Diversity enhancement for collaborative filtering recommendation

Liu Yankai[1,*]

[1]China Mobile Research Institute, 32 Xuanwumen West Street, Beijing, 100000, China

Abstract

To evaluate the user experience of recommendation systems in realistic and complex scenarios, the EvalRS challenge evaluates recommendation algorithms from multiple perspectives such as fairness, diversity. This paper details the diversity-enhanced collaborative filtering recommendation algorithm that won first place in the EvalRS challenge. Our proposed solution has two essential innovations. First, the importance of the user's historical behavior is ranked so as to obtain a high-ranking performance using fewer user behaviors. Second, the recommendation results are re-ranked to enhance the diversity of the recommendation results. In addition, this paper proposes a new evaluation metric, the quantile-based fairness Gini coefficient, to metric the fairness of the recommendation results, as it does not cause drastic fluctuations due to the small number of item interactions.

Keywords

collaborative filtering, recommendation evaluation, recommendation fairness

1. Introduction

Conventional recommendation algorithm evaluation metrics are usually ranking performance metrics such as hit rate, ndcg, MRR, etc., which may lead to recommendation systems considering only certain user preferences and ignoring multiple aspects of user experience. Fairness and diversity are as important as ranking performance metrics. If users frequently interact with a single type of item, the user experience of the recommendation system will be damaged in the long run, which will eventually lead to user churn. Therefore, the recommendation algorithm model needs to be evaluated from multiple perspectives. This paper illustrates the solution of EvalRS challenge[1], including a collaborative filtering algorithm based on frequent item mining, importance ranking of user behavior based on TF-IDF, and diversity-enhanced re-ranking algorithm. The next section presents a brief introduction to this challenge, and section 3 details the solution and the fairness index based on the Gini coefficient; finally, the section 4 shows the experimental results. Furthermore, the specific solution code and documentation are publicly available on GitHub: https://github.com/lazy2panda/cikm2022_solution.

2. The Challenge

EvalRS is one of the challenges of CIKM 2022 AnalytiCup, which is based on the reclist[2] and aims to evaluate recommender systems in terms of key dimensions such as diversity and fairness. The dataset of this challenge is based on LFM-1b[3], a music dataset with about 820,000 songs, 110,000 users, and about 37 million user interactions.The goal of the challenge is to evaluate the recommendation system in multiple aspects, including Standard recommendation system metrics, Standard metrics on a per-group or slice basis, Behavioral and qualitative tests, etc.[1]

Table 1
Results of various experiments

Tests	exp1	exp2	exp3
TRACK_POPULARITY_GINI	0.0046	**0.0008**	**0.0008**
HIT_RATE	**0.0330**	0.0154	0.0154
MRR	**0.0106**	0.0066	0.0058
MRED_COUNTRY	-0.0068	**-0.0040**	**-0.0040**
MRED_USER_ACTIVITY	-0.0085	**-0.0069**	**-0.0069**
MRED_TRACK_POPULARITY	-0.0282	**-0.0020**	**-0.0020**
MRED_ARTIST_POPULARITY	-0.0134	**-0.0016**	**-0.0016**
MRED_GENDER	-0.0027	**-0.0009**	**-0.0009**
BEING_LESS_WRONG	0.3363	0.3240	**0.4248**
LATENT_DIVERSITY	-0.1995	-0.2268	**-0.1216**
AGGREGATE_SCORE	-3.7511	1.3990	**1.7025**

CIKM'22: Proceedings of the 31st ACM International Conference on Information and Knowledge Management

[*]Corresponding author.

liuyankai@chinamobile.com (L. Yankai)
0000-0001-7159-4975 (L. Yankai)

3. Solution

3.1. Model Architecture

The main model used in this paper is the n-gram model, which is mainly applied in the field of natural language processing and is a statistical-based algorithm.[4] Its basic idea is that the content inside the text is operated with a sliding window of size N, forming a sequence of segments of length N. This method first sorts the user history sequence by time denote as H_u and uses the n-gram algorithm to process each user history sequence, where we take N in n-gram as 2; that is, each user history sequence is cut into multiple subsequences of length 2, then the whole training set is 2-gram sequence sliced, and calculate the frequency of all trackid pairs; for trackid pair (i, j), its frequency is denoted as $F(i, j)$; for track i, the track with the highest co-occurrence frequency is denoted as $S(i)$. Next, based on the user's history H_u, the similar track $S(i)$ of each user's history is obtained, and then the recommendation result of the user is obtained by ranking all similar items according to their frequency. In order to reduce the popularity fairness of track in recommendation results, we use the TF-IDF value of track to sort the user history sequence and truncate the user history records, where TF-IDF is calculated as follows, for track t, where C_t^u is the number of plays of t in user u, and C_t^{all} is the number of plays of t in all users, where G denotes the total number of users and $G(t)$ denotes the number of users who have interacted with t.

$$TF - IDF(t) = \frac{C_t^u}{C_t^{all}} * (log\frac{G}{G(t)} + 1) \tag{1}$$

Finally, diversity-enhanced reranking is performed on the basis of the above recommendation results. The user's recommendation results are denoted as R_u, and the user's history is recorded as H_u. The EvalRS challenge for diversity is defined as $D(R_u) = 0.3 * diversity - 0.7 * bias$, where *diversity* is defined as the sum of the differences between each point in the prediction space and the mean of the prediction space, and *bias* is defined as the distance between the ground truth vector and the mean of the prediction vector.[1] In this method, we combine the MMR[5] diversity algorithm with the diversity definition of EvalRS and use the vector mean of H_u as the ground truth vector to calculate the recommended result R_u as follows:

$$M = \arg\max_{t_i \in R_u|S} [0.3 * \text{diversity}([S, t_i])) - 0.7 * \text{bias}([S, t_i]))] \tag{2}$$

3.2. Quantile-based Gini coefficient fairness test

Based on the Gini coefficient, this paper proposes a new test to calculate the fairness of track popularity recommendation. The Gini coefficient originates from the field of economics and is often used to assess the degree of fairness of income distribution of residents. This paper uses the Gini coefficient to assess the degree of variation in the accuracy of items across quartile intervals. Compared to the standard deviation, the Gini coefficient is a more accurate reflection of the difference in fairness between two pairs of items across quartile intervals. The specific calculation is, firstly, dividing the trackid into multiple quantile intervals according to the popularity, secondly, calculating the false positive rate (FPR) of the trackid in different quantile intervals, and finally,the Gini coefficient is calculated as follows:

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} |FPR_i - FPR_j|}{2 * \bar{FPR}} \tag{3}$$

4. Experiments

The experiments part shows the metrics obtained after several iterations. As shown in Table 1, the results of Experiment 1 indicate that HIT_RATE and MRR are the highest, fairness and diversity metrics are the lowest, and the final AGGREGATE_SCORE is the lowest when no samples are performed on user history sequences. Experiment 2 performs TF-IDF sorting on user history sequences and uses only the top 8 tracks for inference. The results show that the fairness metrics are significantly improved compared to Experiment 1, with a 94% improvement in the MRED_TRACK_POPULARITY metric. Based on Experiment 2, Experiment 3 conducted a diversity enhancement ranking. Compared with Experiment 2, Behavioral and qualitative tests were significantly improved, with BEING_LESS_WRONG improved by 31% and LATENT_DIVERSITY improved by 46%, and the others metrics remained the same as in Experiment 2. The AGGREGATE_SCORE of Experiment 3 is 1.7025. In addition, TRACK_POPULARITY_GINI is our custom test metric whose value can reflect the fairness of track popularity. Finally, since the competition requires that the submission must be run on the AWS EC2 p3.2xlarge instance within 90 minutes, with the support of JIUTIAN Artificial Intelligence Platform[1], we completed the experiments and performance tests.

5. Conclusion

This paper details the EvalRS challenge's solution, which is based on the n-gram collaborative filtering algorithm and uses TF-IDF to rank the importance of users' history records, followed by re-ranking for diversity, and finally achieves better results in aggregation metrics.

[1] https://jiutian.10086.cn

References

[1] J. Tagliabue, F. Bianchi, T. Schnabel, G. Attanasio, C. Greco, G. d. S. P. Moreira, P. J. Chia, Evalrs: a rounded evaluation of recommender systems, 2022. URL: https://arxiv.org/abs/2207.05772. doi:10.48550/ARXIV.2207.05772.

[2] P. J. Chia, J. Tagliabue, F. Bianchi, C. He, B. Ko, Beyond NDCG: Behavioral testing of recommender systems with RecList, in: Companion Proceedings of the Web Conference 2022, ACM, 2022. URL: https://doi.org/10.1145%2F3487553.3524215. doi:10.1145/3487553.3524215.

[3] M. Schedl, The lfm-1b dataset for music retrieval and recommendation, in: Proceedings of the 2016 ACM on international conference on multimedia retrieval, 2016, pp. 103–110.

[4] W. B. Cavnar, J. M. Trenkle, et al., N-gram-based text categorization, in: Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval, volume 161175, Citeseer, 1994.

[5] J. Carbonell, J. Goldstein, The use of mmr, diversity-based reranking for reordering documents and producing summaries, in: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 1998, pp. 335–336.