Augmenting Data to Improve the Performance of Recommender Systems

Leticia Freire de Figueiredo[1,2], Joel Pinho Lucas[1], Aline Paes[2]

[1] Globo
Rio de Janeiro, RJ, Brazil

[2]Universidade Federal Fluminense
Niterói, RJ, Brazil

{leticia.figueiredo,joel.pinho}@g.globo, alinepaes@ic.uff.br

Abstract. *News recommendation plays an essential role in suggesting news to users in a personalized way. Most commonly, Recommendation systems (RS) suggest new articles that closely align with topics and themes users have read or engaged with previously. In this context, RS typically benefits from news metadata, providing key news' attributes, enabling the system to find similar news in content, topics, and themes. However, in most production systems, metadata is often manually filled and may not accurately reflect the true context of the news. To address this challenge in video news, we propose an automatic annotation approach powered by BERTopic, enabling precise tagging of news content. The automatically annotated metadata is then applied within a collaborative filtering algorithm that leverages association rules, enhancing the system's ability to identify and recommend relevant news. The proposed approach was experimented within a case study in Globo, where it showed an improvement in video views for user sessions.*

Resumo. *A recomendação de notícias desempenha um papel crucial na sugestão de notícias aos usuários de forma personalizada. A recomendação pode trazer notícias que sejam similares aos temas e tópicos presentes em artigos de notícias que o usuário leu no passado. Em geral, o processo de recomendação se vale de metadados anotados em cada notícia para descrever seus principais atributos. Entretanto, muitas vezes, estes metadados são preenchidos manualmente e podem não refletir de forma precisa o contexto da notícia. Para abordar este problema, propomos um processo de anotação automática para vídeos de notícias utilizando BERTopic. Os metadados anotados automáticos são, ao final, utilizados, em uma recomendação de filtragem colaborativa utilizando regras de associação. A abordagem apresentada foi testada em um caso de estudo na Globo, onde apresentou uma melhora na quantidade de visualizações de vídeos para sessões de usuário.*

1. Introduction

News recommendations have the goal of suggesting news to users in a personalized way [Karimi et al. 2018]. For example, content-based recommender algorithms suggest items similar to those the user has engaged with in the past [Lops et al. 2011]. The pieces of news to be recommended - in our case, news articles or videos - are usually annotated

with metadata that describes their key attributes. However, when the documents' metadata is filled manually, they might not accurately reflect or entirely encompass the context of the document. Moreover, depending on the volume of historical news, manually annotating the metadata is laborious and error-prone, also potentially leading to a lack of accurate metadata.

Given the problem of incorrect tagging, this paper proposes annotating them automatically in news videos, replacing manual annotation to improve the accuracy and effectiveness of recommendations. To do this, we define metadata as news topics automatically extracted using BERTopic [Grootendorst 2022]. The proposed framework processes the title and subtitle of the video to generate topics that reflect the contextual content of the video. Then, the recommender algorithm leverages these annotations so that videos with similar topic distribution will most likely be of interest to the users.

We show the improvement in recommendation through a case study focused on Globo, the largest mass media group in Latin America. Globo's vertical information portals are responsible for providing informative content to more than 100 million unique daily users. In this context, recommendation engines are critical for enhancing user experience and driving publishing revenue based on articles and video consumption. This way, the challenge lies not only in devising recommendations for millions of users with varying engagement levels, profiles and content preferences but also in keeping recommendations relevant, as thousands of news articles and videos are published daily. Our results show a clear improvement in video views, when compared to recommendations based on manual annotation.

2. Case Study: BERTopic for augmenting recommendation metadata

BERTopic is a topic modeling technique that uses embeddings from a pre-trained language model to create clusters from a given *corpus*, while also maintaining the most important words in the topic descriptions for clearer interpretation [Grootendorst 2022]. In [Michiels et al. 2023], the authors used BERTopic to define a topic for each news article, to compute the variety of topics each user was exposed to, by counting the number of unique topics within an observation window.

For our case study, the *corpus* contains the titles and subtitles of each news video. After BERTopic generates the topics from it, each video receives its corresponding topic that will be later used in the recommendation system. The recommendation system is built on a collaborative filtering algorithm that uses association rules, which we refer to as co-occurrence. The association rules focus on identifying patterns that predict the occurrence of one item based on the presence of other items within a transaction [Amatriain et al. 2010]. In our framework, the transaction corresponds to the total number of views within a given period.

In this context, each co-occurrence is restrained to consider videos that share the same topic generated by BERTopic, instead of considering videos from the whole catalog. Within the collaborative filtering approach, the videos recommended to the user will be ones that share the same topic and co-occurred with previous videos they watched.

We compared our method in an AB experiment against an alternative using co-occurrence collaborative filtering, but using manually defined topics. The AB testing

was built with the following setup: 1) the control: a naive approach, recommending the most recent videos from the catalog; 2) the baseline alternative: the previous collaborative filtering approach using manual annotations; 3) the collaborative filtering algorithm implementation using BERTopic. We measure the lift between the alternative metrics with the control alternative, as shown in Table 1. The **Video views per session** metric represents, on average, how many video views there are per user in a session. In the **Conversion rate** metric, we calculate the Click Through Rate (CTR). This metric calculates how many clicks the recommendation obtained [Jannach and Jugovac 2019]. In Globo's context, the CTR is used mostly as a guardrail metric, whereas the video views per session is the primary metric targeting user engagement.

The results demonstrate an improvement on the recommendation metrics when using automatic annotation, compared with the manual annotation. This indicates the recommendation became more accurate for the users.

Tabela 1. Lifts compared with the control alternative

	Alternative w/ automatic annotation	Alternative w/ manual annotation
Video views per session	7,04%	5,68%
Conversion rate	6,03%	5,34%

3. Conclusion

Recommendations in the news play an essential role in delivering personalized suggestions to users. Usually, a recommender algorithm uses document metadata to improve the recommendation. However, when the metadata is manually annotated, the recommendation will likely present bias and be irrelevant to the user because of incorrect tagging. In this paper, we proposed an automatic metadata annotation based on BERTopic, given a video news dataset. This metadata is used to improve an association rule collaborative filtering algorithm. The proposed solution was tested in a case study on Globo. The final results, after an AB experiment showed an improvement in click-through rate and video view metrics. As a future step, this approach will be experimented with annotating news articles, using different article parts - the article title, the article body, or both.

Referências

Amatriain, X., Jaimes*, A., Oliver, N., and Pujol, J. M. (2010). Data mining methods for recommender systems. In *Recommender systems handbook*, pages 39–71. Springer.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Jannach, D. and Jugovac, M. (2019). Measuring the business value of recommender systems. *ACM Transactions on Management Information Systems (TMIS)*, 10(4):1–23.

Karimi, M., Jannach, D., and Jugovac, M. (2018). News recommender systems–survey and roads ahead. *Information Processing & Management*, 54(6):1203–1227.

Lops, P., De Gemmis, M., and Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. *Recommender systems handbook*, pages 73–105.

Michiels, L., Vannieuwenhuyze, J., Leysen, J., Verachtert, R., Smets, A., and Goethals, B. (2023). How should we measure filter bubbles? a regression model and evidence for online news. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 640–651.