Audiovisual, Genre, Neural and Topical Textual Embeddings for TV Programme Content Representation

Saba Nazir*, Taner Cagali †, Mehrnoosh Sadrzadeh *
Department of Computer Science,
University College London, U.K
Email: saba.nazir.19@ucl.ac.uk, 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*—TV programmes have their contents described by multiple means: textual subtitles, audiovisual files, and metadata such as genres. In order to represent these contents, we develop vectorial representations for their low-level multimodal features, group them with simple clustering techniques, and combine them using middle and late fusion. For textual features, we use LSI and Doc2Vec neural embeddings; for audio, MFCC's and Bags of Audio Words; for visual, SIFT, and Bags of Visual Words. We apply our model to a dataset of BBC TV programmes and use a standard recommender and pairwise similarity matrices of content vectors to estimate viewers' behaviours. The late fusion of genre, audio and video vectors with both of the textual embeddings significantly increase the precision and diversity of the results.**

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

I. Introduction

Events and facts of the world and our perceptions and dramatisations of them trigger the production of items that express and explain them. TV programmes are amongst such items. With growing numbers of these and similar items being produced every day, there is need for formally representing their contents and automatically reasoning about them. The difficulty is that there are different types of features associated to each programme: textual information in the title, description, and subtitles; metadata in the form of genre, format and service, and the audiovisual files. In this work, each modality is explored with off-the-shelf-techniques where the novelty lies in combining different modalities in a uniform setting. Therefore, our main goal has been to develop a methodology to combine different forms of content into one unified vectorial representation. Our second goal has been to test the quality of these vectors by using their pairwise matrix similarities and estimate viewers' behaviours. We learn LSI topic vectors and Doc2Vec neural embeddings for subtitles, turn the MFCC acoustic and spectral audio features into Bags of Audio Words, and the SIFT visual data into Bags of Visual Words. Hierarchical genre data are turned into vectors via a tree traversal algorithm of their hierarchies.

We fuse these vectors and turn them into content similarity matrices, each cell of each one of them contains the cosine distance between the content vectors of two programmes. We evaluate these matrices by implementing our model on a dataset of BBC TV programmes and computing degrees of correlation between the pair wise similarities and the behavioural similarities coming from viewers' history. This dataset is chosen because most of the publicly available multimodal datasets do not come with all major forms information (text, audio, video and genres) readily available for the same set of documents. For example, MM-IMDb [1] do not contain subtitles and Youtube 8M dataset [2] is built only on visual features. Because of our data size limitations, we have used off-the-shelf techniques to extract, analyse and combine data obtained from multimodal sources. On the methodological side, this is the first time that neural (Doc2Vec) and topical (LSI) document embeddings are combined with audio (BoAW) and video (BoVW) vectors and vector representations of genres. On the evaluation side, our representations significantly improve the precision and diversity of programme recommendations in a standard recommender and it is the first time such a study has been done for a dataset of TV-only programmes.

Using vectors for representing contents of words and documents originates in the vector semantics of Natural Language Processing and Information Retrieval [3], [4]. Recently, the textual vector representations have been learnt by neural networks, e.g. via the Skipgram algorithm resulting in Word2Vec and Doc2Vec [5]. These have been enriched by audiovisual and cognitive information, e.g. see [6], [7]. The theoretical basis of our model is the work of [7], which we extend from words to documents, and enrich with genre and visual vectors. Using multimodal features in recommendations is common, e.g. see [8] and [9]. Unlike these, we do not learn mappings for users preferences, rather, we use the behavioural similarity of users and the contents of the programmes to compute a degree of correlation between the two. Multimodal content-based and hybrid recommenders have also been considered in [10] for collaborative filtering, and in [11] for e-commerce assortment. The primary focus of our paper has not been improving recommenders, be it hybrid or collaborative. Our

evaluations, however, show that in content-based recommenders, our model is more extensive and conclusive than existing ones, e.g. [12], only considers tags and titles as textual data, [13] combines images with tags, and [14] uses audio and video with subtitles, but ignores the genres, as a result does not improve on the performance of a metadata-only system.

II. Multimodal Content Vectors

Our dataset contains 145 BBC TV programmes with their subtitle and audiovisual files and metadata information. A visualisation of our framework is presented in Figure 1.

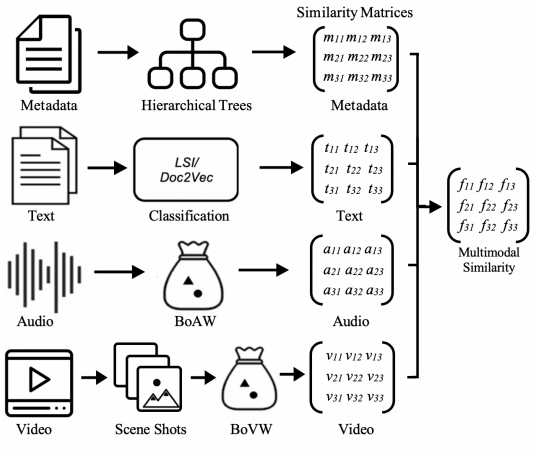

Fig. 1. Multimodal Content Recommendations Framework

A. Metadata Vectorization

These representations are based on editorially-assigned attributes of programmes. Each programme has a genre which is hierarchical with up to three levels (e.g. *factual*, *factual/sci&nature*, *factual/sci&nature/nature&env*) and a match can occur at any level. The hierarchical structure is broken down into a set of attributes by traversing the tree. This set is represented by vectors, where each column represents a genre subtree obtained from a partial tree of the genre hierarchy and each column entry is a binary value denoting the relation between the program and the genre, i.e. whether the programme had that partial tree as part of its genre hierarchy. Using the vectors thus obtained, we computed a metadata similarity matrix, where a complete match receives a score of 1 but the score is halved for each level above.

B. Subtitle Vectorization

Latent Semantic Indexing (LSI) [15], a topic modelling technique, is applied to the subtitles. LSI is a two-step procedure. Firstly, a document- term matrix is generated via a low-rank approximation obtained from the term vector space projections of the Bag of Words vectors. Secondly, Singular Value Decomposition (SVD) is applied to the document-term matrix, where the newly created eigenvectors represent the concepts within the latent space. We worked with 50 dimensional spaces. LSI improves on the term-document matrices, but does not take word order into account. To deal with this, we worked with neural semantics embeddings Doc2vec [16]. Doc2vec is an extension of the neural semantic word embeddings Word2vec [5]. We worked with Paragraph Vector Distributed Memory (PV- DM), which concatenates the unique document ID with the context words with respect to the specified context window over the text and preserves the order of words.

C. Audio Vectorization

The audio files of programmes are a mixture of dialogue and background music. In order to represent both of these information, we chose features used in speech recognition and features from music information retrieval, namely MFCCs; and Spectral Centroid, Zero Crossing Rate, Spectral Flatness and Root Mean Square, all obtained using LibROSA [17], keeping audio sampling rate of 22050 Hz and hop length of 512 samples, with variable lengths of audio tracks averaging on about 30 mins each for a detailed analysis. The extracted features are concatenated, normalised and used as audio vectors for each audio. We then followed [7] and used Bag of Audio Words (BoAW) model to learn abstract audio vector representations via mini-batch K-means clustering with k=50. BoAW is widely used in audio information retrieval and recognition [7], [18] and acoustic event detection [19].

D. Video Vectorization

Videos are represented by a sequence of still images extracted from the middle of each scene. The images were extracted using the SceneDetect application provided by the PySceneDetect library, using the ContentDetector algorithm with threshold=30 and a minimum scene length of 15 frames. The middle images were chosen from each scene because these were most representative and had lower levels of motion blur. A subset of 600 images for each programme was chosen for classification. After careful investigation of local and global image features, SIFT image descriptors were used [20] [21] for scene classification; these at are known for their powerful capability of image matching [22], object detection and recognition [23]. SIFT provides 128D feature vector against each keypoint in an image. Once the features are extracted, BoVW model [24] was used to quantise all image descriptors and create a visual word vocabulary for each programme using K-means with k=300.

E. Fusion

Once the individual vector representations are obtained, they are combined and result in a multimodal representation for each programme. Various fusion techniques have been used in literature to combine the information encoded via different modalities [25], [8]. The most common of techniques are early, middle, and late fusion.

Since the data obtained from these multimodal sources have very different properties, following Kiela and Clark [7], we focused on middle and late fusion. In middle fusion, the feature vectors of all modalities were concatenated with each other. In late fusion the weighted addition of the similarity matrices

of each modality was formed. The optimal set of weights were learnt for each combination. The sparsity of metadata representations tend to reduce system's performance in middle fusion, hence late fusion eventually turned out to be the most effective method.

III. Evaluation and Results

We used a personalised recommender evaluation system based on the MyMediaLite library [26] to evaluate the performance of our representations. It takes binary user-item preference training and testing data, obtained from BBC iPlayer media server logs. In this data, a user's positive preference is recorded when they exceed 5 mins viewing time, a threshold chosen by observing the lapse rate (the rate at which users stop watching a programme). The first week of recorded data is used for training and a subset from the following week is used for testing. We have 1390540 viewings of 33958 users for 145 TV programmes as training data, and 47707 viewings, 10000 users, and 141 programmes as testing data; where the users in testing are a subset of the users in the training data.

Weighted Item-based K Nearest Neighbours (KNN) algorithm provided in MyMediaLite [26] is used to obtain K Nearest Neighbours for each correlation matrix to predict recommendations for testing partitions. The accuracy was measured using Mean Average Precision (MAP): based on the number of correctly predicted viewings found in the top-N recommendations (hits). We also computed the Intra-list diversity (ILD), which measures the diversity of the genres in the recommendations for each individual user.

We evaluated our representations using individual and fused models. The results for textual (LSI, DM), audio (A), video (V), and genre (G) individual matrices are presented in Table I; the combinations of fused textual, genre, and audio matrices in Table II, the combinations of textual, genre, and video matrices in Table III. The combination of all matrices is presented in Table IV. In these tables, the weights associated with each modality represent how much it contributed to the explanatory power of the model. Table I shows that even individually, textual matrices outperform genres in both MAP and ILD. Although audios and videos have lower MAPs, they are very good in diversifying the results as compared to textual and genre representations. Videos showed the highest ILD of 82.05% at rank 20 because of lowest MAP. Our best result was an ILD of 69.88% with a high MAP of 17.55%, for the fully fused model, in Table IV. These results come very close to user-based, with MAP of 18.51% and show the strength of fused representations in estimating users' behaviours.

A short analysis is presented in Figures 2 and 3; here, the fused vector distances are compared with user-based similarity distances between EastEnders and 20 other BBC TV programmes. These programmes are chosen from our gold-standard user-based recommendations. Two important conclusions can be drawn here: 1) Fusing audiovisual vectors with both of the textual vectors and with genre representations boosts system's performance noticeably and 2) The estimations of the fused model are better, for the cases where the genres are the same and even for the cases when they are different; this showcases the better performance of our model in comparison to a genre-only system.

TABLE I
Singular Model Evaluations

Model	MAP@10	ILD@10	MAP@20	ILD@20
Genre (G)	10.78	35.52	12.77	52.72
Doc2vec (DM)	11.76	77.20	13.88	80.37
LSI	11.30	69.89	13.40	76.69
Audios (A)	6.67	77.96	8.11	81.38
Videos (V)	3.88	81.43	4.97	82.05
User-Based	15.60	79.73	18.51	80.90

TABLE II
Fused Textual Audio Genre Evaluations

Model	MAP@10	ILD@10	MAP@20	ILD@20
LSI+ A+ G 0.5 0.3 0.2	12.87	63.10	15.21	71.65
DM+ A+ G 0.7 0.2 0.1	13.78	59.03	16.17	67.87
LSI+ DM+ A+ G 0.7 1.5 0.2 0.65	14.98	61.29	17.45	70.00
User-Based	15.60	79.73	18.51	80.90

TABLE III
Fused Textual Video Genre Evaluations

Model	MAP@10	ILD@10	MAP@20	ILD@20
LSI+ V+ G 1.00 0.13 1.00	13.48	53.00	15.74	64.20
DM+ V+ G 1.8 0.1 1.00	14.23	54.75	16.62	64.68
LSI+ DM+ V+ G 0.7 1.5 0.12 0.65	14.99	60.62	17.45	69.58
User-Based	15.60	79.73	18.51	80.90

TABLE IV
Fused TextualAudio Video Genre Evaluations

Model	MAP@10	ILD@10	MAP@20	ILD@20
LSI+ DM+ A + V+ G 0.7 1.5 0.12 0.1 0.65	15.07	61.17	17.55	69.88
User-Based	15.60	79.73	18.51	80.90

IV. Conclusion

In an earlier version, presented as a poster in the Machine Learning for Media Discovery Workshop of ICML2020, we worked with late fusion of subtitle Doc2Vec and LSI vectors and the audio and genre vectors. In this paper, we enriched these representations with visual feature vectors generated from SIFT and BoVW. The addition of these to the late fusion, increased the precision and diversity of estimating viewers' behaviours and we conclude that this combination is a good candidate for representing contents of TV programmes. Our system can easily be extended to other data modalities such as high level audio and global visual features, and applied to other datasets, such as IMDB– after being extended with subtitles; these are work in progress. [11] jointly learns multimodal representations

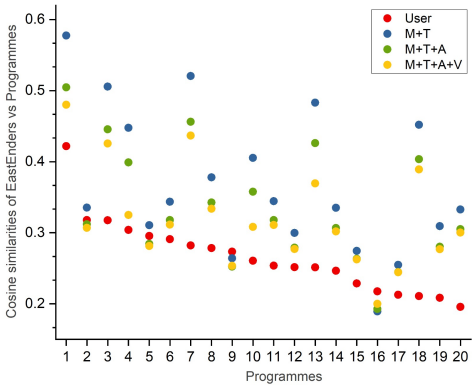

Fig. 2. Effectiveness of adding audiovisual content to textual representations on 20 TV programmes in comparison to EastEnders, the first 10 of these are: Waterloo Road, Outnumbered, Casualty, Uncle, The Voice UK, The truth about webcam girls, Holby City, Sun, Sex & Suspicious Parents, Top Gear, Mrs Brown's Boys.

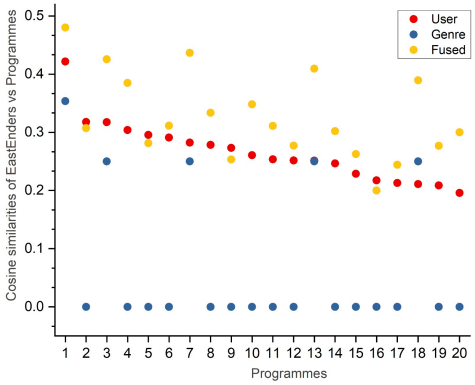

Fig. 3. Comparison of user, genre and fused representations for the same as in Figure2. Even when genres are entirely different from EastEnders, the similarity of fusion based recommendations are relatively higher.

via neural nets for representing content. We only used neural embeddings for textual data; data collections and training of a joint neural network is future work. Making the dataset publicly available is another.

Acknowledgment

We thank BBC R&D's Andrew McParland for discussions, a UCL doctoral scholarship for Nazir, the Royal Academy of Engineering Industrial Scheme Fellowship IF-192058 for Sadrzadeh and, BBC R&D internships for Nazir and Cagail.

References

[1] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González, "Gated multimodal units for information fusion," *arXiv preprint arXiv:1702.01992*, 2017.

[2] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.

[3] T. K. Landauer and S. T. Dumais, "A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge," *Psychological Review*, vol. 104, p. 211–240, 1997.

[4] P. Turney and P. Pantel, "From frequency to meaning: vector space models of semantics," *Journal of Artifical Intelligence Research*, vol. 37, p. 141–188, 2010.

[5] G. C. T. Mikolov, K. Chen and J. Dean, "Efficient estimation of word representations in vector space," in *Proceedings of ICLR*, Scottsdale, AZ, 2013.

[6] N. T. E. Bruni and M. Baroni, "Multimodal distributional semantics," *Journal of Artifical Intelligence Research*, vol. 49, pp. 1–47, 2014.

[7] D. Kiela and S. Clark, "Learning neural audio embeddings for grounding semantics in auditory perception," *Journal of Artificial Intelligence Research*, vol. 60, pp. 1003–1030, 2017.

[8] Q. Zhu, M.-C. Yeh, and K.-T. Cheng, "Multimodal fusion using learned text concepts for image categorization," in *Proceedings of the 14th ACM international conference on Multimedia*, 2006, pp. 211–220.

[9] O. Barkan, N. Koenigstein, E. Yogev, and O. Katz, "Cb2cf: a neural multiview content-to-collaborative filtering model for completely cold item recommendations," in *Proceedings of the 13th ACM Conference on Recommender Systems*, 2019, pp. 228–236.

[10] S. Oramas, O. Nieto, M. Sordo, and X. Serra, "A deep multimodal approach for cold-start music recommendation," in *Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems*, 2017, pp. 32–37.

[11] M. Iqbal, A. Kovac, and K. Aryafar, "A multimodal recommender system for large-scale assortment generation in e-commerce," in *The SIGIR 2018 Workshop On eCommerce co-located with the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2018), Ann Arbor, Michigan, USA, July 12, 2018*, 2018.

[12] B. Yang, T. Mei, X.-S. Hua, L. Yang, S.-Q. Yang, and M. Li, "Online video recommendation based on multimodal fusion and relevance feedback," in *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, ser. CIVR '07. New York, NY, USA: Association for Computing Machinery, 2007, p. 73–80.

[13] H. K. Ekenel and T. Semela, "Multimodal genre classification of tv programs and youtube videos," *Multimedia tools and applications*, vol. 63, no. 2, pp. 547–567, 2013.

[14] K. Bougiatiotis and T. Giannakopoulos, "Enhanced movie content similarity based on textual, auditory and visual information," *Expert Systems with Applications*, vol. 96, pp. 86 – 102, 2018.

[15] C. H. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala, "Latent semantic indexing: A probabilistic analysis," *Journal of Computer and System Sciences*, vol. 61, no. 2, pp. 217–235, 2000.

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

[17] 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.

[18] 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.

[19] A. Plinge, R. Grzeszick, and G. A. Fink, "A bag-of-features approach to acoustic event detection," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 3704–3708.

[20] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2. Ieee, 1999, pp. 1150–1157.

[21] ——, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[22] E. Karami, S. Prasad, and M. Shehata, "Image matching using sift, surf, brief and orb: performance comparison for distorted images," *arXiv preprint arXiv:1710.02726*, 2017.

[23] D. G. Lowe, "Local feature view clustering for 3d object recognition," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1. IEEE, 2001, pp. I–I.

[24] A. A. A. Karim and R. A. Sameer, "Image classification using bag of visual words (bovw)," *Al-Nahrain Journal of Science*, vol. 21, no. 4, pp. 76–82, 2018.

[25] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, vol. 16, no. 6, pp. 345–379, 2010.

[26] MyMediaLite, "Mymedialite recommender system library," http://www.mymedialite.net/, (Accessed on 02/20/2020).