On Building a Podcast Collection with User Interactions

FRANCESCO MEGGETTO, NeuraSearch Laboratory, University of Strathclyde, UK
YASHAR MOSHFEGHI, NeuraSearch Laboratory, University of Strathclyde, UK
ROSIE JONES, Spotify, USA

The podcast is a growing listening medium that has surged in popularity in recent years. Despite the great research opportunities, it has only attracted limited attention from the community so far. This is mainly due to the lack of available data collections that have considerably restricted research in academia. To facilitate it, in 2020, the Spotify Podcast Dataset was released, a corpus of 100k episodes with associated text transcript and metadata. However, no user interactions are available, hence making its usability challenging for certain domains, such as recommendation, personalisation, and user behaviour and consumption analysis. In this position paper, we present various approaches to augment such collection with user interactions, together with their respective strengths and weaknesses. If developed further, this work has the potential of a broader impact on the research community.

Additional Key Words and Phrases: Spotify; Podcast; User Study; User Behaviour

ACM Reference Format:
Francesco Meggetto, Yashar Moshfeghi, and Rosie Jones. 2021. On Building a Podcast Collection with User Interactions. 1, 1 (September 2021), 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Online audio streaming services have a long-lasting connection with the music industry, which has been their main pivotal point for decades. In recent years, a new audio listening medium, i.e. podcasts, has started a rapidly growing process and has swiftly become, although largely under-researched, an essential part of listening habits. As of 2021, there are more than 1 million active podcasts and over 30 million episodes in over 100 languages. They have been sharply rising in popularity such that in the US alone, 75% of the entire population is familiar with the term "podcast", 55% has have listened to one, and 37% are monthly listeners [16].

To foster research in this domain, the Spotify Podcast Dataset [5] was recently released, a large corpus of episodes consisting of an audio file, automatically transcribed text via Google's Cloud Speech-to-Text APIs, and associated metadata. The release was in conjunction with the new TREC Podcast Track [7], where two shared tasks were released: retrieval of fixed two-minute segments and episodes summarisation. Although its great applicability to various tasks in fields such as speech and audio processing, NLP, IR, and computational linguistics, it is not suitable for those where logged user behaviour is required. This is the case of analyses in user information needs, their characteristics and behaviour, relevance, as well as recommendation and personalisation systems. In this position paper, we present and

Authors' addresses: Francesco Meggetto, NeuraSearch Laboratory, University of Strathclyde, Glasgow, UK, francesco.meggetto@strath.ac.uk; Yashar Moshfeghi, NeuraSearch Laboratory, University of Strathclyde, Glasgow, UK, yashar.moshfeghi@strath.ac.uk; Rosie Jones, Spotify, Boston, MA, USA, rjones@spotify.com.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2021 Association for Computing Machinery.
Manuscript submitted to ACM

discuss various approaches with the aim to supplement the existing Podcast Dataset with users interactions, thus working towards creating podcast listening sessions.

The rest of this paper is structured as follows. Section 2 discusses related work in music, podcast, and their intersection. Section 3 outlines the approach of this work and ideas around participants selection and corpus creation. In Section 4 a discussion is presented. Finally, Sections 5 concludes the paper with final remarks on the need for a community effort in the building of this work.

2 RELATED WORK

Research in podcasts is relatively scarce compared to the more established field of music. In the latter, it has been shown that Music Recommendation Systems have unique characteristics in comparison with other domains such as movies: consumption time of the media items, a single item may be consumed repeatedly, and recommendations can be based on groupings by genre, artist, or album [15]. Given the sparsity of user feedback and also rarity of explicit feedback data, audio content analysis [11], contextual features [6] and user interactions [12] are often drawn for improving the recommendation strategy. Academic advances in the field have seen a greater influx with the recent releases of the Million Playlist [4] and the Music Streaming Sessions Dataset [2].

Podcasts are spoken documents, and they differ significantly from other spoken document corpora. In particular, category labels (given by podcast creators) are unreliable, episodes are between half an hour and an hour in length, and the range of style, format, language variety, and variations make their analysis complex and different from any other media type previously analysed [8]. These factors also pose multiple problems for podcast search, at present done via catalogue match using show titles, episode titles, and sometimes metadata provided by podcast creators [3]. Podcasts are substantially different from music. For example, the number of tracks an active music listener typically listens to during the day is much higher than for podcasts, and listening to podcast is also a more significant time investment because of its lengthy nature. Finally, the action of re-listening to a previously listened podcast episode is not as common as re-listening a song. Li et al. [9] state how the functional use of podcasts and music may largely overlap for many people and show how podcast and music consumption compete slightly but do not replace one another. Moreover, users who are both podcast and music listeners demonstrate significantly different consumption patterns in terms of streaming time, duration, and frequency. Despite these major differences with music, Nazari et al. [10] successfully demonstrated how music preferences could be used in a cross-domain setting to address the cold-start problem in podcast recommendations. This is motivated by the fact that many platform users turn to podcasts from music, usually with long prior listening histories that can therefore be leveraged for a cross-domain recommendation. Recently, Benton et al. [1] suggested that podcast listening is sequential and that incorporating as much available information as possible is crucial to building an accurate and successful recommender system. Additionally, in their work, they also show that user behaviour is confined to local trends and that listening patterns tend to be found over short sequences of similar types of shows. In [13, 14] it was found how podcast listening activity in the function of the episode duration is affected by ads and other extraneous material and that linguistic style and textual attributes have an effect on user engagement. Lastly, Yang et al. [17] presented Adversarial Learning-based Podcast Representation (ALPR) to capture non-textual aspects of podcasts. Their model could predict the seriousness, energy, and popularity of podcasts, also showing that the latter can be well predicted within the first five minutes of audio data.

Overall, there is a high level of interest in the emerging podcast domain, with a wide range of opportunities. However, the fundamental lack of availability of user interactions data of podcast listening is limiting academic research, which

we aim to address in the foreseeable future. For a more complete and in-depth review of current challenges and future directions in Podcast Information Access, we refer the readers to [8].

3 APPROACH

The aim of this study is to present different approaches to building and ultimately release a new collection for academic research. It is to be considered as an augmentation of the existing Spotify Podcast Dataset, where, as previously noted, no user interactions data are provided. A successful creation will open up a wide range of research opportunities, including analyses on information needs, users' characteristics and behaviour, and relevance. What follows outlines the various approaches that were identified for the selection of users and then tasked to monitor and collect behavioural data.

3.1 Participants Selection

3.1.1 ***User Simulation.*** The first idea we are going to discuss is statistical analysis and simulation of users' behaviour. Prior work by Yang et al. [17] shows that podcast representation can be enriched by modelling non-textual characteristics. Their model achieved significantly better performance in the task of predicting seriousness, energy, and popularity. Based on this work, we plan to expand the non-textual analysis to characterise and model other properties of podcast content. Identification of other features in the podcast space would open to better shaping of podcasts and thus more effective personalised recommendations. To integrate user behaviour, a user simulation is required, with behaviour driven by statistical models exploiting the most recent findings in users' listening patterns and consumption. However, current research is scarce and not comprehensive. User's interest in a show or episode relies on implicit signals such as subscribe, play, pause, stop. Their identification and characterisation need further investigation, hence making users' simulation not a momentarily straightforward or unbiased process. With automatic inference of users' interaction history not possible with the current research, a better approach may consist of user-based studies, as presented in the next sections.

3.1.2 ***Laboratory-based User Study.*** We perform a laboratory-based user study to collect user interactions from users. A laboratory-based user study is inherently at a small scale, but it allows for high experimental control. In particular, participants can be well instructed in various tasks and scenarios that may be required for dynamic monitoring and modelling of their behaviour. A controlled environment may allow us to truly exploit real user's behaviour, hence making this experiment of great appeal, but at the cost of a greatly reduced number of participants and higher time requirements.

3.1.3 ***Crowdsourcing-based User Study.*** The last feasible option we ought to consider is a crowdsourcing-based user study. Participants can be recruited either on crowdsourcing platforms, such as Amazon Mechanical Turk (MTurk) or on selected academic communities. Despite the former would allow us to collect a high amount of data and, in a short amount of time from recruited workers, it is an unfeasible option given the associated costs. Further, the platform is designed for task-based work and not for research studies. It may lead to low-quality data since workers may not perform the task and interact with the system as accurately as in a real-world setting. Generated podcast listening sessions may therefore contain high levels of interaction noise, and future conclusions or deductions based on this collection may be unreliable and/or inaccurate. In other words, an untruthful reflection of users' information needs.

Instead, we believe a more effective study is performed on a selected crowd that would then enable us to fulfil our objective of producing a new collection similar to Music Streaming Sessions Dataset [2], which is an extremely large

collection consisting of approximately 150 million logged music streaming sessions. We hope to gather an IR community shared effort and contribution in the building of this work.

3.2 Building the Corpus

In this section, we analyse possible mechanisms and tasks that can be used for collecting user behaviour. Irrespective of the selected source of participants, we expect them to interact with an ad hoc system with functionalities and facades similar to modern music streaming platforms. This is for a high familiarity with the system. The *how* participants can navigate through the system, select episodes, and listen to them can be categorised as follows:

- **Search.** Participants can search and listen to a particular episode. A ranked list of indexed episodes may be returned by the system according to their search query.
- **Recommendation**. Instead of participants searching for episodes, they are recommended a set of episodes to listen to for a more dynamic user experience.

Search allows recruited participants a higher control over the content they would like to listen to since they can search and freely choose the content to stream from a large collection of available podcasts. On the other hand, a *recommendation* system exploits user's interests (annotated as part of a questionnaire prior to the experiment's start) and possibly past interactions to suggest episodes they may be enjoying. In this experimental setting, recommendations can be personalised to target specific user groups to monitor their behaviour under a multitude of different scenarios.

Regardless of the mechanism, participants will be given a task of relevance judgement after each listening. Single or pair judgement relevance are two viable alternatives. The first is well-suited for both mechanisms, with an explicit rating on a scale of 1-5 given to each episode. In the latter, a pair of two episodes is displayed to the participant and, after listening to both, they are asked for an evaluation and to provide their preference. The main difference between the two tasks is that with a side-to-side judgement, the system explicitly asks users to give preference to two episodes, which can be manually picked according to certain conditions. Therefore, it creates the opportunity to test various hypotheses that may emerge in user analysis, such as the impact various facets have on episode relevance. Examples may include linguistic style, speakers' tone and rhythm, and acoustic changes. On the other hand, a single-item rating provides a more meaningful and direct evaluation of what participants deem more relevant and less relevant in an episode.

4 DISCUSSION

With the likely impracticability of user simulation, the remaining two types of user studies appear to be the most feasible. Recruitment of workers for crowdsourcing is a fast but unreliable and expensive technique, with a laboratory-based study at the opposite end of the spectrum. We believe a reliable collection of users' interactions and its rapid availability to be desirable characteristics given the high momentum of research in podcast. Therefore, we hypothesise a crowdsourcing-based user study on a research community, such as the Special Interest Group on Information Retrieval (SIGIR) to have the highest potential reward in terms of practicability and expressiveness of the data collected. Moreover, it has the potential to truly exploit real user's behaviour, hence making this type of experiment of greater interest for future research.

The search and recommendation strategies are to be carefully conceived, and they require further investigation and discussion. For example, despite search allowing high customisation of results from participants, a recommendation system can exploit the similarity between users' interests and episodes. This is to provide an initial ground to understand

and model their information needs. However, it adds extra layers of difficulties that may hamper the practicability and final effectiveness of this work.

Finally, an important factor to keep into consideration when designing the experiment is the lengthy nature of podcast episodes. To maintain a high level of interest in the study, only small sections of episodes should be available for listening. We propose to use a sampling strategy to reduce the number of podcasts by creating a smaller set. Important to note is the fact that there are major differences across episodes in terms of length, topic, number of speakers, style, etc. Therefore, it is crucial to accurately design the strategy to account for these factors and to not introduce various biases, such as popularity during the process.

5 CONCLUSION

In this position paper, we presented various approaches with the aim of extending the existing Spotify Podcast Dataset with user interactions data in the format of podcast listening sessions. It is a known limitation that is impairing its adaptability to those domains where this type of data is required. To recruit participants, we explored the following three study types: (i) statistical analyses to simulate users' behaviour, (ii) crowdsourcing-based user study, and (iii) laboratory-based user study. Then, irrespective of the selected framework for recruitment, an ad hoc system with functionalities and facade similar to a modern music streaming platform is used as the platform for conducting the experiment. Search, recommendation, or their combination have been identified as possible ways participants can use to navigate and interact with the system. Last, after listening to small segments of selected episodes, participants are asked to judge the relevance of the content either singularly or in pairs. With a discussion on the strengths and weaknesses of all presented ideas, we believe a community effort is required for a fruitful and insightful discussion on the best strategy to achieve the creation of a new corpus. We also believe in its great potential for fostering future research in podcast recommendation, personalisation, as well as user satisfaction and consumption analysis. Overall, the insights that can be gained from building such a collection could have a broader impact on the community.

REFERENCES

[1] Greg Benton, Ghazal Fazelnia, Alice Wang, and Ben Carterette. 2020. Trajectory based podcast recommendation. *arXiv preprint arXiv:2009.03859* (2020).

[2] Brian Brost, Rishabh Mehrotra, and Tristan Jehan. 2019. The music streaming sessions dataset. In *The World Wide Web Conference*. 2594–2600.

[3] Ben Carterette, Rosie Jones, Gareth F Jones, Maria Eskevich, Sravana Reddy, Ann Clifton, Yongze Yu, Jussi Karlgren, and Ian Soboroff. 2021. Podcast metadata and content: Episode relevance and attractiveness in ad hoc search. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2247–2251.

[4] Ching-Wei Chen, Paul Lamere, Markus Schedl, and Hamed Zamani. 2018. Recsys challenge 2018: Automatic music playlist continuation. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 527–528.

[5] Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, et al. 2020. 100,000 podcasts: A spoken English document corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*. 5903–5917.

[6] Casper Hansen, Christian Hansen, Lucas Maystre, Rishabh Mehrotra, Brian Brost, Federico Tomasi, and Mounia Lalmas. 2020. Contextual and sequential user embeddings for large-scale music recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 53–62.

[7] Rosie Jones, Ben Carterette, Ann Clifton, Maria Eskevich, Gareth J. F. Jones, Jussi Karlgren, Aasish Pappu, Sravana Reddy, and Yongze Yu. 2021. TREC 2020 Podcasts Track Overview. *arXiv:2103.15953 [cs]* (March 2021). http://arxiv.org/abs/2103.15953 arXiv: 2103.15953.

[8] Rosie Jones, Hamed Zamani, Markus Schedl, Ching-Wei Chen, Sravana Reddy, Ann Clifton, Jussi Karlgren, Helia Hashemi, Aasish Pappu, Zahra Nazari, et al. 2021. Current challenges and future directions in podcast information access. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1554–1565.

[9] Ang Li, Alice Wang, Zahra Nazari, Praveen Chandar, and Benjamin Carterette. 2020. Do podcasts and music compete with one another? Understanding users' audio streaming habits. In *Proceedings of the web conference 2020*. 1920–1931.

[10] Zahra Nazari, Christophe Charbuillet, Johan Pages, Martin Laurent, Denis Charrier, Briana Vecchione, and Ben Carterette. 2020. Recommending podcasts for cold-start users based on music listening and taste. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and*

Development in Information Retrieval. 1041–1050.

[11] Aaron Ng and Rishabh Mehrotra. 2020. Investigating the Impact of Audio States & Transitions for Track Sequencing in Music Streaming Sessions. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 697–702.

[12] Bruno L Pereira, Alberto Ueda, Gustavo Penha, Rodrygo LT Santos, and Nivio Ziviani. 2019. Online learning to rank for sequential music recommendation. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 237–245.

[13] Sravana Reddy, Mariya Lazarova, Yongze Yu, and Rosie Jones. 2021. Modeling Language Usage and Listener Engagement in Podcasts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 632–643.

[14] Sravana Reddy, Yongze Yu, Aasish Pappu, Aswin Sivaraman, Rezvaneh Rezapour, and Rosie Jones. 2021. Detecting Extraneous Content in Podcasts. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 1166–1173.

[15] Markus Schedl, Peter Knees, Brian McFee, Dmitry Bogdanov, and Marius Kaminskas. 2015. Music recommender systems. *Recommender systems handbook* (2015), 453–492.

[16] Gavin Whitner. 2021. *The Meteoric Rise of Podcasting*. Retrieved Jul 30, 2021 from https://musicoomph.com/podcast-statistics/

[17] Longqi Yang, Yu Wang, Drew Dunne, Michael Sobolev, Mor Naaman, and Deborah Estrin. 2019. More than just words: Modeling non-textual characteristics of podcasts. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 276–284.