PERSONALISING THE PUBLIC: PERSONALISING LINEAR RADIO AT A PUBLIC SERVICE BROADCASTER

Tim Cowlishaw, Todd Burlington, David Man, Jakub Fiala, Rhiannon Barrington and George Wright

British Broadcasting Corporation, UK

ABSTRACT

With competition from online streaming platforms, many broadcasters have viewed the idea of providing some form of personalisation of their services as a necessity to remain competitive.

However, the idea of introducing personalisation in the context of a public service media organisation presents some unique challenges. This paper presents an outline of the challenges posed by personalised services for PSBs, proposes means of measuring and evaluating them, as well as presenting a novel system for personalised radio services which attempts to address some of these challenges by design.

Our algorithm extends existing approaches, using historic editorial decision making to afford recommendation diversity, as well as automatic explanation and user refinement of decisions taken by the recommendation algorithm.

Finally, we will present the results of an audience-facing evaluation of this system and outline areas for future development of this work.

INTRODUCTION

In recent years, there has been increasing interest in providing personalised services, driven by demand from consumers and increased competition from new entrants to the marketplace (1). As a result, many public service media organisations, including the BBC, have been keen to incorporate personalisation (2) into their services as a response this. There has been renewed interest in the technical challenges associated with personalising audience experiences of broadcast services, most notably the EBU's PEACH project (3).

The personalisation of media services is not without its difficulties. Many have warned that media personalisation presents a challenge to social cohesion and public discourse, notably Pariser (4), who coined the term '*Filter bubble*' to describe the way in which opaquely personalised media services can lead to a partial and biased view of the world on the part of their users. These issues are of particular importance to public broadcast organisations, who are committed to diversity in their output.

We argue that this also presents an opportunity for PSBs to differentiate themselves from their commercial competitors and to provide compelling and novel services with the public interest in mind. This paper presents the results of a project to design a personalised radio service with public service values *embedded* within it. We review the relevant literature to identify the ways in which public service values are challenged by personalised services,

and then identify normative criteria for evaluating how well a personalised service serves public service ends. We describe a prototype developed with these criteria in mind, and present a user-facing evaluation of it, as well as indicating directions for future work.

RELATED WORK

Recommendation / personalisation foundations.

Most approaches to media personalisation are founded in the recommendation systems literature (5), and include *content-based* approaches (6), depending on properties of the recommended items themselves, and *collaborative-filtering* approaches (7), which take a 'wisdom-of-crowds' approach; learning from similarities between the usage habits of different users to produce recommendations. In addition, hybrid approaches have been proposed (8). Our algorithm is one such approach – combining both content metadata and user history to produce recommendations. We make use of matrix-factorisation approaches (9) to the collaborative filtering problem, extending the standard 'Singular Value Decomposition' method to incorporate additional features.

Media personalisation

Recommendation system approaches have been used in audio-visual media almost since their inception – in particular, the Netflix prize (10), a contest to improve the company's film recommendations, resulted in many advances in the field. The sub-field of music recommendation (11) is a very active area of research in both the Music Information Retrieval and Recommender Systems communities. This research has led to innovations in widely-adopted streaming music products such as Spotify (12). Radio broadcasters have begun to explore ways in which they can offer personalised experiences. Casagranda et. al. (13) have proposed a system of 'hybrid content radio' which aims to combine the benefits of both personalised audio services and broadcast radio experiences. While they provide a technical framework for hybrid broadcasting, they do not address the challenge of *how* personalised content can be selected for a given listener, and incorporated into their schedule. We address this challenge, with a particular focus on public service radio broadcasting.

Personalisation and public service

Several authors have provided an overview of the challenges specifically for Public Service Broadcasters of providing personalised services. Scannell (14) argues for the importance of a *shared* experience of public service media, concerns echoed by Sørensen (15) who argues for the role of Public Service Media as a '*social object*' and the importance of editorial curation to public service audiences. Van Es (16) argues that the metricisation involved in providing data-driven personalised services threatens to shift the broadcaster's conception of their audience member from citizen to consumer. Van Es, drawing on the work of Helberger et al (17) identifies a need for diversity in personalised services, as well as stressing the need for tools which allow users to reflect on their consumption habits and the decisions taken based on this data, in order to maintain their status as informed, empowered citizens rather than passive consumers. The themes of diversity, empowerment and transparency inform the values guiding our work in this area.

DISCUSSION OF VALUES AND DEVELOPMENT OF EVALUATION CRITERIA

Based on the literature review, the BBC's public purposes, and indicative feedback from a small panel of audience users interacting with an early prototype, we identified a set of core values which our radio system should embody:

Transparency, Explanation and User feedback

Given that the listener is the sole arbiter of whether a programme delivered by a personalised service is relevant to them, it is important to allow them to feed-back on decisions taken on their behalf in case of an incorrect recommendation. Additionally, it is important to provide users with insight into *how* that recommendation was derived, so they can give specific, meaningful feedback. (In addition, transparency is seen by users as a valuable attribute in recommender systems for their own sake, see (18) for discussion). Our system must allow users to give specific feedback on programme choices, and must also provide explanations of their decisions. Our early prototypes and user testing validated the importance of this for listeners – our panel appreciated the ability to both tailor the recommendations given to them and to be given insight into how those recommendations were decided upon. In our early user testing, participants strongly favoured explanations which were described in terms of properties of the programme *content* (genres, subjects, featured personalities), rather than listener demographics or information about other users listening to that programme.

Diversity

As discussed above, personalisation efforts at a PSB should pay close attention to the diversity of surfaced content. Designing a recommender system to ensure diverse output also has advantages from a user-facing point of view, in terms of maximising recall of relevant programmes (19). In addition, participants in our early research voiced specific concerns about a lack of diversity in existing personalised recommendation systems, citing concerns about 'filter bubbles' as well as the possibility of never seeing potentially relevant content. Diversity in recommendation also affords *serendipity* or *surprise* in encountering new and unusual content which one would not otherwise have encountered, which was an experience frequently cited by participants in our initial survey as being a particularly pleasant aspect of listening to broadcast radio.

MEASUREMENT

Our values of transparency, user agency and diversity described above provide a set of normative principles to guide the development of our personalised service, but crucially, they also offer an insight into how we might measure the success of our system.

In the study of information retrieval and recommender systems, recommender accuracy is measured in terms of *precision* and *recall* – the proportion of returned results which are relevant to the user, and the proportion of relevant items which are returned, respectively. Precision and recall tend to be inversely correlated (20) – one can improve one's precision by sacrificing recall, and vice-versa. Our early research indicating the importance of diversity in our result set, as well as indicating that our users are more concerned about 'missing out' on potentially relevant content than receiving inappropriate recommendations, therefore, suggests that we should seek to maximise the recall of our system as a priority, even if this means sacrificing some precision.

A number of measures for recommendation diversity have been proposed, see Zhou et. al. (21) for an overview. With reference to our recommendation values, two measures in particular seem appropriate – *Surprisal* and *Personalisation*.

The Surprisal metric gives a measure of how novel or unexpected the results of a recommender system are, by calculating the mean self-information over every item in a list of retrieved recommendations:

$$\frac{1}{|items|}\sum_{item\,\in items}\log_2\frac{|users|}{|users_{item}|}$$

This gives a measure of diversity which intuitively corresponds to our need to provide programmes which are unexpected by users, and cover a diverse range of subjects.

By contrast, the *personalisation* metric measures the proportion of recommended items shared by any two users of the system:

$$\frac{1}{|users|^2}\sum_{u_1\in\,users,u_2\in\,users}\frac{\left|items_{u_1}\cap\,items_{u_2}\right|}{k}$$

Where k is the number of recommendations generated for each user. This is an inappropriate measure of diversity for our purposes, as maximising diversity in this case would lead to a heavily individuated service, working against our stated purpose to deliver a shared listening experience between users. However, we can use this metric precisely to measure the degree to which we are providing a shared service. While attempting to simultaneously maximise diversity (as measured by surprisal) and minimise personalisation, we can measure the extent to which both goals are being met.

MODEL DEVELOPMENT

We developed our recommendation model using training data taken from the BBC's iPlayer radio service – anonymised listening logs for a 1% sample of users, who listened to speech radio programmes, taken from a six-month period from January 2017.

Singular Value Decomposition

Initially, we trained a collaborative filtering model on our data, using the *singular value decomposition (SVD)* method of matrix factorization. We construct a $user \times item$ matrix from our training data where each entry corresponds to a binary indicator. The indicator value of '1' indicates the user listened to the entirety of the corresponding programme at least once, and a '0' indicates they did not – we represent this matrix as 'U' in further formulations. We then take the SVD of this matrix, and interpret the resulting values corresponding to each user and item as a 'weight' indicating that user's preference for that item. In order to recover a variable indicating suitability of a programme for a given user, we threshold these weights by comparing them to the threshold parameter t, where any weight $< t$ is assigned to 0 and any weight $> t$ is assigned to 1. The optimal value of t was identified to be 0.2 by cross-validation.

Content metadata based SVD

In order to provide explanations for our recommendations, and promote recommendation diversity, we investigated means of incorporating content metadata into our

recommendation model. The BBC's programme database contains editorially generated programme synopses as free text, entity taggings representing programme subjects, as well as a genre classification taxonomy. These were identified as suitable sources of metadata to incorporate for both user-facing explanatory labels, and to identify latent connections between otherwise unrelated programmes, improving recommendation diversity and serendipity. For each programme, we retrieved editorial metadata and constructed a list of genre labels and programme subject tags applied to each programme. In order to decrease the sparsity of this data, we ran a *named-entity recognition* algorithm over the programme synopsis text, identifying programme subjects and augmenting the editorially provided subject tags (which are inconsistently applied).

From this data, we constructed a second matrix representing incidence of these programme labels across programmes: $label \times programme$. We denote this as 'P' in the formula below.

By combining the data in this matrix with the user-programme indicators in U, we derive a third matrix which relates users to the labels on all the programmes to which they have listened. This is a $user \times label$ matrix we denote as 'L', where each item is the count of programmes listened to by that user which have the corresponding label. This matrix represents an estimate of the listening *interests* of each user, as expressed through the subjects and properties of the programmes to which they have listened.

In order to make recommendations in this new formulation, we can recover a new $user \times programme$ recommendation matrix simply by multiplying L and P. However, while this new matrix LP will successfully recommend programmes based on content metadata associations, it doesn't perform any collaborative filtering in order to learn latent associations between programmes based on the combined information about our audience listening figures, and as a result, significantly underperforms our baseline recommender. To address this, we again apply the singular value decomposition process, but this time only to the matrix L, recovering a lower-rank approximation of L which we denote as L'. Finally, by multiplying L' by P, we recover a new recommendation matrix $L'P$ which provides a weight for each programme for each user. As before, we perform a thresholding process to recover programme relevance indicators for each user.

This new model has the important property that an 'explanation' for each recommended programme can be identified by inspecting the corresponding labels for each programme in the matrix P^{-1}, for which the corresponding entry in L is non-zero. This allows us to present the properties of the programme which were identified as being potentially interesting to them by way of explanation for the recommendation.

With a slight modification, this model can also afford user feedback. We construct a second matrix of the same dimension as L which we denote M, and initialize it to the all-ones matrix. Denoting the elementwise multiplication operation as · we can see that $(L' \cdot M)P$ corresponds exactly to our previous recommendation matrix $L'P$. We may use the matrix M as an indicator matrix which allows users to give feedback on incorrect assumptions – if a user is recommended an item which they do not appreciate, they may indicate, for each label applied to that programme, that it does not interest them, and this data is stored by setting the corresponding entry in M to '0', ensuring this label is not used to make any further recommendations. Our final model affords both explanations about how recommendations are made and allows users to provide explicit feedback on recommendations.

PROTOTYPE DESIGN

We wished to gather insight into how our recommendations were received subjectively, by users, rather than simply assuming that the ability to predict held-out user actions is a good determinant of user satisfaction. In order to do this, we produced a prototype radio player which used our recommendation model to create a personalised, continuous stream of speech radio programmes, which could be operated by participants in a lab trial. This prototype (see Figure 1) was based on the standard iPlayer interface for reasons of user familiarity, but was extended with specific features related to our trial. It continuously plays through recommendations until it is stopped. It displays the recommendation explanations derived from our model, and allows users to refine their recommendations by removing erroneously recommended topics. We also experimented with giving users the ability to skip a programme (but allow it to reoccur later in the stream), and remove a recommended programme from the list entirely.

Figure 1 – Player interface

PROTOTYPE EVALUATION

We evaluated our prototype in one-to-one lab sessions with a panel of 7 regular radio listeners aged 24-32 (we chose to concentrate on younger audiences specifically due to the strategic goals of this project), 4 female and 3 male, with a range of social backgrounds. Before the lab session, each participant was asked to fill out a short survey of their radio listening habits, and previous programmes they had listened to, in order to provide seed data for their recommendations.

Each lab session lasted an hour and consisted of three phases – Initially the participant was guided through the interface by the facilitator, and asked to give their impressions of how they expected each element to behave without interacting with it. They were then left alone for 20 minutes to listen to their radio stream and interact with the player however

they chose, before the facilitator returned to interview them about their experience. Each programme in the player was truncated to a five-minute clip, in order to ensure that each participant would hear multiple programmes during their session. Finally, each participant was asked to provide relevance judgements for twenty programmes in their stream, choosing from four options ("I would be interested in listening to this programme", "I would not be interested in listening to this programme" and "I have already listened to this programme", as well as a "I cannot say" option for which judgements were discarded).

The relevance judgement exercise was also completed by 7 colleagues from our department, none of whom were involved with the project, for a total of n=14 (m=8, f=6). The additional seven participants were asked to provide relevance judgements for the top 20 most popular programmes on iPlayer in addition to their personalised programmes, in order to serve as a baseline for comparison. The 20 programmes were randomised in the questionnaire so participants did not know that any of the 40 programmes they judged were not recommended for them personally.

PROTOTYPE RESULTS

Benchmarks

Evaluating our model against a held-out test set of user data from the same period gave a precision score of 0.32 and recall of 0.27, with surprisal of 9.60 and personalisation of 0.94. This compares to a baseline score for our standard SVD recommender of precision of 0.51 and recall of 0.51, so some accuracy is lost in order to achieve explainability. The surprisal score is vastly improved from a baseline of 2.96 indicating significantly increased recommendation diversity, however the personalisation score is higher than the baseline of 0.82, suggesting that our service is more, rather than less, individuated than the baseline approach.

In addition, while the precision and recall of our model are reduced relative to baseline, they still outperform previous published work on the iPlayer dataset (22).

Relevance judgements

In the subjective relevance test, we defined a 'hit' as a response of either "I would be interested in listening to this programme" or "I have already listened to this programme", and a 'miss' as a response of "I would not be interested in listening to this programme", discarding all "Other / Can't say" responses. This yielded precision of 0.68 on our recommendation model, compared to a precision of 0.39 on the baseline 'most popular' recommendations. Performing a χ^2 test on these results yields a p-value of <0.001 ($\chi^2=21.17$, 2df) indicating a significant improvement over baseline.

Qualitative results

Our lab tests and interviews validated the overall approach we took to the project, but presented some challenges for specific aspects of our prototype. The key findings were as follows:

Participants appreciated receiving a diverse range of recommendations, and were concerned about personalisation limiting the range of content they consume. Many expressed frustration at a lack of diversity in the *ordering* of their recommendations – in particular, if similar programmes appeared alongside each other within an otherwise diverse list.

Participants appreciated the ability to tailor their recommendations themselves, but they expressed a desire to provide both positive *and* negative feedback, rather than just correcting the recommender when it went wrong. Several participants were concerned that the choices offered were too 'binary', and that removing a topic would limit their future recommendations in a way that they found too restrictive.

In general, participants appreciated the provision of explanations about why their recommendations had been chosen, both out of curiosity about how the recommender worked, and in order to discover why surprising recommendations had been provided.

CONCLUSIONS AND FURTHER WORK

We have outlined a series of design principles for personalisation of public service linear radio, including a commitment to explain decisions on behalf of audience members, to accept feedback to further tailor their recommendations, to provide a diverse range of programming and to preserve a shared listening experience, as well as proposing means of measuring the ability of a system to meet these principles. We have proposed a novel recommendation model and a corresponding prototype personalised radio system which is designed to meet these criteria. We have presented the results of both quantitative and qualitative evaluation of the model and prototype.

The evaluation showed that audience members within our target appreciate the provision of a diverse personalised service, as well as the ability to further customise it, and to understand why recommendations are made. It showed that editorially curated metadata can be used effectively to further these goals, providing the means to explain decisions and promote recommendation diversity. However, our current model sacrifices accuracy relative to standard approaches in order to do this, so further work is needed to ameliorate this.

Based on feedback from our qualitative work it appears that effective programme sequencing is paramount to the success of a continuous play personalised radio. We plan to investigate means of learning from the editorially-curated decisions comprising the history of our broadcast schedule, in order to produce an editorially inspired *sequence model* for ordering recommendations.

BIBLIOGRAPHY

1. Deloitte, 2015. Made-to-order: The rise of mass personalisation. Retrieved from: https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/consumer-business/deloitte-uk-consumer-review-mass-personalisation.pdf, Accessed 3rd May 2018.

2. BBC, 2017: Tony Hall's speech at the launch of the Annual plan for 2017/18. Retrieved from: http://www.bbc.co.uk/mediacentre/speeches/2017/tony-hall-annual-plan#heading-a-personalised-uniquely-tailored-bbc, Accessed 3rd May 2018.

3. EBU, 2017: Personalisation for EACH. Retrieved from: https://peach.ebu.io/, Accessed 3rd May 2018.

4. Pariser, E. 2011. The Filter Bubble: What the Internet is Hiding from You, Penguin, UK

5. Ricci, F, Rokach, L, Shapira, B & Kantor, P B, editors, 2011. Recommender Systems Handbook. Springer

6. van Meteren, R & van Someren, M, 2000. Using Content-Based Filtering for Recommendation. Proceedings of ECML 2000 Workshop: Machine Learning in Information Age, pp 47 to 56

7. Resnick, P, Iacovou, N, Suchak, M, Bergstrom, P & Riedel, J, 1994. GroupLens: an open architecture for collaborative filtering of netnews. Proceedings of the 1994 ACM conference on computer supported cooperative work, pp 175 to 186

8. Balabanović, M & Shoham, Y, 1997. Fab: content-based, collaborative recommendation. Communications of the ACM, Volume 40, Issue 3. March 1997. pp 66 to 72

9. Koren, Y, Bell, R and Volinsky, C, 2009. Matrix Factorization Techniques for Recommender Systems. Computer, Volume 42, Issue 8. August 2009.

10. Bennett, J & Lanning, S, 2007. The Netflix Prize. Proceedings of KDD Cup and Workshop. August 2007.

11. Celma, O, 2010. Music Recommendation. Springer.

12. Diaz, F. 2017. Spotify: Music Access At Scale. Proceedings of SIGIR 2017. August 2017. pp 1349 – 1349.

13. Casagranda, P, Erk, A, O'Halpin, S, Born, D, Huijten, W. 2015. Proceedings of the International Broadcasting Convention, 2015, Amsterdam.

14. Scannell, P. 2005. The Meaning of Broadcasting in the Digital Era. Cultural Dilemmas in Public Service Broadcasting (RIPE @ 2005), Nordicom.

15. Sørensen, J K. 2013. Public service broadcasting goes personal: The failure of personalised PSB web pages. MedieKultur 2013, 55. pp 43 to 71.

16. Van Es, K. 2017. An Impending Crisis of Imagination: Data-Driven Personalization in Public Service Broadcasters. Media@LSE Working Paper Series.

17. Helberger, N, Karppinen, K, D'Acunto, L. 2018. Exposure diversity as a design principle for recommender systems. Information, Communication & Society, 21:2 pp 191 to 207.

18. Sinha, R, Swearingen, K. 2002. The role of transparency in recommender systems. CHI '02 Extended Abstracts on Human Factors in Computing Systems pp 830 – 831.

19. Clarke, C L A, Kolla, M, Cormack, G V, Vechtomova, O, Ashkan A, Büttcher, S, MacKinnon, I. 2008. Novelty and diversity in information retrieval evaluation. Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. pp 659 – 666.

20. Buckland, M, Gey, F. 1994. The Relationship between Recall and Precision. Journal of the American Society for Information Science 45 (1). Jan 1994. pp 12-19.

21. Zhou, T Juscik, Z, Liu, J-G, Medo, M, Wakeling, J R, Zhang, Y-C. 2010. Solving the apparent diversity-accuracy dilemma of recommender systems. Proceedings of the National Academy of Sciences 107 (10). March 1010. pp 4511-4515.

22. Paudel, B, Christoffel, F, Newell, C, Bernstein, A. 2017. Updatable, Accurate, Diverse and Scalable Recommendations for Interactive Applications. ACM Transactions on Interactive Intelligent Systems (TiiS) 7 (1). March 2017.