The Music Streaming Sessions Dataset

Brian Brost, Rishabh Mehrotra & Tristan Jehan
Spotify Research
{brianbrost,rishabhm,tjehan}@spotify.com

arXiv:1901.09851v1 [cs.IR] 31 Dec 2018

ABSTRACT

At the core of many important machine learning problems faced by online streaming services is a need to model how users interact with the content. These problems can often be reduced to a combination of 1) sequentially recommending items to the user, and 2) exploiting the user's interactions with the items as feedback for the machine learning model. Unfortunately, there are no public datasets currently available that enable researchers to explore this topic. In order to spur that research, we release the Music Streaming Sessions Dataset (MSSD), which consists of approximately 150 million listening sessions and associated user actions. Furthermore, we provide audio features and metadata for the approximately 3.7 million unique tracks referred to in the logs. This is the largest collection of such track metadata currently available to the public. This dataset enables research on important problems including how to model user listening and interaction behaviour in streaming, as well as Music Information Retrieval (MIR), and session-based sequential recommendations.

1 INTRODUCTION

A long-standing and central challenge of online services is to understand and model how users behave [14]. For web search and online advertising, that led to a large body of work on click modeling, most of which would not have been possible without access to publicly available click logs [6]. Modeling user behaviour is similarly important to streaming services [13, 15], but to the best of our knowledge there are no user interaction datasets currently available to the public. This is particularly limiting when it comes to designing recommender systems, where the use of implicit feedback is often critical [10].

Motivated by the paucity of user interaction logs in streaming, we release the Music Streaming Sessions Dataset (MSSD), which consists of over 150 million listening sessions with associated user interaction information. In addition, we provide audio features and metadata for the approximately 3.7 million unique tracks referred to in the logs, making this the largest collection of such track metadata currently available to the public.

One of the main ambitions of this dataset release is to enable public research on two central challenges facing online streaming services, namely predicting if a user will 1) skip a track, and 2) move from one listening context to another. In addition to these problems, the dataset enables research on other challenging machine learning problems. Finally, sequentially recommending items for users is of particular importance to music streaming services but also to other types of services, such as news or e-commerce, for example in next-basket recommendation. We believe the underlying solutions for these cases would be comparable to those for music.

2 RELATED WORK

Click log releases have played a key role in allowing the development of sophisticated click models for web search and advertising applications, with companies including Microsoft, Yahoo, Yandex, and Criteo all providing such logs to the academic research community. An overview of the click modeling literature, and available datasets is provided in [6].

In the context of recommender systems there are fewer such interaction logs available. Instead, recommender systems datasets have tended to contain explicit ratings [2, 7, 8]. Some streaming logs containing user interactions have been released, for example the XING dataset which was part of the 2016 Recsys Challenge [1]. However, this dataset was in the field of job recommendations and is no longer publicly available.

There are several music listening histories datasets, most of which were crawled from from social media such as last.fm or twitter, however none of these datasets contain user interactions beyond what songs were listened to [4, 9, 11–13]. Similarly, the kkbox dataset from the wsdm challenge provides listening logs [5], but these logs are not timestamped, nor do they provide user interactions beyond what tracks were listened to. Thus, although there are some music listening logs, crucially none of these contain information about how users interacted with the tracks they listened to, showing instead only what tracks users were exposed to, as opposed to showing how much, and how, they actually listened to those tracks.

3 THE MUSIC STREAMING SESSIONS DATASET

The MSSD consists of approximately 150 million streaming sessions with associated user interactions, audio features and metadata describing the tracks streamed during the sessions, and snapshots of the playlists listened to during the sessions. The dataset is hosted at https://www.crowdai.org/challenges/spotify-sequential-skip-prediction

The streaming sessions are stored in a log, where each row of the log contains a session id, a timestamp, contextual information about the stream, the track and context id's, and the timing and type of user interaction within the stream. A schema for the log is provided in Table 1. Each session is defined to be a period of listening with no more than 60 seconds of inactivity between consecutive tracks. Additionally for this dataset we set a cut off of at most 20 tracks per session. Sessions included in the dataset are sampled uniformly at random from listening sessions on the contexts included in the dataset over an 8 week period. We exclude sessions that include tracks which did not meet a minimum popularity threshold.

The sessions in this dataset are sampled from radio, personalized recommendation mixes, the user's own collections, and 100 of the most popular playlists on a major music streaming service. The

Column name	Column description	Example value
session id	unique session identifier	65_283174c5-551c-4c1b-954b-cb60ffcc2aec
session position	position of row within session	4
session length	number of rows in session	11
track id	unique track identifier	t_13d34e4b-dc9b-4535-963d-419afa8332ec
context id	unique context identifier	124321515
skip_1	boolean indicating if the track was only played very briefly	1
skip_2	boolean indicating if the track was only played briefly	1
skip_3	boolean indicating if most of the track was played	0
not_skipped	boolean indicating that the track was played in its entirety	0
context switch	boolean indicating if the user changed context between rows	0
no_pause_before_play	boolean indicating if there was no pause between current and previous playback	1
short_pause_before_play	boolean indicating if there was a short pause between current and previous playback	1
long_pause_before_play	boolean indicating if there was a long pause between current and previous playback	0
n_seekfwd	number of times the user did a seek forward within track	2
n_seekback	number of times the user did a seek back within track	0
shuffle	boolean indicating if the user encountered this track while shuffle mode was activated	0
hour_of_day	hour of day between 0 and 23	22
date	date in YYYY-MM-DD format	2018-09-18
premium	boolean indicating if the user was on premium or not	1
context_type	what type of context the playback occurred within	editorial playlist
reason_start	the user action which led to the current track being played	fwdbtn
reason_end	the user action which led to the current track playback ending	trackdone

Table 1: Schema for the interaction log

track id	duration	release year	us popularity estimate
acousticness	beat strength	bounciness	danceability
dyn range mean	energy	flatness	instrumentalness
key	liveness	loudness	mechanism
mode	organism	speechiness	tempo
time signature	valence	acoustic vector	

Table 2: Features for the track metadata

logs therefore contain a mix of listening sessions based on user's personally curated collections; expertly curated playlists; contextual, but non-personalized recommendations; and finally, personalized recommendations.

For each track contained in the sessions, we provide audio features and metadata describing the track. We provide approximately 3.7 million tracks. These include features like acousticness, a measure of confidence that the track is acoustic, runnability, an estimate of how suitable the track is as running music, and valence, a measure of how positive a track sounds. The schema for the track metadata is provided in Table 2. More detailed descriptions of the features are available in [3] and on our dataset website.

4 RESEARCH PROBLEMS

We contend that this sessions dataset consisting of listening and activity data would would help foster research in a number of directions. A brief list of few key research areas for which we hope this dataset will be useful include:

(1) Skip and context switch prediction
(2) Session based Sequential Recommendations
(3) User Intervention in Automated Systems
(4) Offline Evaluation of Recommender Systems
(5) User Journeys
(6) Proactive Recommendations

5 CONCLUSION

The problems of understanding, modeling and predicting how users interact with content on streaming services has until now been understudied, mainly because of a lack of access to data. With the paper, we provide the only dataset of streaming logs and user interactions currently available to research communities.

Furthermore, we identify important research questions that can be addressed using the dataset. Of particular importance are the tasks of skip prediction, context switch prediction, and session-based sequential recommendations. Such online recommendations are particularly important for streaming services when on-boarding new users.

While the dataset enables new types research directions, compromises were made for privacy and commercial reasons. An interesting future extension of this dataset would be to provide more precision on the skip information than the currently bucketed times. That level of information could allow to predict moments in the track where skips are most likely to occur, which could be of great value for generative and interactive music models.

REFERENCES

[1] Fabian Abel, András Benczúr, Daniel Kohlsdorf, Martha Larson, and Róbert Pálovics. Recsys challenge 2016: Job recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 425–426. ACM, 2016.

[2] Robert M Bell and Yehuda Koren. Lessons from the netflix prize challenge. *Acm Sigkdd Explorations Newsletter*, 9(2):75–79, 2007.

[3] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Ismir*, volume 2, page 10, 2011.

[4] O. Celma. *Music Recommendation and Discovery in the Long Tail*. Springer, 2010.

[5] Yian Chen, Xing Xie, Shou-De Lin, and Arden Chiu. Wsdm cup 2018: Music recommendation and churn prediction. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 8–9. ACM, 2018.

[6] Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. Click models for web search. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 7(3):1–115, 2015.

[7] Gideon Dror, Noam Koenigstein, Yehuda Koren, and Markus Weimer. The yahoo! music dataset and kdd-cupfi11. In *Proceedings of KDD Cup 2011*, pages 3–18, 2012.

[8] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(4):19, 2016.

[9] David Hauger, Markus Schedl, Andrej Košir, and Marko Tkalcic. The million musical tweets dataset: what can we learn from microblogs. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (IS-MIRfi2013)*, pages 189–194, 2013.

[10] Douglas W Oard, Jinmook Kim, et al. Implicit feedback for recommender systems. In *Proceedings of the AAAI workshop on recommender systems*, volume 83. WoUongong, 1998.

[11] Markus Schedl. Leveraging microblogs for spatiotemporal music information retrieval. In *European Conference on Information Retrieval*, pages 796–799. Springer, 2013.

[12] Markus Schedl. The lfm-1b dataset for music retrieval and recommendation. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 103–110. ACM, 2016.

[13] Gabriel Vigliensoni and Ichiro Fujinaga. The music listening histories dataset. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIRfi2017)*, 2017.

[14] Geoffrey I Webb, Michael J Pazzani, and Daniel Billsus. Machine learning for user modeling. *User modeling and user-adapted interaction*, 11(1-2):19–29, 2001.

[15] Hongliang Yu, Dongdong Zheng, Ben Y Zhao, and Weimin Zheng. Understanding user behavior in large-scale video-on-demand systems. In *ACM SIGOPS Operating Systems Review*, volume 40, pages 333–344. ACM, 2006.