CentraleSupélec

Filière Recherche

Étude de cas

Recommandation d'articles d'actualité

Étudiants :
Arthur Gallois
Nathan Chalumeau
Dimitri Martin
Ali Ramlaoui

Encadrante :
Laure Normand
Bich-Liên Doan

19 septembre 2023

Abstract

L'apparition d'Internet dans les années 90 a transformé la manière dont les utilisateurs accèdent aux articles d'actualité : il est devenu simple d'accéder massivement à des articles de toutes natures. Dans ce contexte, la question de l'amélioration de l'expérience utilisateur et de la personnalisation de leur parcours de lecture a rapidement émergé. Des filtres de recommandation ont alors été introduits comme discuté dans l'étude pionnière de Claypool et al. (1999) [1].

Les méthodes habituelles de recommandation (filtrage collaboratif) bien que favorisant la rétention d'utilisateurs posent de nombreux problèmes comme la quantité de données ainsi que la puissance de calcul nécessaire. Elles créent aussi des problèmes de bulles de filtres [2].

Afin d'éviter ces problèmes identifiés dans les algorithmes classiques, Julien Hay propose dans sa thèse une nouvelle méthode de filtrage sur le contenu en explorant la prise en compte du style d'écriture des articles de presse. Notre article se propose de réimplémenter partiellement des méthodes utilisées par Julien Hay et de comparer ses résultats avec les nôtres. Il est important de noter que ce projet a été réalisé avec un temps de moins d'une semaine et une puissance de calcul faible. Nous avons néanmoins réussi à obtenir des résultats favorables à l'emploi de la représentation de style dans les algorithmes de recommandation et encourageant une recherche plus poussée dans ce domaine.

Table des matières

1 Introduction

1.1 Algorithmes de recommandations et méthodes d'évaluation

Avec l'arrivée d'Internet dans les années 90, une quantité massive d'articles d'actualité très variés s'est retrouvée accessible très facilement à toutes les personnes. Plus tard, avec l'arrivée de réseaux sociaux et de sites aggrégateurs, l'utilisation de filtres de recommandation devient nécessaire, évoquée pour la première fois par [1] avec un objectif de personnaliser et d'améliorer l'expérience utilisateur. Ces filtres permettent de donner rapidement accès à un utilisateur aux articles susceptibles de l'intéresser et ainsi de lui épargner une recherche fastidieuse. Actuellement, les plus grosses plateformes de flux de contenu (Facebook, YouTube, etc.) utilisent des algorithmes collaboratifs, s'appuyant sur leurs bases d'utilisateurs et d'interaction massives [2]. Ces méthodes sont très efficaces pour la rétention d'utilisateurs, mais montrent des limites tant par la quantité de données et la puissance de calcul nécessaire que par la génération de « bulles de filtre », qui empêchent les utilisateurs de tomber sur du contenu nouveau pour eux [2]. Le développement de nouvelles méthodes et critères d'évaluations se justifie alors pour pallier ces différents problèmes.

Les principaux algorithmes de recommandation basés sur le texte ont traditionnellement fait appel à des outils de traitement du langage naturel consistant à observer et comparer les fréquences d'apparition des mots entre les documents historiques qui ont été lus et appréciés par les utilisateurs et les nouveaux documents afin d'attribuer un score. L'algorithme ***BM25*** proposé par *Robertson et Walker* [4] est le plus répandu, car il permet d'obtenir une bonne précision sur les jeux de données utilisés pour le *benchmark* [2]. Cependant, de nouvelles méthodes émergent [3] basées par exemple sur des réseaux de neurones et des architectures composées de *transformers* et de couches d'attention et permettent la représentation vectorielle du texte de manière plus simple grâce à l'augmentation des puissances de calcul. L'idée est alors de mettre en avant l'ensemble de ces outils afin d'augmenter à la fois la précision des recommandations d'articles tout en évitant autant que possible les bulles de filtre.

Il existe 3 méthodes d'évaluation différentes :

— Les méthodes Offline
— Les méthodes Online
— Les méthodes Qualitatives

Les méthodes offline consistent en l'utilisation d'un jeu de données figé pour évaluer le modèle : on étudiera sa capacité à prédire des interactions ayant déjà eu lieu. Les méthodes online s'appuient sur un feedback réel de l'utilisateur : les articles vont être recommandés directement à l'utilisateur et son interaction avec vont être étudiée. Enfin, les méthodes d'évaluation qualitative se basent sur de nouveaux critères comme la diversité, la nouveauté, la sérendipité (articles surprenants et nouveaux) ou la détection de bulles de filtres.

1.2 Analyse de style

La thèse de Julien Hay [2] se concentre sur la recommandation d'articles d'actualité en prenant en compte la représentation du style écrit de l'auteur. L'idée est de mesurer l'importance du style dans la manière dont les utilisateurs apprécient et partagent le contenu qu'ils lisent. Pour ce faire, de multiples modèles et méthodes ont été développés afin de représenter le texte dans un espace permettant de distinguer les styles utilisés par l'auteur du texte. L'entraînement de tels modèles est basé sur une l'hypothèse de généralisation du style selon laquelle il est possible de projeter le style sur un espace constitué de plusieurs auteurs et que le style d'un nouvel auteur pourra être obtenu en combinant les styles présents dans l'ensemble utilisés pour l'entraînement. Un des objectifs de la thèse est alors de valider cette hypothèse en comparant les précisions des modèles de recommandations d'articles d'actualité utilisant la représentation du style aux modèles plus sémantiques basés par exemple sur la fréquence d'apparition des mots. Deux principaux modèles ont été développés pour apprendre à représenter le style : le modèle ***SNA*** (*Stylometric Neural Attention*) et ***DBERT-ft***. Alors que le premier modèle consiste à entraîner un réseau de neurones à identifier les auteurs de différents textes et à ensuite extraire les représentations apprises par une des couches du réseau, la seconde méthode est basée sur un *fine-tuning* du modèle ***DBERT*** utilisé pour la représentation de texte sur un corpus adapté à l'apprentissage du style.

2 Résultats

Nous avons tenté d'implémenter les idées principales de la thèse de Julien Hay [2] afin de vérifier les résultats qu'il est possible d'obtenir en utilisant les méthodes présentées ci-dessus et d'obtenir une implémentation minimaliste des algorithmes. Du fait de la contrainte de temps et de puissance de calcul, les tests ont été réalisés en utilisant une version réduite du jeu de données *Twinews* [2] permettant de s'assurer qu'il est possible de faire tourner nos nouvelles implémentations et d'obtenir des résultats cohérents. Nous proposons le code contenant les détails de l'implémentation adaptée de *Twinews* afin de pouvoir rapidement effectuer des tests minimes. Nous avons remarqué que l'utilisation du modèle ***DBERT-ft*** était en accord dans notre cas avec les résultats de la thèse et qu'il est possible d'obtenir de meilleurs résultats en utilisant l'algorithme ***Reswhy*** [2] sur ***BM25*** et ***DBERT-ft*** confirmant la complémentarité dans l'information apprise par les modèles basés sur la représentation du style.

Le jeu de données utilisé pour obtenir les résultats est la première version du *Twinews* qui consiste en 25.000 articles d'actualités partagés par 500 utilisateurs sur *Twitter*. L'idée est donc de garder, pour chaque utilisateur, un ensemble d'articles dits "historiques" qui vont constituer les articles que l'utilisateur a aimé et sur lequel on va baser les goûts de l'utilisateur et des articles sur lesquels on va tester l'algorithme afin de vérifier les scores obtenus.

BM25 se base sur la représentation sac de mots, c'est à dire qu'il ne s'appuie pas sur la structure de la phrases mais seulement sur le nombre d'occurrence des mots. Cet algorithme se base donc sur des critères tel que l'IDF (Inverse Document Frequency), la longueur des documents ou des hyperparamètres. L'IDF correspond à l'inverse de la fréquence d'apparition d'un terme dans l'historique d'un utilisateur. Une des étapes délicates de cet algorithme est la lemmatisation.

DBERT-ft est une version affinée par Julien Hay [2] du modèle ***DBERT*** développé par Google. Pour chaque utilisateur, nous allons calculer la représentation vectorielle de ces textes historiques (données d'entraînement), puis nous allons regrouper les articles que l'utilisateur a partagé mais n'apparaissant pas dans les textes historiques avec des articles aléatoires non partagé par l'utilisateur pour former 100 articles différents que nous allons également vectoriser. L'idée est ensuite de calculer la similarité cosinus entre chacun de ces 100 articles et l'ensemble des articles historiques de l'utilisateur pour obtenir un score de similarité qui correspond au score de recommandation attribué à l'article. Cela génère ainsi un ordonnancement des 100 articles que nous pouvons utiliser pour calculer les métriques telles que le score *nDCG* et le *MRR*. .

	nDCG	**MRR**
BM25	0.45	0.27
DBert-ft	0.53	0.46
Reswhy	0.55	0.45

FIGURE 1 – Comparaison des résultats entre BM25, DBert-ft et Reswhy

Nous utilisons deux métriques pour comparer les deux modèles :
D'une part *nDCG*, une métrique très utilisée pour comparer deux ordonnancements. Elles est calculée comme ci-dessous :

Pour tout ordonnancement de n articles, on note rel le vecteur de taille n tel que $rel_i = 1$ si et seulement si l'utilisateur a aimé l'article i et sinon $rel = 0$ et tel que les articles $1, ..., n$ sont classés dans l'ordre décroissant du score attribué par le modèle utilisé. On donne alors :

$$nDCG(rel) = \frac{DCG(rel}{iDCG(rel)}$$

Avec le score DCG :

$$DCG(rel) = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

et le score DCG idéal :

$$iDCG(rel) = \sum_{i=1}^{\text{sum}(rel)} \frac{1}{\log_2(i+1)}$$

D'autre part, la métrique *MRR* (Mean Reciprocal Rank) se base sur la position de l'élément le plus pertinent dans l'ordonnancement généré par l'algorithme. Pour chaque ordonnancement généré, il s'agit alors de chercher la position du premier élément pertinent de l'utilisateur dans cet ordonnancement. Le *MRR* est ensuite obtenu en effectuant la moyenne de l'inverse de cette position sur l'ensemble des utilisateurs considérés. R' est l'ensemble des ordonnancement générés pour tous les utilisateurs, T' l'ensemble des éléments pertinents pour chaque utilisateur, et *minrank* la fonction renvoyant la position du premier élément pertinent dans l'ordonnancement considéré.

$$\text{MRR(R',T')} = \sum_{i=1}^{|R'|} \frac{1}{\text{minrank}(R'_i, T'_i)}$$

Une des raisons de l'écart entre le score MRR attendu et le score obtenue provient de la taille du dataset utilisé et de la variance élevé du score MRR.

3 Conclusion

Ainsi, cet article propose une réimplémentation d'une partie de la thèse de Julien Hay [2]. Les résultats obtenus prouvent que la méthode présentée par Julien Hay est plus performantes que les méthodes précédentes dans certaines situations. Nous avons obtenu des résultats satisfaisants sur une base de donnée de taille réduite, cohérents avec ceux obtenus par Julien Hay sur une base de plus grande taille. Un travail sur une plus longue durée avec une base de donnée plus large et plus de puissance de calcul aurait pu permettre de valider les résultats plus en profondeur. Il serait également très intéressant de pouvoir comparer le lien entre les différents scores obtenus et les avis des utilisateurs sur les recommandations proposées dans un environnement de test réel de type *online*. Ce modèle nous paraît être une bonne solution pour la génération en mode offline de recommandations d'articles. De plus, cette méthode évite, dans l'idéal et en partie, les risques de bulles de filtres et permet de concevoir un filtre favorisant la sérendipité plus simplement que les filtres sur le contenu conventionnels.

Références

[1] Mark Claypool et al. "Combining Content-Based and Collaborative Filters in an Online Newspaper. SIGIR'99 Workshop on Recommender Systems : Algorithms and Evaluation. Berkeley, CA". In : *Berkeley, CA* (1999).

[2] Julien Hay. "Apprentissage de la représentation du style écrit, application à la recommandation d'articles d'actualité". Theses. Université Paris-Saclay, mars 2021. URL : `https://theses.hal.science/tel-03420487`.

[3] Marius Kaminskas et Derek Bridge. "Diversity, Serendipity, Novelty, and Coverage : A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems". en. In : *ACM Transactions on Interactive Intelligent Systems* 7.1 (mars 2017), p. 1-42. ISSN : 2160-6455, 2160-6463. DOI : `10.1145/2926720`. URL : `https://dl.acm.org/doi/10.1145/2926720`.

[4] Stephen Robertson et al. "Okapi at TREC-3." In : 1994.