Estimating the Magic Barrier of Recommender Systems: A User Study

Alan Said, Brijnesh J. Jain, Sascha Narr, Till Plumbaum, Sahin Albayrak, Christian Scheel
Technische Universität Berlin
Berlin, Germany
{alan, jain, narr, till, sahin, scheel}@dai-lab.de

ABSTRACT

Recommender systems are commonly evaluated by trying to predict known, withheld, ratings for a set of users. Measures such as the Root-Mean-Square Error are used to estimate the quality of the recommender algorithms. This process does however not acknowledge the inherent rating inconsistencies of users. In this paper we present the first results from a noise measurement user study for estimating *the magic barrier* of recommender systems conducted on a commercial movie recommendation community. The magic barrier is the expected squared error of the optimal recommendation algorithm, or, the lowest error we can expect from any recommendation algorithm. Our results show that the barrier can be estimated by collecting the opinions of users on already rated items.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval - Information filtering, Retrieval models

Keywords

recommender systems, evaluation, noise

1. INTRODUCTION & RELATED WORK

Recommender Systems (RS) are usually evaluated in offline settings [4], meaning on previously collected transactional data between users and items, such as movie ratings, product purchases etc. Evaluation of RSs is commonly performed in one of two scenarios; *top-N* or *rating prediction*. In the first scenario, the recommended list of probable items is compared to a set of withheld transactions. The quality of the recommendation is then expressed through an accuracy metric such as Precision or Recall based on the withheld transactions. The second scenario, rating prediction, attempts to estimate the ratings of a number of withheld items. Accuracy is measured in terms of prediction errors, such as the Root-Mean-Square Error (RMSE), or Normalized Discounted Cumulative Gain (NDCG). The presumption is that the lower the error measure is, the better the RS performs.

These two evaluation methods have been the de facto standard in the RS community, and common optimization approaches aim at either raising the precision values, or lowering the error measures, or a combination of both [4]. However, none of these methods address the concept of the *magic barrier* introduced by Herlocker et al. [4]. The magic barrier is a theoretical boundary for the level of optimization that can be performed on a recommender algorithm on known transactional data. The evaluation models assume that recorded transactions, whether ratings on movies, purchases of items, etc., reflect a *ground truth*; that the historical transactions users have performed are *free of noise*. These concepts have been addressed previously by Amatriain et al. [1, 2] in a user study performed in a synthetic environment, where a service was set up for the purpose of capturing this noise. To our knowledge, the first to mention this phenomenon in a RS setting, Hill et al. [5], discussed the reliability of users in terms of rating consistency already in 1995.

Since then, great advances have been made in the field of RS in terms of algorithmic development [3], much of it thanks to challenges such as the Netflix Prize[1], and yearly versions of the KDD Cup[2]. However, evaluation methodologies have not had a similar evolution, and most RS evaluation still uses traditional information retrieval measures and methods, even though these might not always reflect the actual quality of the recommendation [6].

In this paper, we present a user study, performed on a *real-life* movie RS. The study is intended to capture the natural level of noise in users' ratings. The initial results from the study confirm the notion that users' ratings are inherently noisy, and that the ratings should not be seen as an absolute truth. Based on this we propose a method for estimating a maximum optimization level for RSs based on error measures in a rating prediction scenario.

2. ESTIMATING THE MAGIC BARRIER

RMSE is commonly used for evaluating the accuracy of a rating function f on a set R of ratings

$$E(f|R) = \sqrt{\frac{1}{|R|} \sum_{(u,i) \in R} \big(f(u,i) - r_{ui}\big)^2}, \qquad (1)$$

where the sum runs over all user-item pairs (u, i) where $r_{ui} \in R$.[3]

Assuming the availability of additional transactions for (u, i) given at a different point in time, hereafter *opinions*, the error between an opinion o_{ui} and a rating r_{ui} of a user u for an item i is defined by $\varepsilon_{ui} = o_{ui} - r_{ui}$. We can suppose there is an unknown true rating function f_* that knows the true opinions o_{ui} of each user u about any item i. We derive an estimate of the the RMSE of f_* on the basis of R as

$$E(f_*|R) = \sqrt{\frac{1}{|R|} \sum_{(u,i) \in R} \big(o_{ui} - r_{ui}\big)^2}. \qquad (2)$$

which is equal to the standard deviation of E where $\varepsilon_{u,i} \in E$. It is possible that there are rating functions f with lower RMSE on R

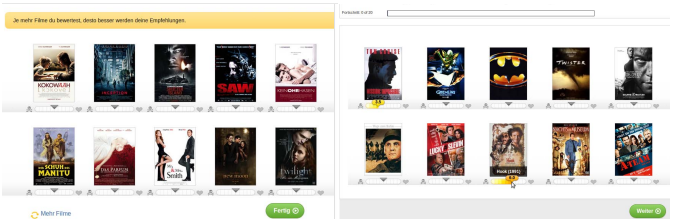
[1] http://www.netflixprize.com
[2] http://www.sigkdd.org/kddcup/
[3] For the sake of brevity, we abuse notation and write $(u, i) \in R$ for user-item pairs (u, i) for which $r_{ui} \in R$.

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

than f_*. Those functions, however, tend to overfit the given rating set R and are likely to degrade on the complement of R.

3. USER STUDY

In order to empirically estimate the magic barrier, a user study on the real-life commercial movie recommendation community moviepilot[4] was performed. moviepilot provides its users with personalized movie recommendations based on their previous ratings. The community counts its users in hundreds of thousands, ratings in dozens of millions and movies in tens of thousands. Information about the study was circulated to users through newsletters and on the community forums. The purpose was withheld so to not affect the outcome. The study was performed through a webpage mimicking the look-and-feel of the moviepilot website, on this page users were presented with a random selection of movies they had previously rated, with the ratings withheld. Users who chose to participate in the study were asked to give *opinions* on movies, the term opinion was used in order to mitigate the implicit meaning of *re-rating*, e.g. changing one's mind. A "next" button allowed for loading additional movies, not imposing users to give opinions on any movies shown at a specific page in order to not force users' opinions on movies they did not remember. After 20 opinions were collected the next button terminated the study.



(a) moviepilot (b) Opinion collecting interface

Figure 1: moviepilot's rating interface and the opinion interface used in the study. The opinion interface mimics the look-and-feel of moviepilot in order to create a feeling of familiarity lowering the level of UI-induced noise.

The study, which ran during April and May 2011, resulted in a dataset consisting of 6, 299 opinions by 306 users on 2, 329 movies.

4. RESULTS AND DISCUSSION

Figure 2 summarizes the opinion polling study. It shows the standard deviation of the error for (a) the deviation between opinion and ratings over all collected opinions, (b) for ratings above each user's average rating value, and (c) for ratings below the personal average rating value. The deviations are shown with respect to the internal rating scale of moviepilot (0 to 100 in steps of 5).

The estimated magic barrier of moviepilot is ± 12.01, as shown in Figure 2, which implies that improvements of predictions differing on average about 25 rating steps or less, on the $0, \ldots, 100$ scale, from the actual ratings are likely to be meaningless. The RMSE of 12.01 is an early stage approximation of the barrier. The study collected only one opinion per user-movie pair, which could introduce potential noise in the opinion values. In a more thorough study collecting several opinions of a user about the same item, we expect to reach a lower estimate of the magic barrier due to lower noise when averaging several user-opinion pairs.

When calculating the RMSE, we performed a temporal analysis, comparing the RMSE between rating and opinions based on the time which had passed between the rating event and the opinion event. The results showed that only if the rating and opinion had been performed within a very short timespan of each other (a few days) did the RMSE become (little) lower. In other cases, the RMSE stayed at a constant level. Due to the temporal aspect being of little difference in terms of RMSE, the results of the study argue for the fact that difference between ratings and opinions were not introduced by time (e.g. users' tastes changing).

[4] http://www.moviepilot.de

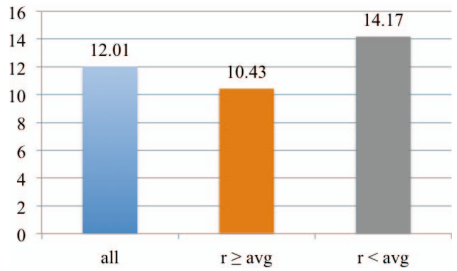


Figure 2: Standard deviation of the error, where *all* refers to the deviation over all opinions; $r \geq avg$ and $r < avg$ refer to the deviation over all ratings above and below average.

5. CONCLUSION & FUTURE WORK

In this paper we have presented a study on the inherent noise found in the rating values given by users in a commercial movie recommendation system. Our assumption, that the so-called *magic barrier* of recommender systems can be better assessed by noise estimation in scenarios similar to ours, seems to hold. Based on the collected data, we have presented an early model for the magic barrier and the level of accuracy a recommender systems can achieve without over-fitting to the noise in the data.

We are currently in the process of processing the data from the study, and preparing a larger scale study in a similar environment in order to gain enough empirical data for a final model for magic barrier estimation.

6. ACKNOWLEDGMENTS

The authors would like to thank the moviepilot team for their support. The presented work was conducted in the scope of the KMulE project, which was sponsored by the German Federal Ministry of Economics and Technology (BMWi).

7. REFERENCES

[1] X. Amatriain, J. M. Pujol, N. Tintarev, and N. Oliver, 'Rate it again: increasing recommendation accuracy by user re-rating', in *3rd ACM conf. on Recommender systems*. ACM, (2009).

[2] Xavier Amatriain, Josep M. Pujol, and Nuria Oliver, 'I like it... i like it not: Evaluating user ratings noise in recommender systems', in *Proc. of 17th Intl. Conf. on User Modeling, Adaptation, and Personalization*. Springer-Verlag, (2009).

[3] Robert M. Bell and Yehuda Koren, 'Lessons from the netflix prize challenge', *SIGKDD Explor. Newsl.*, **9**, (12/2007).

[4] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl, 'Evaluating collaborative filtering recommender systems', *ACM Trans. Inf. Syst.*, **22**, (2004).

[5] Will Hill, Larry Stead, Mark Rosenstein, and George Furnas, 'Recommending and evaluating choices in a virtual community of use', in *Proc. SIGCHI conf. on Human factors in computing systems*, New York, NY, USA, (1995). ACM.

[6] G. Shani and A. Gunawardana, 'Evaluating recommendation systems', in *Recommender Systems Handbook*, (2011).