Measuring the Concentration Reinforcement Bias of Recommender Systems

Panagiotis Adamopoulos
padamopo@stern.nyu.edu

Alexander Tuzhilin
atuzhili@stern.nyu.edu

Peter Mountanos
peter.mountanos@nyu.edu

Department of Information, Operations, and Management Sciences
Leonard N. Stern School of Business, New York University

ABSTRACT

In this paper, we propose new metrics to accurately measure the concentration reinforcement of recommender systems and the enhancement of the "long tail". We also conduct a comparative analysis of various RS algorithms illustrating the usefulness of the proposed metrics.

Keywords

Dispersion; Diversity; Long Tail; Popularity Reinforcement

1. INTRODUCTION

Even though many researchers have focused on developing efficient algorithms for generating more accurate recommendations, there is increasing interest in metrics that go beyond this paradigm [1, 2] and evaluate various other properties and dimensions of recommender system (RS) algorithms, including the popularity bias and dispersion of recommendations. However, following the currently established evaluation protocols and simply evaluating the generated recommendation lists in terms of dispersion and inequality of recommendations does not provide any information about the concentration reinforcement and popularity bias of the recommendations (i.e., whether popular or long-tail items are more likely to be recommended) since these metrics do not consider the prior popularity of the candidate items. Focusing on improving the current evaluation protocols of RSes through alleviating this problem, we propose new metrics to accurately measure the concentration reinforcement and "long-tail enhancement" of recommender system algorithms.

2. RELATED WORK

Several measures have been employed in prior research in order to measure the concentration reinforcement and popularity bias of RSes as well as other similar concepts. These metrics include catalog coverage, aggregate diversity, and the Gini coefficient. In particular, catalog coverage measures the percentage of items for which the RS is able to make predictions [9] while aggregate diversity uses the total number of distinct items among the top-N recommendation lists across all users to measure the absolute long-tail diversity of recommendations [5]. The Gini coefficient [7] is used to measure the distributional dispersion of the number of times each item is recommended across all users; similar are the Hoover (Robin Hood) index and the Lorenz curve.

However, these metrics do not take into consideration the prior popularity of candidate items and, hence, do not provide sufficient evidence on whether the prior concentration of popularity is reinforced or alleviated by the RS. Moving towards this direction, [3, 4] employ a popularity reinforcement measure M to assess whether a RS follows or changes the prior popularity of items when recommendations are generated. To evaluate the concentration reinforcement bias of recommendations, [3, 4] measure the proportion of items that changed from "long-tail" in terms of prior sales (or number of positive ratings) to popular in terms of recommendation frequency as: $M = 1 - \sum_{i=1}^{K} \pi_i \rho_{ii}$, where the vector π denotes the initial distribution of each of the K popularity categories and ρ_{ii} the probability of staying in category i, given that i was the initial category. In [3, 4], the popularity categories, labeled as "head" and "tail", are based on the Pareto principle and hence the "head" category contains the top 20% of items (in terms of positive ratings or recommendation frequency, respectively) and the "tail" category the remaining 80%. However, this metric of concentration reinforcement (popularity) bias entails an arbitrary selection of popularity categories. Besides, all items included in the same popularity category are contributing equally to this metric, despite any differences in popularity.

3. CONCENTRATION REINFORCEMENT

To precisely measure the concentration reinforcement (popularity) bias of RSes and alleviate the problems of the aforementioned metrics, we propose a new metric as follows:

$$CI@N = \sum_{i \in I} \frac{1}{2} \frac{s(i)}{\sum_{j \in I} s(j)} \ln \left(\frac{\frac{s(i)+1}{\sum_{j \in I} s(j)+1}}{\frac{r^N(i)+1}{N*|U|+|I|}} \right) + \frac{1}{2} \frac{r^N(i)}{N*|U|} \ln \left(\frac{\frac{r^N(i)+1}{N*|U|+|I|}}{\frac{s(i)+1}{\sum_{j \in I} s(j)+1}} \right),$$

where $s(i)$ is the prior popularity of item i (i.e., the number of positive ratings for item i in the training set or correspondingly the number of prior sales of item i), $r^N(i)$ is the number of times item i is included in the generated top-N recommendation lists, and U and I are the sets of users and items, respectively.[1] In essence, following the notion of Jensen-Shannon divergence in probability theory and statistics, the proposed metric captures the distributional divergence between the popularity of each item in terms of prior sales (or number of positive ratings) and the number of times each item is recommended across all users. Based on this metric, a score of zero denotes no change (i.e. the number of times an item is recommended is proportional to its prior popularity) whereas a (more) positive score denotes that the generated recommendations deviate (more) from the prior popularity (i.e., sales or positive ratings) of items.

In order to measure whether the deviation of recommendations from the distribution of prior sales (or positive ratings) promotes long-tail rather than popular items, we also

[1] Another smoothed version of the proposed metric is: $CI@N = \sum_{i \in I} \frac{1}{2} s_{\%}(i) \ln \left(\frac{s_{\%}(i)}{\frac{1}{2} s_{\%}(i) + \frac{1}{2} r^N_{\%}(i)} \right) + \frac{1}{2} r^N_{\%}(i) \ln \left(\frac{r^N_{\%}(i)}{\frac{1}{2} s_{\%}(i) + \frac{1}{2} r^N_{\%}(i)} \right)$, where $s_{\%}(i) = \frac{s(i)}{\sum_{j \in I} s(j)}$ and $r^N_{\%}(i) = \frac{r^N(i)}{N*|U|}$.

RecSys 2015 Poster Proceedings, September 16–20, 2015, Austria, Vienna.

propose a measure of "long-tail enforcement" as follows:

$$LTI_\lambda@N = \frac{1}{|I|}\sum_{i\in I}\lambda\left(1-\frac{s(i)}{\sum_{j\in I}s(j)}\right)\ln\left(\frac{\frac{r^N(i)+1}{N*|U|+|I|}}{\frac{s(i)+1}{\sum_{j\in I}s(j)+1}}\right) + (1-\lambda)\frac{s(i)}{\sum_{j\in I}s(j)}\ln\left(\frac{\frac{s(i)+1}{\sum_{j\in I}s(j)+1}}{\frac{r^N(i)+1}{N*|U|+|I|}}\right),$$

where $\lambda \in (0,1)$ controls which items are considered long-tail (i.e., the percentile of popularity below which a RS should increase the frequency of recommendation of an item). In essence, the proposed metric rewards a RS for increasing the frequency of recommendations of long-tail items while penalizing for frequently recommending already popular items.

4. EXPERIMENTAL RESULTS

To empirically illustrate the usefulness of the proposed metrics, we conduct a large number of experiments comparing various algorithms across different performance measures. The data sets we used are the MovieLens 100k (ML-100k), 1M (ML-1m), and "latest-small" (ML-ls), and the FilmTrust (FT). The recommendations were produced using the algorithms of association rules (AR), item-based collaborative filtering (CF) nearest neighbors (ItemKNN), user-based CF nearest neighbors (UserKNN), CF ensemble for ranking (RankSGD) [10], list-wise learning to rank with matrix factorization (LRMF) [12], Bayesian personalized ranking (BPR) [11], and BPR for non-uniformly sampled items (WBPR) [6] implemented in [8].

Figure 1 illustrates the results of the comparative analysis of the different algorithms across various metrics. In particular, Fig. 1 shows the relative ranking in performance for each algorithm based on popular metrics of predictive accuracy and dispersion as well as the newly proposed metrics; green (red) squares indicate that the specific algorithm achieved the best (worst) relative performance among all the algorithms for the corresponding dataset and metric.[2]

Based on the results, we can see that the proposed metrics capture different performance dimensions of an algorithm compared to the relevant metrics of Gini coefficient and aggregate diversity. Comparing the performance based on the proposed concentration bias metric ($CI@N$) with the metric of Gini coefficient, we see that even though on aggregate an algorithm might distribute more equally than another algorithm the number of times each item is recommended, it might still achieve this by deviating less from the prior popularity (i.e., number of sales or positive ratings) of each item separately (e.g., green color for Gini coefficient and red color for concentration reinforcement). Nevertheless, the differences among the LTI_λ performance and the other metrics (e.g., aggregate diversity) indicate that even though some algorithms might recommend fewer (more) items than others or distribute how many times each item is recommended less (more) equally among the recommended items, they might achieve this by frequently recommending more (fewer) long-tail items rather than more (fewer) popular items (e.g., red color for Gini coefficient and green color for "long-tail enforcement"). Hence, the two proposed metrics should be used in combination in order to evaluate i) *how much the recommendations of a RS algorithm deviate from the prior popularity of items* and ii) *whether this deviation occurs by promoting long-tail rather than already popular items.*

[2]We have reversed the scale of the Gini coefficient for easier interpretation of the results (i.e., the green color corresponds to the most uniformly distributed recommendations).

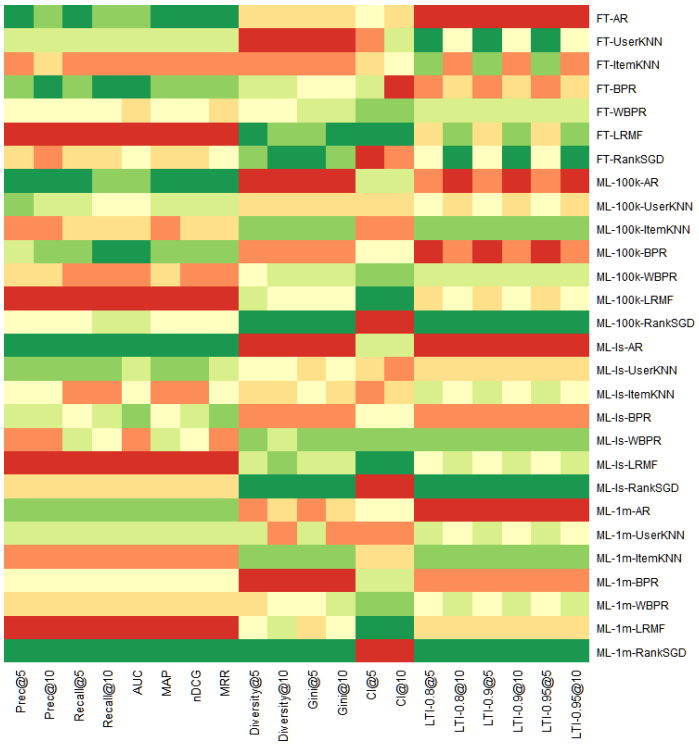

Figure 1: Performance (ranking) of various RS algorithms.

5. CONCLUSIONS

We propose new metrics to accurately measure the concentration reinforcement and "long-tail enforcement" of recommender systems. The proposed metrics capture different performance dimensions of an algorithm compared to existing metrics of RSes as they take into consideration the prior distribution of positive ratings and sales of the candidates items in order to accurately measure the effect of a RS. We also conduct a comparative analysis of various RS algorithms illustrating the usefulness of the proposed metrics.

6. REFERENCES

[1] P. Adamopoulos. Beyond rating prediction accuracy: On new perspectives in recommender systems. In *RecSys*. ACM, 2013.

[2] P. Adamopoulos. On discovering non-obvious recommendations: Using unexpectedness and neighborhood selection methods in collaborative filtering systems. In *RecSys*. ACM, 2014.

[3] P. Adamopoulos and A. Tuzhilin. Probabilistic neighborhood selection in collaborative filtering systems. *Working Paper: CBA-13-04, NYU*, 2013. http://hdl.handle.net/2451/31988.

[4] P. Adamopoulos and A. Tuzhilin. On over-specialization and concentration bias of recommendations: Probabilistic neighborhood selection in CF systems. In *RecSys*. ACM, 2014.

[5] G. Adomavicius and Y. Kwon. Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Trans. on Knowl. and Data Eng.*, 24(5):896–911, 2012.

[6] Z. Gantner, L. Drumond, et al. Personalized ranking for non-uniformly sampled items. In *Proceed. of KDD Cup*, 2012.

[7] C. Gini. Measurement of inequality of incomes. *The Economic Journal*, pages 124–126, 1921.

[8] G. Guo, J. Zhang, Z. Sun, and N. Yorke-Smith. Librec: A java library for recommender systems. In *UMAP'15*, 2015.

[9] J. Herlocker, J. Konstan, et al. Evaluating collaborative filtering recom. systems. *ACM Trans. Inf. Syst.*, 22(1), 2004.

[10] M. Jahrer and A. Töscher. Collaborative filtering ensemble for ranking. In *Proceedings of KDD Cup competition*, 2012.

[11] S. Rendle, C. Freudenthaler, et al. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *UAI '09*, 2009.

[12] Y. Shi, M. Larson, et al. List-wise learning to rank with matrix factorization for collaborative filtering. In *RecSys '10*, 2010.