# Enhancing Personalised Recommendations with the Use of Multimodal Information

Taner Cagali*, Mehrnoosh Sadrzadeh*,
Department of Computer Science,
University College London, U.K
Email: taner.cagali.20@ucl.ac.uk,
m.sadrzadeh@ucl.ac.uk

Chris Newell ‡,
BBC Research and Development
The Lighthouse, White City Place, London, U.K.
Email: chris.newell@bbc.co.uk

***Abstract*—Whenever we watch a TV show or movie, we process a substantial amount of information that is conveyed to us via various multimedia mediums, in particular: visual, textual, and audio. These data signify distinctive properties that aid in creating a unique motion picture experience. In effort to not only produce a more personalised recommender system, but also tackle the problem of popularity bias, we develop a system that incorporates the use of multimodal information. Specifically, we investigate the correlation between features that are extracted using state-of-the-art techniques and deep learning models from visual characteristics, audio patterns and textual subtitles for a given programmes. We evaluate this framework in a dataset of 145 British Broadcasting Corporation (BBC) TV programmes against genre and user baselines. Along with metadata, we demonstrate that personalised recommendations can not only be improved with the use of multimodal information, but also outperform the user-based model in terms of diversity, whilst maintaining matching levels of accuracy. These results also show that the use of more advanced neural techniques for encoding textual and visual data, in the form of NTM-R neural topics and Res-Net image features, does improve the performance.**

***Index Terms*—Multimedia systems; Information filtering; Recommender systems; Content-based retrieval**

## I. Introduction

Modern recommender systems tend to be driven by either collaborative filtering or content-based filtering methods. Collaborative filtering systems are based upon a programme's valuation from an assortment of users [1]. They are capable of producing sufficient recommendations by simply distinguishing the intersecting interests amongst their users. Collaborative filtering emphasizes the notion of community, as the overall performance of the recommender is heavily influenced by the choices made by its users. Though, it does make the simplifying assumption that the target user will favour content that is akin to the content of a user with comparable preferences. Content-based recommender systems are based on item descriptors or features [2], this is usually in the form of metadata (e.g., genre, title, cast, tags). Within this system recommendations are regarded as a user-specific problem, in which the model learns the users' preferences via item features.

The fundamental objective of a recommender system is to be able to provide pertinent suggestions to its users, thus, such systems ought to be considering the underlying content of the items to be able to truly capture the purpose that is driving a user's inclination to certain content. While extracting features from the visual, textual and audio domains is well researched within the field of multimedia, to date, recommender systems mostly only take advantage of metadata, even and especially in the most state-of-the-art systems. Representing items directly from its content, however, poses additional advantages; collaborative filtering systems are unable to produce recommendations for new or rare items, due to the lack of user interaction, this is what is known as the cold-start problem [3]. This is a growing concern as recommender systems are continuously expanding with new items and users. Moreover, metadata can be absent or lacking for new items making it problematic in providing adequate recommendations. Collaborative based systems also suffer from what it is known as popularity bias [4], whereby frequently consumed items receive a lot of exposure, whilst less popular items are under-represented within the recommendations.

In this paper, we propose a multimodal similarity based TV programme recommender system that incorporates multimodal features to capture the unique interests of its users, and alleviate the issue of popularity bias that arises with collaborative filtering-based systems. We model it to be a similarity problem, whereby the model learns to predict the top-n set of programmes that are most likely to appease a user depending on their viewing history.

This paper is a third in our line of work. In [5], we presented a framework that modelled content using textual and audio data, as well as genre metadata. In [6], we extended that framework with SIFT video features. In this paper, we improve on our previous methodology by (1) training an NTM-R neural topic model, instead of using old fashioned LSI topic models, and (2) working with a range of new video features, from chromatic and luminance descriptors to neural features using pre-trained Res-Net and VGG models. These improvements led to better accuracy and also diversity of recommendations. The improvement on the topic models was from 11.30% to 13.5% in MAP and from 69.89% to 71.27% in ILD. Similarly on the video content, we increased the MAP from 3.88% to 8.48%, but ILD slightly decreased. The overall MAP of the model, obtained by a late fusion of the similarities between each content model, was from 14.98% to 15.76%, and ILD also increased from 61.29% to 64.02%.

The remainder of the paper is structured as follows. In the next section we summarise the structure of the dataset and generation of the genre and user baselines. In section 3 we go over the feature retrieval processes. Within Section 4 the architecture of the recommender systems will be outlined. The experimental results will be put forth within section 5. Finally, our conclusions will be presented in section 6.

## II. Dataset

### A. Data Description

To validate the proposed model, a dataset over a two-week period was compiled from the BBC iPlayer server logs. The first week comprises of 145 programmes of a single episode each, consisting of 1390540 viewings from 33958 users, which will be utilised as training data. The data gathered within following week will be used for testing and consists of 141 programmes, with 47707 viewings from 10000 users. The users within the testing data are a subset of the users within the training data. A viewing is recorded if a minimum viewing time of 5 minutes is exceeded. This threshold was chosen because the number of viewers falls rapidly in the first few minutes, but then stabilises. Thus, a viewing time over 5 minutes was interpreted a positive preference for a programme.

### B. Genre and User Baselines

In order to assess the performance of the multimodal features, we implemented two baseline recommenders, the first based on content metadata, the second using collaborative filtering, against which our results can be compared. The genre for each programme is represented by a hierarchical tree with up to three levels (e.g., factual, factual/sci & nature, factual/sci & nature/nature & env). The hierarchical structure is broken down into a set of attributes. The set is then used to represent each programme by a binary feature vector, where each column represents a genre subtree. Likewise, using the training set, binary item-user feature vectors were also created, where each entry denotes whether a user has viewed a programme for more than 5 minutes.

## III. Method

### A. Textual Analysis

We start our textual analysis by learning the distributed representation of each set of subtitles. This is achieved by using the Doc2Vec model proposed by Le and Mikolov [7]. Which is an extension of the neural semantic Word2Vec model [8]. Doc2Vec enhances the methodology of learning neural word embeddings to sentences, paragraphs and documents. Similar to the Word2Vec model, Doc2Vec has two primary architectures: Paragraph Vector – Distributed Bag-of-Words (PV-DBOW) and Paragraph Vector – Distributed Memory (PV-DM). We specifically worked with PV-DM, where the idea is to concatenate a unique document ID with the context words to predict the following word [7]. We use averaging as the means to combine the vectors.

In the next stage we deploy a topic modelling algorithm, namely Neural Topic Model (NTM) with topic coherence regularisation (NTM-R) [9]. The topic coherence regularisation term leverages pre-trained word embeddings, as they carry contextual similarity that is highly related to mutual information [9]. Topic modelling is an information retrieval technique that finds the abstract "topics" within a set of documents via their latent semantic structure. The core principle is that rather than projecting documents in high dimensional term space, we project them within a much lower dimensional space. The commonly used topic models are: Latent Semantic Indexing (LSI) [10] which makes the use of matrix factorisation to find a low-rank approximation of a document-term matrix, and Latent Dirichlet Allocation (LDA) [11], a probabilistic generative model, where each topic is modelled as distribution over a set of terms. However, since the introduction of Variational Autoencoders [12] there has been advancements within the field of topic modelling. Models such as GSM [13], ProdLDA [14] and NTM [9] have proven to be able to construct much more coherent topics in comparison to the traditional methods. The document-term matrix is constructed by way of bag-of-words (BoW) and the pre-trained word embeddings used are GloVe [15].

### B. Audio Analysis

The audio signal contains a mixture of various forms of information from speech to music tracks and sound-effects. In attempt to represent such information, the following audio descriptors were extracted: Mel-frequency Cepstral Coefficients, Spectral Centroid, Zero Crossing Rate, Spectral Flatness and Root Mean Square. All of which was obtained using LibROSA [16]. We then deploy the bag-of-audio-words (BoAW) method [17]. Within this approach a set of "audio words" is learned via K-mean clustering. The centroids of the resulting clusters are perceived as the audio-words, and the original feature vectors are replaced by a single index indicating the closest cluster centroid to the feature vector. The "bag" is generated by creating a histogram of these indices. The audio descriptors are concatenated with the histograms and then normalised, to assure that the dispersion of the distributions is conserved.

### C. Visual Analysis

A video consists of successive frames that are highly similar and correlated. Therefore, prior to feature extraction, keyframes are detected following [18]. Utilising keyframes for feature extraction is not only computationally less expensive, but they also provide a more detailed summary of the video content. In regards to the feature extraction process, we focus specifically on three types of video features: appearance features from colour, illumination, objects and scenes, action features inferred via human gestures and motion, as well as emotion features derived from facial expressions.

We begin the feature extraction process by transforming each keyframe to the Hue, Saturation, Value (HSV) colour-space and generate 16-bin histograms for every coordinate to capture lighting and chromatic characteristics. The average over all histograms is then computed to obtain the video-level feature representation for each video. Next, we employ a Res-Net 152 model pre-trained with ImageNet [19], a Res-Net 50 model pre-trained with Places365 [20], and a VGG19 model pre-trained with FER-2013 [21], to extract object, scene and emotion features respectively. The feature representation is acquired by averaging the results from the last convolutional layer for each keyframe. Lastly, we use a 3D Res-Net model pre-trained with Kinetics-400 [22] to extract action features. The video file is first segmented into 14 consecutive keyframes. Subsequently, the output from the last convectional layer is taken, which then the average is computed over all segmentations. The features are concatenated, and then dimensionality reduction is performed via Principal Component Analysis (PCA), in order to avoid the curse of dimensionality [23].

## IV. Recommender Framework

We gauge the effectiveness of our representations using a weighted item k-nearest neighbours (k-NN) approach, based on an earlier framework MyMediaLite [24]. The recommender works as follows: for every item, an aggregated score is computed based on the top-n most similar programmes via their similarity score, items that are part of the user's viewing history (in the training set) are omitted from the recommendations. The recommendations are then made based on the rank of the aggregated scores. In effort to combine the various modalities, late fusion is performed by taking the weighted average of the aggregated scores. The performance of the recommender is then determined by producing recommendations for each user found in the testing data and comparing these with the items they consumed. The framework supports two accuracy metrics:

- **Mean Average Precision (MAP)** reflects the number of correctly predicted viewings found in the top-N recommendations (hits) [25]
- **Normalised Discounted Cumulative Gain (NDCG)** Is similar to MAP but takes into account the position of hits within the top-N. Hits at the top of the recommendations are weighted higher than those at the bottom of the list [25]

The framework also supports four diversity metrics:

- **Intra-list diversity (ILD)** measures the diversity of the genres in the recommendations for each individual user [26]
- **Surprisal** reflects how many low popularity items are present in the recommendations [27]
- **Personalisation** measures how each user's recommendations differ from other users [27]
- **Coverage indicates** how many of the available programmes are present in the recommendations [28]

Quality of recommendations if often measured by accuracy, though, the most accurate recommendations may not always be the most useful to users, and other factors which may impact recommendations should be taken into consideration. For instance, a significant metric from the user's perspective would be ILD, as recommendations which only include one or two genres would be unappealing and monotonous. Furthermore, personalisation will indicate how unique the set of recommendations are to a user, and if the system is tending to just recommend the same programmes. Coverage and surprisal will be able to provide insight on how fairly the programmes are being treated, i.e., whether if some programmes will appear in anybody's recommendations and whether or not low popular programmes are incorporated within the recommendations Maintaining high diversity measures will assure a even distribution of recommendations, as well as reduce popularity bias, ensuring a more personalised recommender system.

## V. Evaluation and Results

We evaluate the similarity rankings both individually and fused at a rank of 10, i.e., the first 10 recommendations. Tables 1, 2 and 3 present the results for the genre and user baselines, individual modalities and finally the multimodal representations. Regarding the baseline recommenders, we can see that the user-based model produces the best recommendations in regards to MAP, NDCG and ILD, though, does fall quite behind in surprisal, personalisation and coverage. It clear to see the model suffers severely from popularity bias. This is only natural, as the recommendations are based purely upon audience behaviour. The genre-based model is able to overcome this issue, however, it struggles to produce recommendations containing varying genres, as portrayed by the ILD score of 35.52%, which is only expected given the nature of the data. Additionally, MAP and NDCG scores of 10.78% and 19.71% convey that is unable to predict user viewings as effectively as the user-based model. Concerning the content-based models, the textual features are able to produce the most adequate recommendations when compared to the audio and video-based models. Further, the NTM-R-based model is able to maintain high diversity metric scores, whilst only performing 2.1% and 3.35% lower in terms of MAP and NDGC, than the user-based model. Despite the fact that the content-based models get around the issue of popularity bias, greater MAP and NDCG is still desirable.

TABLE I
Baseline Evaluations

| Model | MAP | NDCG | ILD | Surprisal | Personalisation | Coverage |
|---|---|---|---|---|---|---|
| Genre (G) | 10.78% | 19.71% | 35.52% | 0.59 | 85.61% | 98.58% |
| User | 15.60% | 27.20% | 79.73% | 0.32 | 64.41% | 74.47% |

TABLE II
Singular Modality Evaluations

| Model | MAP | NDCG | ILD | Surprisal | Personalisation | Coverage |
|---|---|---|---|---|---|---|
| DM | 11.76% | 21.46% | 77.20% | 0.57 | 85.17% | 100.00% |
| NTM-R | 13.50% | 23.85% | 71.27% | 0.53 | 80.56% | 100.00% |
| NTM-R+DM (T) | 14.75% | 25.59% | 71.13% | 0.53 | 81.13% | 100.00% |
| Audio (A) | 6.67% | 13.61% | 77.96% | 0.57 | 83.73% | 100.00% |
| Video (V) | 8.48% | 16.61% | 69.62% | 0.53 | 75.27% | 100.00% |

Upon inspection of table 3, the most noteworthy conclusion is that the hybridisation of the content features does in fact enhance the performance of the recommender. The combination of text, audio and video provides an improvement of 11.1%

TABLE III
MULTIMODAL EVALUATIONS

| Model | MAP | NDCG | ILD | Surprisal | Personalisation | Coverage |
|---|---|---|---|---|---|---|
| V+A | 9.50% | 18.14% | 70.11% | 0.52 | 74.53% | 100.00% |
| T+A | 14.82% | 25.67% | 70.49% | 0.52 | 80.79% | 100.00% |
| T+V | 14.90% | 25.75% | 70.44% | 0.53 | 80.81% | 100.00% |
| T+A+V | 15.00% | 25.89% | 70.02% | 0.52 | 80.61% | 100.00% |
| T+A+V+G | 15.76% | 26.92% | 64.02% | 0.52 | 81.57% | 100.00% |

over MAP and 8.6% to NDCG, whilst still being able to maintain high diversity scores, when compared to the best singular model (namely NTM-R). Further, it is able to achieve accuracy scores close to that of the user-based model. In more detail, fusing with the genre-based model does provide a boost to accuracy, although, it does degrade the ILD score considerably. It is important to note that even models with low metric scores are still able to provide a boost to the performance of the recommender. This is probably the most intriguing outcome of this experiment, as it demonstrates that there is much diversity between the modalities, and that the combination of them allows the recommender to effectively capture the overlying set of recommendations between them. This is encouraging, as it illustrates that the use of multimodal information is able enhance the overall performance of a content-based recommender system.

## VI. Conclusion

In this paper a multimodal similarity based TV programme recommender system was presented. We make use of various state-of-the-art models and techniques to exploit multimodal features. To incorporate the various modalities within the recommendations, we proposed a weighted item k-NN recommender. The model is then evaluated on two weeks worth of BBC iPlayer log data. We use both accuracy (MAP and NDCG) and diversity (ILD, surprisal, personalisation and coverage) metrics to measure the ability the model. The results presented suggest that a multimodal-based recommender system could be an alternate option to a collaborative-based one, as it able to match the performance of a user-based model in terms of accuracy, whilst still being able to sustain high diversity scores.

This work provides a foundation for future study, as additional feature extraction methods can be explored to obtain richer content descriptors. In addition, more sophisticated fusion schemes could also be examined.

Our results also improved on results of previous work [5], [6] by taking advantage of neural topic models and neural video features, emphasising the repeatedly learnt lesson of state of the art machine learning, that neural models do indeed learn better quality features and thus also improve the overall results in tasks of interest, and in our case in content learning and generating recommendations based on them.

## References

[1] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen, "Collaborative filtering recommender systems," in *The adaptive web*. Springer, 2007, pp. 291–324.

[2] P. Lops, M. De Gemmis, and G. Semeraro, "Content-based recommender systems: State of the art and trends," *Recommender systems handbook*, pp. 73–105, 2011.

[3] B. Lika, K. Kolomvatsos, and S. Hadjiefthymiades, "Facing the cold start problem in recommender systems," *Expert Systems with Applications*, vol. 41, no. 4, pp. 2065–2073, 2014.

[4] H. Abdollahpouri, "Popularity bias in ranking and recommendation," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 529–530.

[5] *Cosine Similarity of Multimodal Content Vectors for TV Programmes*, 2020.

[6] *Audiovisual, Genre, Neural and Topical Textual Embeddings for TV Programme Content Representation*, 2020.

[7] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*. PMLR, 2014, pp. 1188–1196.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *ICLR*, 2013.

[9] R. Ding, R. Nallapati, and B. Xiang, "Coherence-aware neural topic modeling," *arXiv preprint arXiv:1809.02687*, 2018.

[10] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[12] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[13] Y. Miao, E. Grefenstette, and P. Blunsom, "Discovering discrete latent topics with neural variational inference," in *International Conference on Machine Learning*. PMLR, 2017, pp. 2410–2419.

[14] A. Srivastava and C. Sutton, "Autoencoding variational inference for topic models," *arXiv preprint arXiv:1703.01488*, 2017.

[15] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[16] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015.

[17] Y. Liu, W.-L. Zhao, C.-W. Ngo, C.-S. Xu, and H.-Q. Lu, "Coherent bag-of audio words model for efficient large-scale video copy detection," in *Proceedings of the ACM international conference on image and video retrieval*, 2010, pp. 89–96.

[18] C. Lv and Y. Huang, "Effective keyframe extraction from personal video by using nearest neighbor clustering," in *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, 2018, pp. 1–4.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[20] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

[21] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *International conference on neural information processing*. Springer, 2013, pp. 117–124.

[22] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[23] R. E. Bellman *et al.*, "Dynamic programming, ser," *Cambridge Studies in Speech Science and Communication. Princeton University Press, Princeton*, 1957.

[24] MyMediaLite, "Mymedialite recommender system library," http://www.mymedialite.net/, (Accessed on 02/18/2020).

[25] C. Manning, P. Raghavan, and H. Schütze, "Xml retrieval," in *Introduction to Information Retrieval.* Cambridze University Press, 2008.

[26] B. Smyth and P. McClave, "Similarity vs. diversity," in *International conference on case-based reasoning*. Springer, 2001, pp. 347–361.

[27] T. Zhou, Z. Kuscsik, J.-G. Liu, M. Medo, J. R. Wakeling, and Y.-C. Zhang, "Solving the apparent diversity-accuracy dilemma of recommender

systems," *Proceedings of the National Academy of Sciences*, vol. 107, no. 10, pp. 4511–4515, 2010.

[28] M. Ge, C. Delgado-Battenfeld, and D. Jannach, "Beyond accuracy: evaluating recommender systems by coverage and serendipity," in *Proceedings of the fourth ACM conference on Recommender systems*, 2010, pp. 257–260.