Comparative Evaluation of Recommender System Quality

Paolo Cremonesi

Franca Garzotto

Sara Negro

Alessandro Papadopoulos

Politecnico di Milano (DEI)
Via Ponzio 34/5
20133 Milano, Italy

[first.last]@polimi.it

Roberto Turrin

Moviri srl (Italy), R&D

roberto.turrin@moviri.com

Abstract

Several researchers suggest that the Recommendation Systems (RSs) that are the "best" according to statistical metrics might not be the most satisfactory for the user. We explored this issue through an empirical study that involved 210 users and considered 7 RSs using different recommender algorithms on the same dataset. We measured *user's perceived* quality of each RS, and compared these results against measures of *statistical quality* of the considered algorithms as they have been assessed by past studies in the field, highlighting some interesting results.

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.
ACM 978-1-4503-0268-5/11/05.

Keywords

Recommender systems, quality metrics, user study.

ACM Classification Keywords

H3.3. Information search and retrieval; H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms

Measurement, Experimentation, Human Factors

Introduction

Recommender Systems (RSs) play an increasingly important role in online applications characterized by a very large amount of data - e.g., multimedia catalogs of music, products, news, images, or movies. Their goal is to filter information and to recommend to users only the items that are likely of interest to them. Traditionally, the quality of a RS is defined in terms of *statistical metrics* - e.g., error metrics and accuracy metrics - which do not involve users and are evaluated algorithmically, using well-known techniques developed in the fields of information retrieval and machine learning. More recently, *user-centric* approaches to RS quality evaluation have received some interest in the research and industry arena of RS and HCI communities [7,10]. Some works pinpoint that the

quality of the User eXperience (UX) with a RS as determined by its pragmatic factors (e.g., usability) or hedonic characteristics (e.g., aesthetics and "fun") are as (or even more) important than algorithmically assessed quality to determine the user's attitudes towards a RS, and are more influential on users' decisions to implement a system's recommendations (e.g., to "purchase" recommended items). The paper provides a contribution to this discussion presenting an empirical study that involved 210 users and considered 7 RSs, which share the *same* dataset and user interface, but implement 7 different baselines and state-of-the-art recommender algorithms. We measured the *user's perceived quality* of each RS, and compared our results against the *statistical* quality of the considered algorithms, as it has been assessed by past studies in the field based *accuracy metrics*.

User-based study of RS quality

Research Variables

The study was designed as a between subjects controlled experiment, in which we measured *perceived quality,* decomposed into a number of attributes (*dependent variables*) in seven different experimental conditions, each one using a system that supports the *same user interface,* employs the *same dataset* in the movie domain, but implements a *different* recommender *algorithm* (*independent variable).* We refer to the ResQue model [10] as conceptual framework for RS user-centric quality evaluation but, to better scoping our research, we focus our analysis on three attributes:

1. *Perceived accuracy* (also called *Relevance*) - how much the recommendation matches the users' interests, preferences and tastes.
2. *Novelty* - the extent to which users receive "new" recommended items. We define a recommended movie to be *novel* for a user only if the user has no knowledge of it. According to this definition, this attribute can be regarded as a sub-dimension of ResQue concept of novelty (which is less stringent and also considers some aspects related to serendipity).
3. *Overall users' satisfaction* - the global users' feeling of the experience with the RS.

Our study considered several state-of-the-art recommender algorithms: (i) five *collaborative filtering* algorithms - Correlation Neighborhood (*CorNgbr*), Non-Normalized Cosine Neighborhood (*NNCosNgbr*), Asymmetric SVD (*AsySVD*), and *PureSVD* in two versions – (ii) a *non-personalized* one referred to as Top Popular (*TopPop*), and (iii) a *content-based* one – Latent Semantic Analysis (*LSA*) - very shortly described below (further details can be found in [2] and in the papers quoted therein). *TopPop* (Top Popular) implements a simple, non-personalized estimation rule, which recommends the most popular items to any user, regardless his or her profile. *CorNgbr* and *NNCosNgbr* are two item-based *k*-nearest-neighborhood (kNN) algorithms, whose rating prediction is based on the collaborative similarity among items. *AsySVD* and *PureSVD* are two algorithms based on latent-factor, i.e., users and items are represented into a low-dimensional space. This family of algorithms has been leading the Netflix contest thanks to its performance in terms on Round Mean Squared Error (RMSE). Finally, *LSA* is a content-based algorithm whose recommendation rule is based on domain specific item characteristics, such as director, actors, or genre.

Figure 1. ContentWise interface: catalog exploring and movie (up); movie details (down)

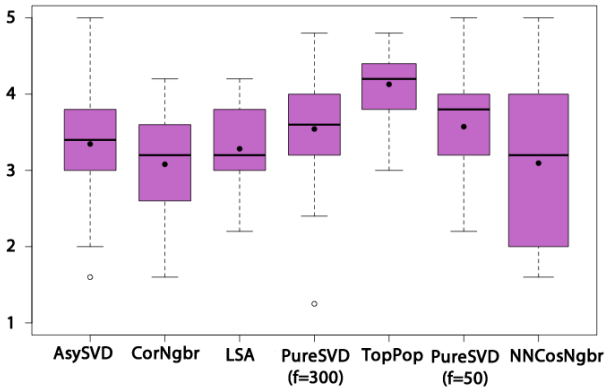

Figure 2. Perceived accuracy for each RS.

Instruments

We used a web-based commercial recommender *framework* - called ContentWise (Fig. 1). Its modularization and customization features allow us to easily create different experimental conditions and to evaluate user's perceived quality of the RSs using different algorithms, while sharing the same interface and dataset. Furthermore, it allows us to select a specific recommender algorithm among the seven ones we considered, and supports users with a wide range of typical RS functionalities, such as browsing a catalog of products, retrieving the detailed description of each item, getting recommendations and rating their relevance. The *dataset* is formed by 2137 movies and about 7.7 million ratings given by 49,969 users. We used a subset of the well-known large-scale *movie* dataset Netflix, integrated with data and metadata collected online (*e.g.*, movie plot, images, actors, director and genre).

Participants

Data collection was carried on by a team of 14 master students (two per experimental conditions) at our School of Information Engineering. Students were trained to perform the study, were given written instructions on the evaluation procedure, and were regularly supervised by a teaching assistant during their activities. Students were motivated in performing the evaluation to the best of their capabilities, as the work accounted for 50% of their mark at the courses. After a pre-screening among school mates, friends and relatives, evaluators recruited a group of *thirty* subjects for each algorithm, almost uniformly distributed w.r.t. gender and age. Overall, the study involved *210* users aged between 20 and 50, 54% male and 46% female. None of them had been previously used a RS.

Procedure

The evaluation took place in informal environments such as university (15%), interviewer's place (32%), and interviewee's place (31%). Each session lasted from 15 to 35 minutes. Each participant was initially invited to browse the movie catalog of ContentWise (pre-customized on a specific algorithm) and to freely select five known (not necessarily watched) movies, rating the degree of appreciation or interest for them on a 1-5 point scale. The user was then invited to explore the *five recommendations* returned by the system and to reply to a set of questions related to the quality of the recommendations. In order to compute novelty for a single suggested item, for each one we asked the question "Have you ever watched this movie or heard about it?" If the answer was "yes", novelty was set to 0. If the answer was "no" or "perhaps", the user was invited to explore information related to the movie to refresh memory. Therefore, if the final answer was still "no", novelty was set to 1, otherwise to 0. In the case of *perceived accuracy* for each suggested item, if the user has already watched the recommended movie, (s)he was asked to rate how much (s)he liked/disliked (on a 1-5 scale). Otherwise, (s)he was invited to look at the trailer and other available information on the movie (plot, director, cast, etc.) to form a more conscious opinion, in order to provide a

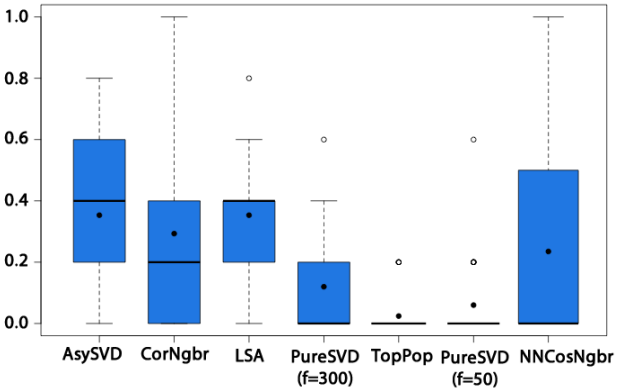

Figure 3. Novelty for each RS

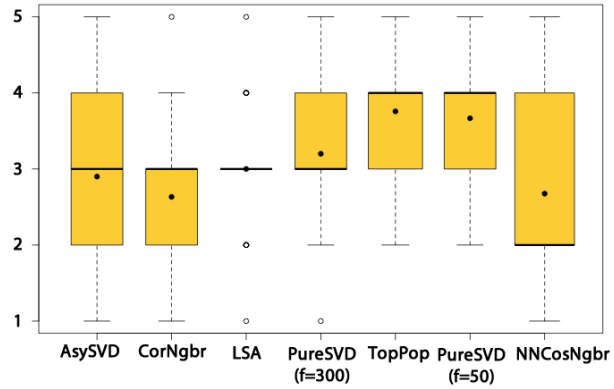

Figure 4. Global satisfaction for each RS

rating expressing his or her potential interest (on a 1-5 scale). For each user, novelty and accuracy were then calculated as the average on the respective values assigned to each recommended item. Finally, *overall satisfaction* was computed through questions devoted to assess user's global feeling about the set of recommendations, ranking how much (s)he likes/disliked the set of recommendations on a 1-5 scale.

User Study Results

Fig. 2 shows the box plot of the perceived relevance for each algorithm. Upper and lower ends of boxes represent *75th and 25th percentiles*. Whiskers extend to the most *extreme data point* which is no more than 1.5 times the interquartile range. *Median* is depicted with a solid line, *mean* with a dot. Outliers are represented with empty circles. A similar analysis has been performed for perceived novelty (Fig.3) and global satisfaction (Fig. 4).

All the algorithms have an average perceived accuracy between 3 and 4, which is quite good, and *TopPop* – the simplest one - have the highest perceived accuracy both in term of mean and median. This result may provide evidence against the real usefulness of sophisticated recommender algorithms, a hypothesis that will be further analyzed in the following of the paper.

In order to better compare the results, we first used 1-way ANOVA. The test suggests that, for each of the dependent variables, at least one of the algorithms differs significantly with respect to the others. We run multiple pair-wise comparison post-hoc tests using Tukey's method. All tests were run using a significance level $\alpha = 0.05$. Although along no quality dimension we could establish any "winning" RS, we can at least identify a partial order, outlined in Table 1.

	Accuracy	**Novelty**	**Global satisfaction**
Maximal	TopPop	AsySVD LSA CorNgbr	PureSVD50 TopPop
	$p < 0.01$	$p < 0.05$	$p < 0.05$
Intermediate	AsySvd PureSVD300 PureSVD50	NNCosNgbr PureSVD300	PureSVD300 LSA
	$p < 0.05$	$p < 0.05$	$p < 0.05$
Minimal	NNCosNgbr LSA CorNgbr	PureSVD50 TopPop	AsySVD CorNgbr NNCosNgbr

Table 1: Partial Ordering of RSs w.r.t. the various quality attributes; *p*-values between groups is shown

Evaluation of statistical quality

The RS performances are usually measured by methodologies based on accuracy metrics (e.g., recall and fallout) and error metrics (e.g., RMSE and MAE). Some of the algorithms tested in this study (*TopPop*, *NNCosNgbr* and *PureSVD*) cannot be evaluated with error metrics [4]. Hence, we have considered only accuracy metrics. In particular we have focused our attention on recall *r* (the conditional probability of suggesting a movie given it is relevant for the user)

and fallout *f* (the conditional probability of suggesting a movie given it is irrelevant for the user).

A good algorithm should have high recall (i.e., it should be able to recommend items of interest to the user) and low fallout (i.e., it should avoid to recommend items of no interest to the user). A measure that combines recall and fallout is the F-measure, defined as the harmonic mean of precision and recall. Precision can be estimated from recall and fallout by using the definition provided in. The testing methodology adopted in this study is similar to the one described in [7]. Table 2 presents the statistical accuracy of the tested algorithm. Algorithms in the table are ordered in decreasing order of recall.

	Recall	**Fallout**	**F-measure**
PureSVD50	0.29	0.005	0.45
PureSVD300	0.25	0.005	0.40
AsySVD	0.13	0.001	0.23
NNCosNgbr	0.12	0.010	0.21
TopPop	0.11	0.025	0.20
CorNgbr	0.08	0.010	0.15
LSA	0.01	0.002	0.02

Table 2. Recall, fallout and F-measure on the Netflix dataset computed for Top-5 recommendation lists

Recall and F-measure suggest *PureSVD* as being the most accurate algorithm. Second in line are *AsySVD*, *NNCosNgbr* and the non-personalized *TopPop* algorithms, all of them with a similar recall. The content-based *LSA* algorithm has the worst statistical accuracy both in terms of recall and F-measure. If we look at the fallout, *AsySVD* and *LSA* obtain the best results, while *NNCosNgbr* and *TopoPop* are the algorithms with the largest error rate.

Discussion and Conclusions

The analysis of the results presented in the previous section suggests a number of interesting considerations.

1. No algorithm is significantly better (or worse) than all the other in terms of *perceived relevance*. However, the partial ordering among the algorithms (Table 1) highlights that *TopPop* is the algorithm with the best perceived relevance (this is unexpected) and with the best novelty (as expected), thus its utility is limited because oftentimes the user has already watched the suggested items. Still, *TopPop* (together with *PureSVD300*) is at the top level in terms of global user satisfaction. In summary: *simple non-personalized TopPop recommendations are better perceived by the users with respect to other more sophisticated and personalized recommender algorithms,* although users are aware of the low utility of such recommendations*.* Global user satisfaction seems mainly driven by the perceived accuracy than by the novelty of the recommendations. This is a somehow surprising result, especially if we consider the large academic and industrial effort in the development of new and more sophisticated recommender algorithms.

2. The perceived *novelty* of content-based recommendations is equal or even better with respect to collaborative recommendations; Table 1 highlights that *AsySVD*, *CorNgbr* and *LSA* are the algorithms with the best perceived novelty, while *TopPop* and *PureSVD50* are the algorithms with the worst perceived novelty. *This result is in contrast with most of the existing literature in RSs which considers content-based algorithms as not able to recommend novel items*. For a better interpretation of this result, we should consider

that collaborative algorithms, by design, tend to recommend popular items, thus reducing the chances of novel recommendations.

3. *Statistical accuracy metrics (e.g., recall and fallout) are not always good predictors of users' perceived quality*. It is useful to consider that statistical metrics compute accuracy of recommendations by (i) exploiting previously-rated movies, i.e., user's rankings about known movies, and (ii) sampling all the ratings in the dataset - the majority of which concerns few popular movies. Consequently, statistical metrics focus their attentions on measuring the quality of an algorithm when recommending popular items and might not be particularly effective for measuring the quality of the same algorithm when recommending novel, unrated items.

Our research has its weaknesses, most notably (i) the limited number of user-centric quality attributes considered; (ii) the relatively limited sample size (30 participants) used for each RS. The fact that we have replicated the study in seven experimental conditions using the same methodological framework partially compensates for this drawback and strengthens the reliability of our results.

In spite of the above limitations, our work provides contributions both from a research and a practical perspective. To our knowledge, this is the first work that systematically compares perceived quality in a significant number – 7 – of different RSs isolating a precise factor – the underlying recommender algorithm - and analyses the results against statistical measures of quality. For the practice of RS design and evaluation, our work may promote further approaches that move beyond the attention to conventional accuracy metrics and shift the emphasis to more user-centric factors.

References

[1] L. Chen and P. Pu. A cross-cultural user evaluation of product recommender interfaces. *Proc.* RecSys '08, 75-82, New York, NY, USA, 2008. ACM.

[2] P. Cremonesi, Y. Koren, and R. Turrin. Performance of recommender algorithms on top-N recommendation tasks. Proc. *RecSys '10,* 39–46, Barcelona, Spain, 2010. ACM.

[3] J.L. Herlocker, J.A. Konstan, A. Borchers, and J. Riedl. An algorithmic framework for performing collaborative filtering. In *Proc. SIGIR Conf. R&D in Information Retrieval*, 230–237, New York, NY, USA, 1999. ACM.

[4] J.L. Herlocker, J.A. Konstan, L.G. Terveen, and J.T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Sys*, 22(1):5–53, Jan 2004.

[5] R. Hu and P. Pu. Potential acceptance issues of personality-based recommender systems. *Proc.* RecSys '09, 221-224, New York, NY, USA, Oct 2009. ACM.

[6] N. Jones and P. Pu. User Technology Adoption Issues in Recommender Systems. *Proc. Networking and Electronic Commerce Research Conf. 2007*, 379-394, Riva del Garda, Italy, 2007.

[7] S.M. McNee, J. Riedl, and J.A. Konstan. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI'06 Extended Abstracts,* 1097–1101, New York, NY, USA, 2006. ACM.

[8] P. Pu, L. Chen, and P. Kumar. Evaluating product search and recommender systems for e-commerce environments. 8:1-27, Jun 2008.

[9] P. Pu, M. Zhou, and S. Castagnos. Critiquing recommenders for public taste products. *Proc. RecSys '09*, 249-252, New York, NY, USA, 2009. ACM.

[10] P. Pu and L. Chen. A User-Centric Evaluation Framework of Recommender Systems. In *Proc. RecSys 2010 workshop on User-Centric Evaluation of RSs and Their Interfaces (UCERSTI)*, Barcelona, Spain, Sept. 2010.

[11] K. Swearingen and R. Sinha. Beyond algorithms: An HCI perspective on recommender systems. In *Proc. ACM SIGIR 2001 Workshop on Recommender Systems*. ACM, 2001.