Beyond Metadata for BBC iPlayer: an autoencoder-driven approach for embeddings generation in content similarity recommendation

Simone Spaccarotella

September 2024

Contents

1 Introduction and background

I am a Software Engineer at the BBC, Team Lead for the Sounds web team, and I have been training as a Data Scientist, working in attachment with the iPlayer Recommendation team.

I built a machine learning model pipeline that produces content-to-content (C2C) similarity recommendations of video on-demand (VOD), for the "More Like This" section on BBC iPlayer [1]. This project is relevant to me because it is about recommendations, and I have been crossing paths with this world multiple times during my career at the BBC. I had a tangent encounter back in 2015, while working for a team that was building an initial recommender for BBC News and an API to provide recommendations using 3rd party engines. During a Hack Day some time later, I produced and presented a talk called "Recommendation Assumptions" [3], which was about recommendations and external factors affecting them, contextual to the consumption of the content itself.

The BBC is a well-known British broadcaster, and it is always evolving to remain relevant to its audience. Its mission is to inform, educate and entertain, and it operates within the boundaries set by the Royal Charter [2]. The current media landscape requires the BBC to deliver digital-first content that is relevant to the audience, and this involves investements in data and personalised services, not to mention a certain revolution in machine learning that is keeping everyone busy.

2 Outline of the issue or opportunity and the business problem to be solved

The BBC produces and stores a vast amount of metadata for its content, and it is surfaced by countless services and APIs. One of the priority for the BBC is to increae the adoption across the business of "Passport", an internal service that generates, stores and provides access to a richer dataset of metadata annotations for multi modal content (audio, video and text). The usage in production is very low if not existent, and its adoption would make the access to metadata consistent, removing duplications and reducing effort and costs.

Furthermore, the similarity score of the current C2C recommender is directly proportional to the number of values in common between any pairs of items on a per-feature basis. But the commonality is calculated with an exact string equality, hence it ignores any relationship between different strings expressing a similar concept (e.g. "comedy", "stand-up comedy"). The number and types of tags are not enough to sufficiently describe the content. The data distribution of each feature is severely skewed towards the most popular category and no pre-processing is applied. Lastly, each feature similarity score is multiplied by a weight to manually set the importance, but unfortunately, it doesn't solve the polarising effect of a skewed distribution and because they are hyperparameters and not learned weights, the model can't improve its performances by minimising them against a cost function.

To address these issues, the aim of this project is:

- **To improve the quality of the recommendations**. My hypothesis was that by using in input a richer set of metadata that better describes the content, and by reducing the high-dimensional data to a lower-dimensional latent manifold, the model is able to generate embeddings that can improve the performances of the model by mapping the item similarity problem to a geometric distance calculation between vectors in a multi-dimensional Euclidean space.

- **To reduce the costs to generate C2C similarity recommendations**. I sourced the data from Passport, to build a BBC-wide general solution that can be used by any product and can be applied to any content recommendation, because they all share the same set of annotations.

3 Methods used & justification

4 Scope of the project and Key Performance Indicators

5 Data selection, collection & pre-processing

6 Survey of potential alternatives

7 Implementation - performance metrics

8 Results

9 Discussion & conclusions/recommendations

10 Summary of findings

11 Implications

12 Caveats & limitations

13 Appendices

References

[1] BBC. Bluey - "more like this" tab. https://www.bbc.co.uk/iplayer/episodes/m000vbrk/bluey?seriesId=more-like-this.

[2] BBC. The Royal Charter. https://www.bbc.com/aboutthebbc/governance/charter, 2017. Valid until 31 December 2027.

[3] Simone Spaccarotella. Recommendation assumptions. https://www.slideshare.net/slideshow/recommendations-assumptions/236291920, 2015. Presented at Prototyping Day @ Mozilla London on 3 September 2015, at Engineering Summit @ BBC on 7 March 2018, uploaded on SlideShare on 27 June 2020.