maximisation des traits. Le comportement de cette méthode est comparé avec celui de la méthode de référence LDA, habituellement utilisée pour cette tâche. Nous utilisons comme données expérimentales de grandes collections de publications scientifiques et nous mesurons la qualité des sujets extraits en utilisant le score usuel de cohérence. Nos expériences montrent une amélioration supérieure à 50% de la qualité des sujets obtenus par rapport à la méthode de référence, permettant en conséquence une bien meilleure interprétation de ces derniers par les experts humains.*

.

Mots-Clés – *Apprentissage non supervisé, Extraction de sujets, Clustering, LDA*

1 Introduction

Les méthodes d'extraction automatique de sujets jouent un rôle primordial dans l'analyse des grandes collections de documents. Elles permettent en effet de mener à bien de nombreuses tâches stratégiques, comme la veille technologique, si elles sont appliquées sur des collections de brevets, ou encore, l'évaluation des activités scientifiques, si les données considérées sont des articles scientifiques. Ces méthodes supportent l'analyse diachronique et facilitent également les tâches de résumé automatique en général. De nombreuses méthodes ont été proposées pour l'extraction de sujets. Ce sont généralement des méthodes non supervisées. La plus populaire d'entre elle est la méthode LDA (Latent Dirichlet Analysis) [1] qui comprend de nombreuses variantes. L'amélioration des méthodes de modélisation thématique reste cependant une préoccupation majeure, notamment du fait que des méthodes de type LDA dépendent d'hyperparamètres, ce qui les rend également dépendantes des corpus étudiés. LDA est également basée sur des hypothèses discutables, comme le fait que les documents sont des combinaisons de sujets et que les sujets extraits des documents doivent nécessairement être indépendants les uns des autres, ce qui n'est pas le cas lorsque les documents contiennent des sujets complexes qui sont le plus souvent en interaction.

Nous proposons ci-après une approche alternative d'extraction de sujets sans paramètres basée sur le clustering neuronal et la maximisation des traits. Nous présentons brièvement ces deux techniques avant de passer à la description de notre protocole expérimental. Puis, nous exposons nos résultats d'expérience et nos conclusions.

2 Méthodologie

Le clustering est une technique connue de regroupement non supervisé des données, qui comprend de nombreuses variantes. Parmi celles-ci, le clustering neuronal est une méthode de regroupement direct basée sur l'apprentissage hebbien qui présente l'avantage d'être moins sensible aux conditions initiales ainsi qu'aux données marginales que les méthodes de clustering usuelles telles que k-means. Dans ce travail, nous utilisons la sous-variante GNG (Growing Neural Gas) [2] du clustering neuronal, une méthode à topologie libre, pour constituer des groupes de documents susceptibles de porter les mêmes sujets.

Nous calculons plusieurs modèles GNG avec un nombre variable de clusters et nous appliquons dans un second temps une méthode originale de recherche de modèle de clustering optimal [3] qui permet de déterminer le nombre de clusters adéquat pour les données analysées.

L'identification des caractéristiques du sujet porté par chacun des clusters est opérée en utilisant la métrique de maximisation des traits (FMax) [4]. Cette métrique représente une alternative aux mesures usuelles de distance telles que la distance euclidienne, la corrélation cosinus ou le Chi2. Elle permet notamment d'évaluer des similarités entre des données dans des conditions où les distances classiques ne s'avèrent plus discriminantes, comme dans le cas du traitement des données fortement multidimensionnelles et éparses, cas typique des données textuelles, notamment lorsqu'elles sont représentées en sacs de mots. Dans les textes, les traits représentent des mots et FMax est basé sur l'estimation de la F-mesure de trait en tant que moyenne harmonique (1) du rappel de trait, qui se concentre sur le pouvoir de discrimination des mots concernant des données groupées ; et (2) de la prédominance de trait, qui se concentre sur la capacité de généralisation des mots concernant les mêmes données.

Cette métrique possède également des capacités complémentaires de pondération et de sélection de variables et ne nécessite pas l'utilisation de paramètres. Elle s'est avérée très utile dans de nombreuses tâches d'apprentissage automatique, y compris dans la fouille de graphes [5].

L'ensemble des étapes mentionnées constitue une nouvelle méthode globale que nous nommons CFMf (neural Clustering and Feature Maximization with feature F-measure).

3 Protocole expérimental

Nous avons appliqué la méthode CFMf avec succès sur plusieurs corpora de taille importante dont un corpus relatif aux données de recherche en sciences de la science en Chine sur 40 ans (Nb docs=2790) [6]. Dans cet article, nous discutons plus spécifiquement de la comparaison des performances de notre approche avec celle de LDA, nous appliquons ces deux méthodes à un corpus de référence d'articles de philosophie des sciences en texte intégral (Nb docs =16 917). Le corpus a été nettoyé et prétraité de manière standard (les textes en langue étrangère ont été traduits mécaniquement en anglais). Seuls les noms, les verbes, les adverbes et les adjectifs ont été conservés après l'étiquetage des parties du discours et la lemmatisation (paquet TreeTagger [7] (Schmid, 1994) avec les ensembles d'étiquettes Penn TreeBank [8] (Marcus et al., 1993)) et les mots apparaissant dans moins de 50 phrases dans le corpus ont été supprimés. Tous les documents ont ensuite été vectorisés, ce qui a permis d'obtenir une matrice terme-document. Pour pouvoir établir une comparaison entre les méthodes, le nombre de sujets a été fixé à 25, valeur choisie à la suite d'expériences précédentes menées avec la méthode LDA [9].

L'évaluation des résultats se base sur une des mesures de référence pour l'extraction de sujets, à savoir la mesure de cohérence C_v de Roêder [10]. C_v compte les

cooccurrences d'un certain nombre de mots principaux (généralement 10 à 20) dans une fenêtre glissante (généralement de taille 110); les cooccurrences apparaissant dans la fenêtre sont utilisées pour calculer l'information mutuelle entre les mots principaux, ce qui donne des vecteurs pour chacun d'entre eux ; la moyenne arithmétique des similitudes cosinus entre chaque vecteur de mot principal et la somme de tous les vecteurs de mots principaux est ensuite calculée.

4 Résultats

Les résultats prometteurs (figure 1) montrent des améliorations significatives des mesures de performance quantitatives clés telles que la cohérence, indépendamment du nombre de sujets, jusqu'à 55% meilleure que dans le cas de LDA. Les comparaisons qualitatives montrent également clairement des améliorations dans la cohérence des sujets et leur interprétabilité à la lumière des connaissances des experts.

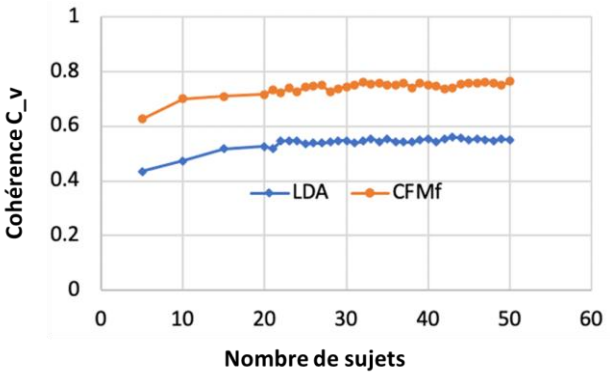

Figure 1 – Cohérence C_v en fonction du nombre de sujets.

5 Conclusion

Nous avons présenté une nouvelle approche prometteuse pour l'extraction de sujets basé sur une combinaison de regroupement neuronal, de maximisation des traits et de mesure F1 : CFMf. Nous avons également partagé les résultats de nos première expériences comparatives. Cependant, malgré le potentiel évident de CFMf pour la réalisation d'études à grande échelle, telles que les études scientifiques [6], plusieurs adaptations et expériences supplémentaires sont encore possibles. Nous chercherons en particulier à savoir si la LDA peut être améliorée en utilisant des composants-clés de la présente méthode, notamment en exploitant la maximisation des traits FMax pour pondérer les termes dans les sujets de la LDA. Nous examinerons la comparaison de notre modèle avec des approches de modélisation de sujets basées sur enchâssements de mots. Nous envisageons également d'étudier la possibilité de choisir un nombre optimal k de sujets [3].

6 Remerciements

Les auteurs expriment leur gratitude à Christophe Malaterre, Francis Lareau (UQAM, Montréal) et Pascal Cuxac (INIST/CNRS) pour leur contribution indirecte à ce travail.

Références

[1] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3(Jan), 993–1022.

[2] Fritzke, B. (1994). A growing neural gas network learns topologies. Advances in Neural Information Processing Systems, 7.

[3] Dugué, N., Lamirel, J.-C., & Chen, Y. (2021). Evaluating clustering quality using features salience: A promising approach. Neural Computing and Applications. https://doi.org/10.1007/s00521-021-05942-7

[4] Lamirel, J.-C., Cuxac, P., Chivukula, A. S., & Hajlaoui, K. (2015). Optimizing text classification through efficient feature selection based on quality metric. Journal of Intelligent Information Systems, 45(3), 379–396. https://doi.org/10.1007/s10844-014-0317-4

[5] Prouteau, T., Connes, V., Dugué, N., Perez, A., Lamirel, J.-C., Camelin, N., & Meignier, S. (2021). SINr: Fast Computing of Sparse Interpretable Node Representations is not a Sin! In P. H. Abreu, P. P. Rodrigues, A. Fernández, & J. Gama (Eds.), Advances in Intelligent Data Analysis XIX (Vol. 12695, pp. 325–337). Springer International Publishing.

[6] Lamirel, J.-C., Chen, Y., Cuxac, P., Al Shehabi, S., Dugué, N., & Liu, Z. (2020). An overview of the history of Science of Science in China based on the use of bibliographic and citation data: A new method of analysis based on clustering with feature maximization and contrast graphs. Scientometrics, 125(3), 2971–2999. https://doi.org/10.1007/s11192-020-03503-8

[7] Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. Proceedings of International Conference on New Methods in Language Processing, 44–49.

[8] Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics, 19(2), 313–330. https://doi.org/10.21236/ADA273556

[9] Malaterre, C., & Lareau, F. (2022). The early days of contemporary philosophy of science: Novel insights from machine translation and topic-modeling of non-parallel multilingual corpora. Synthese, 200(3), 242. https://doi.org/10.1007/s11229-022-03722-x

[10] Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15, 399–408. https://doi.org/10.1145/2684822.2685324