SCIENCE CHINA
Information Sciences

• HIGHLIGHT •

January 2018, Vol. 61 018101:1–018101:3
doi: 10.1007/s11432-016-9078-0

Accurate inference of user popularity preference in a large-scale online video streaming system

Xiaoying TAN[1], Yuchun Guo[1], Yishuai CHEN[1]* & Wei ZHU[2]

[1]*School of Electronics and Information Engineering, Beijing Jiaotong University, Beijing* 100044*, China;*
[2]*PPTV Inc., Shanghai* 200120*, China*

Received 21 March 2016/Revised 4 January 2017/Accepted 16 March 2017/Published online 12 July 2017

Citation Tan X Y, Guo Y C, Chen Y S, et al. Accurate inference of user popularity preference in a large-scale online video streaming system. Sci China Inf Sci, 2018, 61(1): 018101, doi: 10.1007/s11432-016-9078-0

With the fast growth of online video services, the service providers pursue to satisfy users' personal preferences. Most of them have noticed the diversity of users' preferences on video content but not that on video popularity. Only Goel et.al. [1] proved in other domains that users have different popularity preferences (PPs) and Oh et.al. [2] used the statistics of users' PPs to improve recommendation performances. However, the statistical method to obtain users' PPs is biased when the available historical records are so limited as that in an online video recommendation system. In this article, we characterize users' PPs in a large-scale online video streaming system from China and propose two collaborative filtering (CF) [3] based algorithms to infer users' PPs. Compared with the statistical method, our proposed algorithms largely enhance the PP accuracy, and the enhancement gets larger with the fewer training data. Our work is beneficial for providing better personalized services.

Dataset. We base our study on a a large-scale dataset from the client of PPTV, one of the largest typical online video streaming systems in China. In the dataset, we filter out the sessions shorter than 30 s where users might not be purposeful watching out of interest, and filter out the users with less than 20 records to ensure that we have enough data to evaluate the accuracy of our inference algorithm. The resulted dataset collected from March 23rd to 28th in 2011 including more than 20 thousands of movie videos, 90 thousands of users and more than 2 million of sessions.

Characterization. We assign each user a PP sequence whose elements are the ordered popularity rankings of each video one has watched yet. We characterize an individual user's PP sequence with the respective of three statistical terms: central tendency (measured by Median), dispersion tendency (by coefficient of variation (CV)) and skewness (by a normalized metric defined to be (Mean−Median)/Standard Deviation). These three characteristics above complement each other. Any single one, such as only the central tendency examined in literature [1], would be not enough to characterize the users' PPs.

To examine whether the users' PPs are homogenous, we compare the distributions of the three PP characteristics in the real dataset and those in a null model which assumes that the users select the videos at a probability proportional to the video's popularity homogeneously. We find the observations as below.

(i) Most real users in PPTV prefer the popular videos averagely but not as significantly as that assumed in the null model, as shown in Figure 1(a). Such a gap is different in different systems. For example, the majority of users in Netflix (a movie rental system), as shown in Figure 5(a) in the literature [1], averagely prefers more popu-

* Corresponding author (email: yschen@bjtu.edu.cn)
The authors declare that they have no conflict of interest.

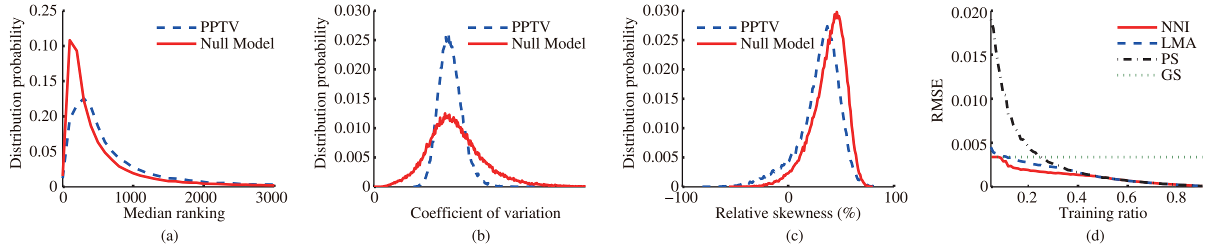

Figure 1 (Color online) Distributions of the three characteristics, (a) median, (b) CV and (c) relative skewness, of the PP sequences in PPTV and that in the null model. (d) is the PP inference accuracy of our proposed algorithms and the baseline algorithms.

lar videos than that assumed in the corresponding null model, whereas the typical medians in PPTV is a little larger than that assumed in the null model. It may because that watching online is more effortless than renting movies so that the users have more opportunities to watch the cold videos in PPTV. For another, the real medians in Web search system, as shown in Figure 5(c) in the literature [1], distribute much more dispersal than that in PPTV. One reason for the difference is that the navigational search behavior in the Web search system is more purposeful than the browsing behavior in PPTV. The differences between systems validate the necessity of a specific study on the online video streaming system.

(ii) Most users have wide preference ranges, and the ranges are different among the users. A distribution with a CV larger than 1 is considered to have a higher variance than that of an exponential distribution. Through measurement, we find that more than 60% of the CVs are larger than 1 and is incompatible with the null model. Furthermore, as shown in Figure 1(b), the ranges distribute more dispersal in PPTV than that assumed in the null model, as the interquartile range in PPTV ([0.89, 1.38]) is more than twice as wide as that in the null model ([0.97, 1.19]).

(iii) Most users are popular-video-biased within their own PP ranges as the same as that assumed in the null model, as shown in Figure 1(c), which may be caused by the current common practice in an online video system that recommend the popular videos on the front page.

On the above, users have diverse preferences on video popularity, although sharing some common characteristics. It suggests service providers satisfy the users' personal PP rather than only provide the most popular videos. Furthermore, we find that the distribution gaps among the real dataset and the corresponding null model in our system are somehow different from those in other systems (e.g., movie rental system and web browsing), which implies that our system needs a specific research.

Inference algorithm. In practical, it is necessary to infer a user's preference accurately at the early stage, instead of doing that at the end of the user's life cycle, in order to take advantage of such preference for providing the user with better personalized services. In this case, the challenge of inferring an individual user's PP accurately is that we could not observe all the watching records but only the quite limited records during the user's whole life cycle, so that the statistical PP observed from the limited records is likely to be biased.

To address this challenge, we propose to utilize the idea of the CF algorithms [3], a technique widely used in recommendation system (RS), referring to the collection of the preferences of the user's similar users to correct the bias. We propose two CF-based PP inference algorithms, namely user-based K-nearest neighbor inference (NNI) algorithm and low-rank matrix approximation (LMA) algorithm respectively. Formally, after packing the videos into r (set to be 15 here) ranking bins according to their logarithmic popularities, we assign each user a r-dimension PP vector whose elements are the statistical frequency of the user watching the videos at the corresponding ranking bins. Thus, for an individual user, our proposed CF-based inference algorithms are to find the user's accurate PP vector, provided with the training PP vectors of the user's own and the other users that are calculated on their observed limited records.

In the NNI algorithm, we first select the neighbor users whose similarities with the active user are larger than a threshold (thd). The similarity is measured by the cosine coefficient of their training PPs. Different from the traditional KNN algorithm used in RS that only utilizes the other neighbors' interest, we also take into account the active user's own observed PP, since the observed training PP of this user is valuable although somehow inaccurate. Then, we infer the active user's PP to be the weighted average of the training PPs of this user's own and the neighbors. The weights of the neighbors are proportional to their similarities with the active user.

In the LMA algorithm, we collect the training PP vectors of all the m users to be a training $m \times r$ PP matrix, and apply the singular value decomposition (SVD) technique [4] to infer the accurate PP matrix. Compared with another technique, nonnegative matrix factorization (NMF) [5], which works well in the face of a quite sparse training matrix, the SVD method is more appropriate for our inference since our training PP matrix is not sparse and the SVD method is more reliable than NMF method where the learning of the parameters is sometimes trapped in local optimum. Through factorizing the training PP matrix, the SVD technique extracts the most k principle characteristics of the users and the ranking bins respectively and filters out the disturbance of statistic noise brought by the limited observations.

Evaluation. We randomly select a certain ratio of our data to calculate users' training PPs and use the whole dataset to calculate the true accurate PPs. The training ratio varies in each separate round. The optimal parameters, i.e., thd and k, are chosen against the different training ratios.

We compare the performances of our algorithms with two baselines: (1) the personal statistic (PS) which supposes the inferred PP matrix equals the training one, and (2) the global statistic (GS) method which supposes the users prefer videos in proportion to the videos' global popularity homogenously.

As the root mean square error (RMSE) results shown in Figure 1(d), when the training ratio is set from 9% to 48%, our proposed algorithms outperform both of the baseline methods. For example, when the training ratio is 30%, our NNI algorithm reduces the RMSE by 39% on the PS method, 51% on the GS algorithm and by 39% on the PS algorithm. The improvement on the PS gets more significant when the training ratio is smaller. When the training ratio gets smaller than 9% or larger than 48%, they perform as well as the GS and the PS, respectively, as the tuned optimal parameters in these cases make our algorithms degenerate into the baseline methods.

The results agree with our intuitions that collaborative opinions are more essential when an individual user has the fewer records. When the amount of the training records is large enough, the statistical PP is accurate enough; otherwise, the statistical PP is biased. Our proposed algorithms make up the bias with the aid of a collection of similar users' PPs.

Furthermore, the NNI algorithm makes the improvement more significantly than the LMA algorithm. For example, when the training ratio is 30%, the NNI algorithm brings another 20% improvement on the LMA algorithm averagely. Such an observation is opposite to that in RS where the NMF algorithm usually performs better than the KNN algorithm. That is because the matrix factorization-like algorithm is good at dealing with sparse data, like the user-item matrix in RS usually, but the PP matrix in our PP inference problem is not sparse.

Application. Our work is beneficial to the personalization services, e.g., personal recommendation. As proved in literature [2], the statistical user PP could improve the recommendation accuracy. We believe such an improvement could be more significant with the aid of the more accurate user PP obtained by our inference algorithm.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61271199, 61301082, 61572071).

References

1 Goel S, Broder A, Gabrilovich E, et al. Anatomy of the long tail: ordinary people with extraordinary tastes. In: Proceedings of the 3rd ACM International Conference on Web Search and Data Mining, New York, 2010. 201–210

2 Oh J, Park S, Yu H, et al. Novel recommendation based on personal popularity tendency. In: Proceedings of the 11th International Conference on Data Mining (ICDM), Vancouver, 2011. 507–516

3 Su X, Khoshgoftaar T M. A survey of collaborative filtering techniques. Adv Artif Intell, 2009, 2009: 1–19

4 Zhang S, Wang W, Ford J, et al. Using singular value decomposition approximation for collaborative filtering. In: Proceeding of the IEEE International Conference on E-Commerce Technology, Munchen, 2005. 257–264

5 Zhang S, Wang W, Ford J, et al. Learning from incomplete ratings using non-negative matrix factorization. In: Proceedings of SIAM International Conference on Data Mining, Bethesda, 2006. 549–553