Une approche sémantique hybride pour la recommandation des articles d'actualité à large échelle

Hemza Ficel,
Mohamed Ramzi Haddad,
Hajer Baazaoui Zghal

Laboratoire Riadi , École Nationale des Sciences de l'Informatique,
Université de la Manouba, La Manouba 2010, Tunisie
hemza.ficel@ensi-uma.tn
haddad.medramzi@gmail.com
hajer.baazaouizghal@riadi.rnu.tn

Résumé. Les portails d'actualités en ligne produisent un flux d'information ayant un volume et une vélocité importants. Dans ce contexte, il devient plus difficile de proposer en temps réel des recommandations dynamiques adaptées aux intérêts de chaque utilisateur. Dans cet article, nous présentons une approche hybride pour la recommandation des articles d'actualité reposant sur l'analyse sémantique du contenu disponible. L'approche est basée sur l'hybridation de plusieurs approches personnalisées et non personnalisées pour remédier au problème de démarrage à froid. L'expérimentation de notre approche dans un environnement à large échelle et à fortes contraintes temps réel dans le cadre du challenge NEWSREEL a permis d'évaluer la qualité de ses recommandations et de confirmer l'apport de la sémantique dans le processus de recommandation.

1 Introduction

C'est à travers des messages du type « La rédaction vous conseille », « Lire aussi », « Sur le même sujet », « Vous pourriez également être intéressé par », que les portails d'actualités en ligne proposent aux utilisateurs des sélections personnalisées de contenus pouvant les intéresser afin de les assister dans leur exploration des articles d'actualités. Dans ce domaine d'application, les systèmes de recommandation font face à un flux abondant et continu d'actualités qu'il faut collecter, analyser et mettre à disposition des utilisateurs selon leurs préférences et dans les plus brefs délais. Par ailleurs, la recommandation des articles d'actualité diffère des domaines d'application traditionnels des systèmes de recommandation qui se caractérisent par une parcimonie de données reflétant les comportements des utilisateurs à cause de l'inaccessibilité de certains produits (produits non disponibles ou ayant un coût d'acquisition élevé). En effet, les articles d'actualité étant gratuits et accessibles à tous les utilisateurs, les approches de recommandation à adopter doivent être capables de prendre en considération un flux abondant d'interactions entre les utilisateurs et le contenu. À titre d'exemple, sur la plateforme de

recommandation de Plista [1], le nombre d'événements observés par seconde peut dépasser les 5000 et le temps de réponse maximal pour fournir une recommandation est limité à 100ms. Dans ce contexte, il est nécessaire qu'un système de recommandation adopte une approche temps-réel, capable de monter en charge et proposant des recommandations dynamiques qui s'adaptent rapidement aux comportements observés chez chaque utilisateur. De plus, vu la nature non structurée du contenu traité, il est judicieux d'adopter des approches sémantiques capables d'extraire les concepts et les thématiques invoquées dans les articles d'actualité afin de mieux cerner les préférences et les intérêts des utilisateurs.

Les contraintes fortes de la problématique de la recommandation continue, temps réel et à large échelle des articles d'actualité ont conduit à une large adoption d'approches non personnalisées et non sémantiques fondées sur critères de popularité et de nouveauté (Lommatzsch et al., 2016) pour leurs faibles complexités. Plusieurs approches récentes se focalisent sur les technologies de traitement des données massives (p.ex. Apache Spark [2] et Flink [3]) afin d'assurer le passage à l'échelle (Domann et al., 2016) sans prendre en considération l'aspect sémantique du contenu ni les préférences personnelles de chaque utilisateur. D'autres travaux ont intégré les connaissances extraites à partir du contenu dans le processus de recommandation pour mieux assimiler les facteurs et les thématiques qui influencent les attitudes du consommateur et améliorer ainsi la qualité des suggestions (Capelle et al., 2013). Néanmoins, ces travaux ne tiennent pas compte des contraintes liées à la recommandation en ligne en temps réel. En effet, les approches proposées sont généralement évaluées hors-ligne sur des jeux de données statiques et limités ne reflétant pas les conditions réelles de ce domaine d'application nécessitant une gestion continue du flux de données entrant et une capacité à générer en temps réel et à large échelle des recommandations personnalisées et dynamiques.

C'est dans ce contexte que s'inscrit ce travail dont l'objectif est de proposer une approche de recommandation personnalisée d'articles d'actualité traitant l'aspect sémantique du contenu et respectant les contraintes temps réel et de passage à l'échelle du domaine d'application. Notre approche est basée sur l'hybridation de plusieurs approches personnalisées et non personnalisées afin d'éviter le problème de démarrage à froid. Pour valider l'intérêt de la proposition, notre système de recommandation a été évalué pendant le challenge NEWSREEL 2017 dont l'objectif est de faire la recommandation d'articles d'actualité en temps réel et à large échelle pour un ensemble de magazines et de journaux en ligne.

Le reste de cet article est organisé comme suit. La section 2 détaille l'approche proposée ainsi que les composantes du système développé. La section 3 présente les différentes expérimentations menées ainsi que les résultats obtenus. Enfin, nous concluons cet article avec un résumé de la contribution et des travaux futurs.

2 Notre système de recommandation

La figure 1 présente l'architecture du système de recommandation intégrant l'approche proposée et qui inclut un composant pour le traitement des données (A) et un composant pour le filtrage et la recommandation des articles d'actualité (B). Dans cette section, nous détaillons le système proposé, ses composants ainsi que les techniques utilisées.

1. https ://www.plista.com/
2. https ://spark.apache.org/
3. https ://flink.apache.org/

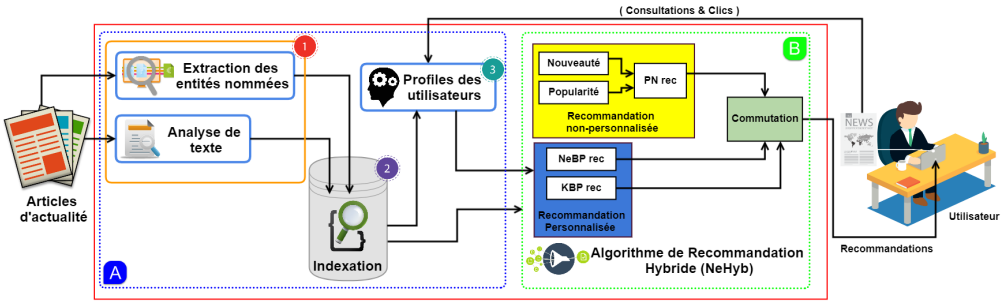

FIG. 1 – *Architecture du système de recommandation proposé.*

2.1 Le traitement des données

Le contenu textuel des articles d'actualité passe tout d'abord par une phase de prétraitement linguistique et d'indexation. La première étape (figure 1 (1)) consiste à analyser le texte pour en extraire d'une part les mots-clés normalisés après lemmatisation et d'autre part les entités nommées (p.ex. personnes, pays, organisations, etc...) qui y sont mentionnées. Les mots-clés et les entités nommées sont ensuite indexés (figure 1 (2)) pour faciliter et accélérer les recherches et les requêtes ultérieures. Pour un document d de n mots, la phase d'analyse et d'indexation a une complexité linéaire (i.e. $\mathcal{O}(n)$). Les connaissances ainsi extraites serviront par la suite à modéliser les intérêts des utilisateurs (figure 1 (3)) en analysant les contenus des articles avec lesquels ils ont interagi.

2.2 Les algorithmes de recommandation

Pour faire face aux défis de personnalisation des recommandations et du démarrage à froid, nous proposons une approche hybride par commutation (Burke, 2002) qui sélectionne, selon le cas à traiter, exclusivement une de ses deux approches sous-jacentes. La première approche, nommée PN, est non personnalisée dans la mesure où elle se base sur les critères de nouveauté et de popularité pour évaluer la pertinence d'un article d'actualité. Cette approche est adoptée pour faire face au problème de démarrage à froid et qui se présente dans les cas où l'utilisateur actif est inconnu ou lorsque l'article considéré est nouveau et n'a pas encore été lu ou évalué. Par contre, dans les cas où les données sur les l'utilisateur ou l'article ne sont pas parcimonieuses, l'approche a recours au filtrage par similarité de contenu. Dans ce travail, deux variantes du filtrage par contenu ont été évaluées. La première, nommée KBP, modélise l'article d'actualité avec ses mots-clés alors que la deuxième, nommée NeBP, n'a recours qu'aux entités nommées mentionnées dans le texte.

La proposition de l'approche non personnalisée PN est motivée par des études récentes portant sur la recommandation d'actualité et affirmant la pertinence des critères de popularité et de nouveauté comme indicateurs de l'intérêt que susciterait un article d'actualité chez le consommateur (Kille et al., 2016). Pour estimer la pertinence $PN(a)$ d'un article d'actualité a en fonction de sa nouveauté $N(a)$ et sa popularité $P(a)$, nous proposons l'équation 1 qui adopte une forme exponentielle (Ding et Li, 2005) pour modéliser la décroissance progressive

de la popularité au fil du temps et qui a été observée par des travaux existants (Kille et al., 2016). La modélisation de la nouveauté est motivée par le fait que les utilisateurs sont attirés par l'actualité tandis que le facteur popularité permet de mettre en avant les articles ayant suscité un grand intérêt brusque (i.e. un grand nombre de lectures) rendant leurs probabilités de lecture supérieures à celles des articles supposés intéresser l'utilisateur.

Dans la formulation proposée, $P(a)$ et $N(a)$ représentent respectivement le nombre de lectures de l'article a et son âge (en jours), tandis que K est un paramètre spécifiant le taux de dégradation de la popularité (deux à quatre jours (Kille et al., 2016)). Cette formulation permet ainsi de sélectionner les articles ayant le meilleur compromis entre nouveauté et popularité et a le mérite d'être adaptée au contexte de la recommandation temps-réel grâce à sa complexité constante ($\mathcal{O}(1)$).

$$PN(a) = P(a) * e^{-kN(a)} \tag{1}$$

Les approches KBP et NeBP proposent des recommandations personnalisées basées sur le contenu des articles d'actualité puisqu'elles suggèrent à un utilisateur donné les articles les plus similaires à ceux qu'il a déjà lus dans le passé. Cependant, les deux approches diffèrent au niveau de la représentation des articles et donc celle des profils des utilisateurs. En effet, KBP a recours à un vecteur de termes pondérés par leurs importances pour représenter un article ou l'historique des lectures d'un utilisateur alors que NeBP se limite aux entités nommées. Dans les deux variantes, nous adoptons la mesure de pondération TF-IDF (Pazzani et Billsus, 2007) pour quantifier l'importance d'un mot-clé ou d'une entité nommée dans un article d'actualité. Dans ce contexte, le profil de l'utilisateur s'enrichit au fur et à mesure des articles lus par les termes ou les entités nommées qui y sont présents. Enfin, nous avons recours à une variante de la mesure de similarité cosinus [4] pour calculer la concordance entre un article candidat et le profil de l'utilisateur actif et filtrer ainsi les actualités potentiellement capables de l'intéresser. Le calcul de similarité entre deux documents d_1 et d_2 ou entre un document d et un utilisateur u est proportionnel à la taille de leurs vecteurs représentatifs (resp. $\mathcal{O}(|d_1|+|d_2|)$ et $\mathcal{O}(|d|+|u|)$). Dans ce travail, le choix du cosinus comme mesure de similarité a été effectué sur la base de leur complexité et sa popularité dans le domaine de la recherche d'information. Son apport par rapport à d'autres mesures devrait être validé par des expérimentations comparatives contrôlées hors ligne.

3 Expérimentations

Les expérimentations ont été menées sur la plateforme de recommandation d'articles d'actualité de Plista [5] dans le cadre de la compétition NEWSREEL [6] de la conférence CLEF 2017. Cette plateforme collecte les contenus ainsi que les données d'usage sur les sites des éditeurs d'actualités partenaires, les partage en temps réel avec les systèmes de recommandation participants qui doivent en retour fournir des suggestions ciblées en moins de 100ms. La pertinence des systèmes participants est alors évaluée en fonction du nombre de consultations que leurs

4. `https://lucene.apache.org/core/4_6_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html`
5. Open Recommandation Platform, `https://orp.plista.com/`
6. http ://www.clef-newsreel.org/

Algorithme	Nombre de requêtes	Nombre de clics	CTR (%)
KBP	7891	59	0.7477
NeBP	34490	259	0.7509
NeHyb	75535	764	1,0115

TAB. 1 – *Résultats comparatifs des différents algorithmes testés pendant l'expérimentation.*

recommandations ont provoqué. En effet, plus le taux de clics sur les articles recommandés augmente, plus il est considéré comme performant.

Pour valider notre proposition, nous avons implémenté plusieurs prototypes basés sur quatre composants principaux à savoir (1) un service de communication avec la plateforme pour l'ingestion des données et la livraison des recommandations, (2) une base de données NOSQL (MongoDB) pour stocker les profils des utilisateurs, (3) un composant basé sur le framwork FOX [7] pour extraire les entités nommées sous forme d'une liste de triplés RDF et enfin (4) un index distribué (ElasticSearch) pour organiser les mots-clés et les entités nommées issus des articles de manière à accélérer les recherches et les requêtes lors de la phase d'inférence. Ces prototypes ont été déployés lors de la compétition sans phase d'apprentissage préalable et donc sans connaissance du domaine d'application.

Le Tableau 1 illustre les valeurs des taux de clics relatives à chaque algorithme que nous avons évalué pendant ce challenge. Les résultats montrent que l'algorithme NeHyb surpasse les stratégies de base : KBP et NeBP. De plus, nous pouvons constater que toutes les solutions à base de sémantique (NeHyb et NeBP) fournissent des recommandations plus pertinentes que l'autre solution (KBP). Cela montre que la modélisation des préférences d'un utilisateur avec des connaissances sémantiques dérivées des entités nommées permet de mieux cerner ses intérêts et donc de présenter les meilleures prédictions.

Le système de recommandation implémentant l'approche hybride NeHyb a été déployé sans connaissances à priori. Au fil du temps, il a collecté et analysé des données émanant de 82336 articles et de 3755547 utilisateurs. Ces traitements ont été menés sur une machine dotée de deux processeurs Intel Xeon E5-26xx avec une fréquence de 2GHz chacun, une mémoire cache de 4Mo, 4Go de RAM et 94Go d'espace de stockage. L'expérimentation nous a permis de valider l'intérêt de l'approche proposée puisque le système déployé a été capable de traiter le flux abondant et véloce des données tout en respectant la contrainte temps réel. En effet, le système a pu traiter jusqu'à 4000 requêtes par minute avec un temps de réponse moyen de 47ms, une occupation moyenne du processeur de 6% et une occupation moyenne de l'espace mémoire de 240Mo.

4 Conclusions et perspectives

Dans ce travail, nous avons proposé et évalué en conditions réelles une approche hybride, sémantique et personnalisée pour la recommandation des articles d'actualité en ligne. L'approche est capable de respecter les contraintes de la recommandation en temps réel tout en permettant le passage à l'échelle grâce à sa faible complexité et à sa capacité à traiter les flux

7. http ://aksw.org/Projects/FOX.html

continus de données dès leur arrivée tout en faisant l'apprentissage de manière incrémentale et non en lot à l'instar des approches basées sur les voisinages comme le filtrage collaboratif.

Nos futurs travaux porteront sur l'extraction, l'analyse et l'interprétation approfondies du contexte à partir du contenu non structuré des articles d'actualité et des comportements observés chez les consommateurs. La résolution de ce verrou scientifique permettrait de mieux raffiner les recommandations effectuées en les adaptant aux contexte de consommation.

Références

Burke, R. (2002). Hybrid Recommender Systems : Survey and Experiments. *User Model User-Adap Inter 12*(4), 331–370.

Capelle, M., F. Hogenboom, A. Hogenboom, et F. Frasincar (2013). Semantic News Recommendation Using Wordnet and Bing Similarities. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, SAC '13, New York, NY, USA, pp. 296–302. ACM.

Ding, Y. et X. Li (2005). Time Weight Collaborative Filtering. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, New York, NY, USA, pp. 485–492. ACM.

Domann, J., J. Meiners, L. Helmers, et A. Lommatzsch (2016). Real-time News Recommendations using Apache Spark. In *CLEF (Working Notes)*, pp. 628–641.

Kille, B., A. Lommatzsch, F. Hopfgartner, M. Larson, J. Seiler, D. Malagoli, A. Sereny, et T. Brodt (2016). CLEF NewsREEL 2016 : Comparing Multi-Dimensional Offline and Online Evaluation of News Recommender Systems. CEUR workshop proceedings.

Lommatzsch, A., N. Johannes, J. Meiners, L. Helmers, et J. Domann (2016). Recommender Ensembles for News Articles based on Most-Popular Strategies. In *CLEF (Working Notes)*, pp. 657–668.

Pazzani, M. J. et D. Billsus (2007). Content-Based Recommendation Systems. In *The Adaptive Web*, Lecture Notes in Computer Science, pp. 325–341. Springer, Berlin, Heidelberg.

Summary

Online news portals produce a huge amount of content in high velocity streams. In this context, it becomes more difficult to provide dynamic real-time and large-scale recommendations that best suit each user's interests. In this article, we present a hybrid news recommendation approach based on the semantic analysis of news articles' content. The approach exploits several personalized and non-personalized approaches to alleviate the cold start problem. Experiment results in an active large-scale news delivery platform during the NEWSREEL challenge show that our system produces significantly better quality recommendations than non-semantic recommenders.