RecSys Challenge 2018: Automatic Music Playlist Continuation

Ching-Wei Chen
Spotify
New York, USA
cw@spotify.com

Paul Lamere
Spotify
New York, USA
paul@spotify.com

Markus Schedl
Johannes Kepler University
Linz, Austria
markus.schedl@jku.at

Hamed Zamani
University of Massachusetts
Amherst, USA
zamani@cs.umass.edu

ABSTRACT

The ACM Recommender Systems Challenge 2018 focused on automatic music playlist continuation, which is a form of the more general task of sequential recommendation. Given a playlist of arbitrary length, the challenge was to recommend up to 500 tracks that fit the target characteristics of the original playlist. For the Challenge, Spotify released a dataset of one million user-created playlists, along with associated metadata. Participants could submit their approaches in two tracks, i.e., main and creative tracks, where the former allowed teams to use solely the provided dataset and the latter allowed them to exploit publicly available external data too. In total, 113 teams submitted 1,228 runs in the main track; 33 teams submitted 239 runs in the creative track. The highest performing team in the main track achieved an R-precision of 0.2241, an NDCG of 0.3946, and an average number of recommended songs clicks of 1.784. In the creative track, an R-precision of 0.2233, an NDCG of 0.3939, and a click rate of 1.785 was realized by the best team.

KEYWORDS

Recommender Systems; Automatic Playlist Continuation; Music Recommendation Systems; Dataset; Challenge; Benchmark; Evaluation

ACM Reference Format:
Ching-Wei Chen, Paul Lamere, Markus Schedl, and Hamed Zamani. 2018. RecSys Challenge 2018: Automatic Music Playlist Continuation. In *Twelfth ACM Conference on Recommender Systems (RecSys '18), October 2–7, 2018, Vancouver, BC, Canada.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3240323.3240342

1 TASK DESCRIPTION AND MOTIVATION

The task participants had to solve in the ACM Recommender Systems Challenge 2018 was a music information retrieval task [5], more precisely the task of automatic music playlist continuation (APC).[1] This task consists of adding one or more tracks to a music playlist in a way that fits the target characteristics of the original playlist [2, 6]. APC is a useful feature for music streaming services not only because it can extend listening session length, but also because it can increase engagement of users on their platform by making it easier for users to create playlists that they can enjoy and share.

Participants had to devise algorithms that predict, for a given playlist, an ordered list of 500 recommended candidate tracks. Performance was evaluated against a challenge set (cf. Section 2) of user-created playlists, where different combinations of playlist titles and some numbers of tracks were withheld. Their algorithms could either use only the data in the provided training dataset or may additionally exploit publicly available external data sources. Submissions provided by algorithms of the former kind were considered for the main track, those of the latter kind for the creative track. Both tracks were assessed individually and independently, and the top 3 teams in each track were awarded monetary prizes. We, however, enforced that an algorithm used in the main track was not submitted again to the creative track.

To ensure reproducibility and to assess compliance to the challenge rules, we required participating teams to provide their algorithms as open source and the links to the used data sources (applicable for the creative track).

2 DATASET

For algorithm development and testing, we released a dataset of one million user-created playlists from the Spotify platform, dubbed the *Million Playlist Dataset* (MPD). Statistics of the MPD can be found in Table 1. The dataset includes, for each playlist, its title as well as the list of tracks (including album and artist names), and some additional metadata such as Spotify URIs and the playlist's number of followers. The playlist titles in the dataset were unmodified, however for reporting in Table 1, playlist titles were lightly normalized by converting to lowercase, and removing spaces and common non-alphanumeric symbols.

A separate *challenge dataset* was used to validate the quality of the elaborated algorithms. It consisted of a set of playlists from which a number of tracks had been withheld. The challenge set was composed of 10,000 incomplete playlists and covered a total of 10 scenarios (1,000 playlists for each): (1) title only, no tracks, (2) title and first track, (3) title and first 5 tracks, (4) no title and first 5 tracks, (5) title and first 10 tracks, (6) no title and first 10 tracks, (7)

[1] http://2018.recsyschallenge.com

RecSys '18, October 2–7, 2018, Vancouver, BC, Canada

ACM ISBN 978-1-4503-5901-6/18/10.
https://doi.org/10.1145/3240323.3240342

Property	Value
Number of playlists	1,000,000
Number of tracks	66,346,428
Number of unique tracks	2,262,292
Number of unique albums	734,684
Number of unique artists	295,860
Number of unique playlist titles	92,944
Number of unique normalized playlist titles	17,381
Average playlist length (tracks)	66.35

Table 1: Basic statistics of the Million Playlist Dataset.

title and first 25 tracks, (8) title and 25 random tracks, (9) title and first 100 tracks, and (10) title and 100 random tracks.

The task was then to predict the missing tracks in those playlists, and participating teams were required to submit their predictions for those missing tracks (as list of 500 ordered predictions). The withheld tracks were used by the organizers as ground truth, i.e. to compute the performance measures for each submission.

3 EVALUATION

To assess the quality of submissions, we computed three metrics and averaged them across all playlists in the challenge dataset: R-precision, normalized discounted cumulative gain (NDCG) [4], and recommended songs clicks. *R-precision* measures the fraction of recommended relevant items among all known relevant items (i.e., the number of withheld tracks) and is invariant of the order in which tracks are retrieved. The *R-precision* is calculated on both the track and the artist level, with artist matches contributing a partial score (of 0.25) even if the track is incorrect. Let G_T and G_A be the set of unique track IDs and artist IDs in the ground truth respectively. Let S_T be the subset of tracks IDs in the top-$|G_T|$ tracks recommended in the submitted playlist, and S_A be the set of unique artist IDs in the same set. Then:

$$\text{R-precision} = \frac{|S_T \cap G_T| + 0.25 \cdot |S_A \cap G_A|}{|G_T|}$$

In contrast, *NDCG* assesses the ranking quality of the recommended tracks and increases when relevant tracks are placed higher in the recommendation list [1]. *Recommended songs clicks* is a user-centric beyond-accuracy measure that relates to a Spotify feature called Recommended Songs. Given a set of tracks in a playlist, this feature recommends 10 tracks to add to the playlist. The list can be refreshed to produce 10 more tracks. The recommended songs clicks measure is the number of refreshes needed before the first relevant track is encountered. It is formalized as shown in Equation 1, where R is the list of tracks recommended by a participant's algorithm and G is the ground truth, i.e., the omitted tracks from the real playlist.

$$clicks = \left\lfloor \frac{argmin_i\{R_i : R_i \in G\} - 1}{10} \right\rfloor \quad (1)$$

If there is no relevant track in R, a value of 51 is picked, which is 1 plus the maximum number of clicks possible. To aggregate the individual scores for the three metrics, Borda rank aggregation [3] is used.

4 STATISTICS OF PARTICIPATION

The Challenge was well received: 1,791 people registered; 1,430 with an academic affiliation and 361 from industry. These people formed a total of 410 teams. Out of these, 117 teams were active, i.e., submitted at least one run (113 and 33, respectively, to the main and to the creative track). In total we received 1,467 submissions, out of which 1,228 were submitted to the main track and 239 to the creative track.

5 RESULTS

The highest performing team in the main track achieved an R-precision of 0.2241, an NDCG of 0.3946, and an average number of recommended songs clicks of 1.784. In the creative track, an R-precision of 0.2233, an NDCG of 0.3939, and a click rate of 1.785 was realized by the best team. Final results of all participating teams for the main track[2] and the creative track[3] are available online. A naïve baseline approach was implemented, which took the 500 most commonly appearing tracks in the training set, and recommended them for all playlists in the challenge set. The results of this approach achieved an R-precision of 0.0458, NDCG of 0.0993, and Clicks of 13.217.

6 CONCLUSION

We provided details about the ACM Recommender Systems Challenge 2018 on music playlist continuation. We presented the datasets used for algorithm development and for validation, detailed the evaluation metrics, and reported statistics about participation in the Challenge as well as on the obtained results. The outcomes of the Challenge provided interesting insights and will contribute to the next generation of music recommender systems.

7 ACKNOWLEDGMENTS

We would like to thank everyone at Spotify who was involved in the Challenge, including Ben Carterette, Christophe Charbuillet, Cedric de Boom, Jean Garcia-Gathright, James Kirk, James McInerney, Vidhya Murali, Hugh Rawlinson, Sravana Reddy, Marc Romejin, Romain Yon, and Yu Zhao. Furthermore, we greatly appreciate the help provided by previous organizers of the Challenge, in particular by Yashar Deldjoo, Mehdi Elahi, and Alan Said.

REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval – the concepts and technology behind search.* Addison-Wesley, Pearson, Harlow, England, 2nd edition, 2011.
[2] G. Bonnin and D. Jannach. Automated generation of music playlists: Survey and experiments. *ACM Computing Surveys (CSUR)*, 47(2):26, 2015.
[3] J.-C. de Borda. Mémoire sur les élections au scrutin. *Histoire de l'Académie Royale des Sciences*, 1781.
[4] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, Oct. 2002.
[5] M. Schedl, E. Gómez, and J. Urbano. Music information retrieval: Recent developments and applications. *Foundations and Trends in Information Retrieval*, 8(2–3):127–261, 2014.
[6] M. Schedl, H. Zamani, C.-W. Chen, Y. Deldjoo, and M. Elahi. Current challenges and visions in music recommender systems research. *International Journal of Multimedia Information Retrieval*, 7(2):95–116, Jun 2018.

[2] https://recsys-challenge.spotify.com/static/final_main_leaderboard.html
[3] https://recsys-challenge.spotify.com/static/final_creative_leaderboard.html