User Profiling vs. Accuracy in Recommender System User Experience

Paolo Cremonesi
DEI-Politecnico di Milano
Via Ponzio 34/5
20133 Milano, Italy
(+39) 02 23993517
paolo.cremonesi@polimi.it

Francesco Epifania
Universita' degli Studi di Milano
Via Comelico 29
20135 Milano, Italy
(+39) 02 50316357
francesco.epifania@dis.unimi.it

Franca Garzotto
DEI-Politecnico di Milano
Via Ponzio 34/5
20133 Milano, Italy
(+39) 02 23993505
franca.garzotto@polimi.it

ABSTRACT

A Recommender System (RS) filters a large amount of information to identify the items that are likely to be more interesting and attractive to a user. Recommendations are inferred on the basis of different user profile characteristics, in most cases including explicit ratings on a sample of suggested elements. RS research highlights that profile length, i.e., the number of collected ratings, is positively correlated to the accuracy of recommendations, which is considered an important quality factor for RSs. Still, gathering ratings adds a burden on the user, which may negatively affect the UX. A design tension seems to exist, induced by two conflicting requirements – to raise accuracy by increasing the profile length, and to make the profiling process smooth for the user by limiting the number of ratings. The paper presents a wide empirical study (1080 users involved) which explores this issue. Our work attempts to identify which of the two contrasting forces influenced by profile length – recommendations accuracy and burden of the rating process - has stronger effects on the perceived quality of the UX with a RS.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: Multimedia Systems, User Interfaces; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Design, Experimentation, Human Factors

Keywords

Recommender System, User Experience Quality, UX Design, Accuracy, Profile Length, Empirical Study

1. INTRODUCTION

Recommender Systems (RSs) attempt to improve the quality of the UX with large amount of information by proposing items that are likely to be more interesting and attractive to the user. Such recommendations are inferred on the basis of different characteristics of the user profile that in most systems include explicit ratings, i.e., users' judgments on their degree of interest for a sample of items. Such recommendations are inferred on the basis of different characteristics of the user profile that in most systems include explicit ratings, i.e., users' judgments on their degree of interest for a sample of items.

One frequently adopted measure of the quality of a RS is accuracy, i.e., the degree of "match" between the characteristics of the recommended items and the user's tastes [8] [7][9], and most RSs attempt to maximize this factor. Research on RSs highlights that accuracy is positively correlated to the so called "profile length", i.e., the number of explicitly collected ratings: the more the items rated by the user, the higher the accuracy [17][18]. Still, the rating process adds a burden on the user. Several studies show that the risk in requiring users to rate many items is to annoy them [10], or to have them give up the rating process [12][14]. Hence, RS designers face a design tension induced by two conflicting requirements: to raise accuracy by increasing the profile length, and to make the profiling process smooth for the user by limiting the number of ratings. This tension raises an interesting research question: "which of the two contrasting "forces" that depend on the profile length – accuracy and burden of the rating process – have stronger effects on the perceived quality of the UX?" Answering this question can help designer prioritize design decision towards one or the other constraint.

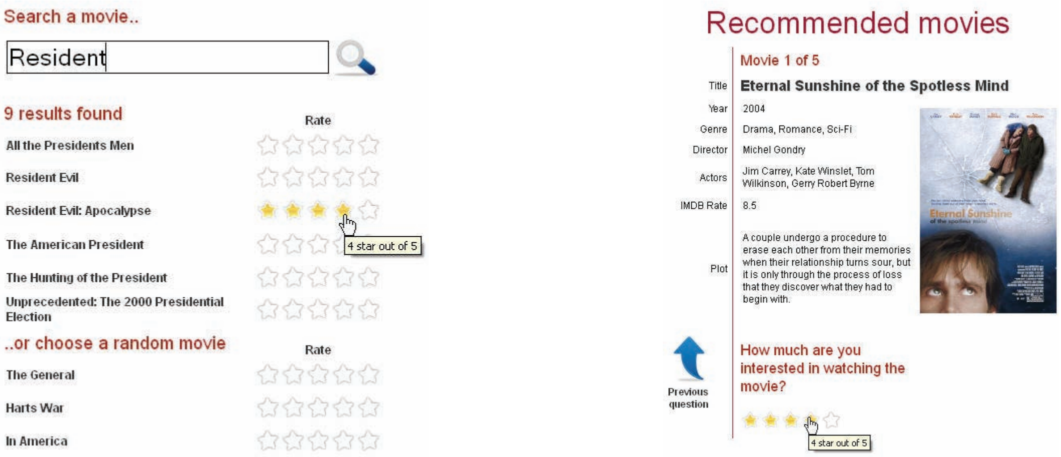
This paper presents a large empirical research (1080 participants) where we have explored the above issue. Our results show that the perceived quality of the UX with the RS does not significantly change with the variation of the profile length. In other words, the two contrasting forces mutually compensate: the potentially negative effect of a long profile is mitigated by an increase of accuracy, and the potentially positive effects of increased accuracy resulting from a longer profile is eroded by the burden of a more demanding rating process. The rest of the paper discusses the design of our study, and its main results, pinpointing their implications for research and design practice in the RS domain.

2. DESIGN OF THE STUDY

2.1 Study Variables and Research Hypothesis

The research was designed as *three replicated* between-subjects studies. In each study, we measured user's *perceived quality* of a RS in the movie domain in *two* different experimental conditions. In each experimental condition we used a recommender system having the same dataset, algorithm, and user interface, but a *different rating process*, asking users to rate a *different* number of movies. In other words, the two experimental conditions were

AVI '12, May 21-25, 2012, Capri Island, Italy



Search a movie..

Resident

9 results found

	Rate
All the Presidents Men	☆☆☆☆☆
Resident Evil	☆☆☆☆☆
Resident Evil: Apocalypse	★★★★☆
The American President	☆☆☆☆☆
The Hunting of the President	☆☆☆☆☆
Unprecedented: The 2000 Presidential Election	☆☆☆☆☆

4 star out of 5

..or choose a random movie

	Rate
The General	☆☆☆☆☆
Harts War	☆☆☆☆☆
In America	☆☆☆☆☆

(a) Providing explicit ratings during navigation

Recommended movies

Movie 1 of 5

Title	Eternal Sunshine of the Spotless Mind
Year	2004
Genre	Drama, Romance, Sci-Fi
Director	Michel Gondry
Actors	Jim Carrey, Kate Winslet, Tom Wilkinson, Gerry Robert Byrne
IMDB Rate	8.5
Plot	A couple undergo a procedure to erase each other from their memories when their relationship turns sour, but it is only through the process of loss that they discover what they had to begin with.

How much are you interested in watching the movie?

★★★★☆

4 star out of 5

(b) Providing relevance scores for a recommended item

Figure 1

characterized by different profile lengths (*independent variable*), respectively 5 and 20. The choice of these values was been motivated by considering that most research works on RSs assume to have users with at least 20 ratings (all the works based on the popular MovieLens [15] and Netflix datasets [3]) or 5 ratings (all the works based on the Book-Crossing dataset [19]).

User's perceived quality was operationalized in terms of two measurable metrics (*dependent variables*): *perceived relevance* and *global user satisfaction*, which are acknowledged as important quality factors in user centric approaches to RS quality (e.g., the ResQue model [1]). Perceived relevance measures how well the user believes that recommendations match his or her interests, preferences, and taste. Global user satisfaction measures how users feel about the experience with the RS.

Considering the strong emphasis that the RS literature put on improving accuracy, and the role on profile length for this purpose [17][18], we may assume that the effects on UX quality of a longer profile (and the consequent higher accuracy) "wins" over the negative effect (rating burden). Hence the research question defined in sect. 1 (Introduction) can be reformulated in terms of the following research hypotheses, which our empirical studies aimed at validating or confuting:

H1a: perceived relevance measured in the long profile condition is higher than in the short profile conditions

H1b: global satisfaction measured in the long profile condition is higher than in the short profile conditions

2.2 Instruments

We used the web-based recommender and evaluation framework PoliRec [1], shown in Figure 1, and powered by the ContentWise [2] recommendation engine. PoliRec supports users with a wide range of functionalities that are common in on-line DVD rental services such as Netflix, Lovefilm and Blockbuster. Users can browse a catalog of 2137 movies, retrieving the detailed description of each item, rating it, and getting recommendations. In each experimental condition, the modularization and customization features of PoliRec allowed us to i) select and apply a specific recommender algorithm among the three that we considered, and ii) set the desired profile length, i.e., the minimum number of ratings a user has to provide before receiving recommendations. PoliRec also embeds an on-line questionnaire system that allows researchers to collect quantitative and qualitative from the user in a relatively easy way.

2.3 Participants

The overall empirical research involved 1080 subjects. In *each* of the three replicated study, we recruited 360 subjects who were split in two groups of the same size, and randomly assigned to either experimental condition 1 (short profile) or experimental condition 2 (long profile). The same demographic characteristics were maintained in each subgroup: subjects aged between 20 and 50, evenly distributed into three age categories: 20-30, 30-40, 40-50. Overall, 52% of the subjects were male and 48% female. None of them had been previously exposed to the system used in our study, and none of them had technical knowledge about RSs.

2.4 Procedure

Our first study used a PoliRec version implementing a *non-personalized* algorithm. Algorithms of this kind present the same predefined list of items to everybody, regardless his or her user profile. We used a simple, non-personalized algorithm (TopPop), which recommends top-N items with the highest popularity (largest number of ratings) [3]. This initial study was used as baseline for verifying the internal validity of our research. As the accuracy of recommendations is not affected by profile length, the only measured force playing in the two experimental conditions - long and short profile, is the rating burden. Hence, in the first study, we expected that both research hypotheses were negated.

The second study was executed using a different version of the same recommender system, where dataset and interface were unchanged, but a *collaborative* algorithm denoted PureSVD was used. Collaborative algorithms recommend items on the basis of the ratings collectively provided by groups of users. For instance, a collaborative algorithm might predict a user's preference on an item based on the ratings that that item has received from users with similar taste. Previous research has shown that the accuracy of PureSVD is one of the highest, among the most well-known collaborative algorithms [3].

In order to strengthen the external validity of our research, we finally *replicated* the *same* study using another different version of the same RS, implementing a recommender algorithm of a strongly different nature - a *content-based* algorithm denoted by

[1] www.polirec.org

[2] www.contentwise.tv

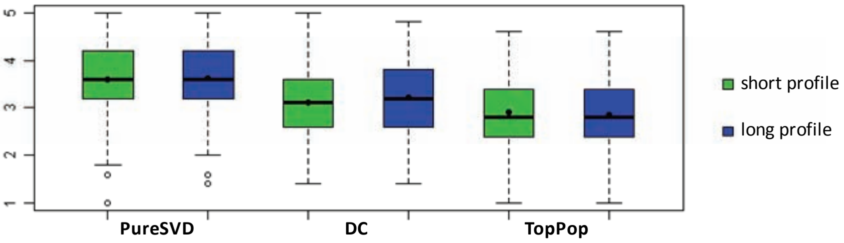


Figure 2. Perceived relevance.

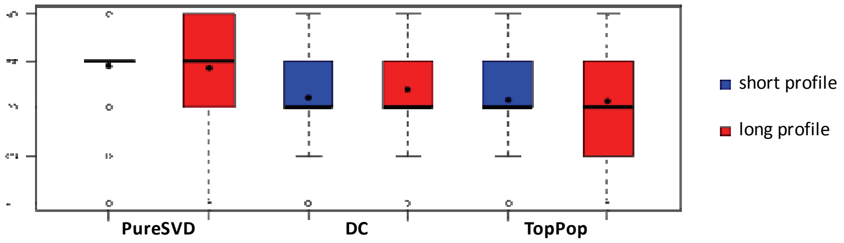


Figure 3. Global satisfaction.

as DirectContent (DC). As any recommender algorithm of the content-based category, DC recommends items whose content is similar to the content of items the user has positively rated in the past [11]. For instance, in the case of movies the content can be the title, the playing actors, the director, the genre, and the summary. DirectContent is a simplified version of the LSA algorithm described in [1].

In each study, participants were initially asked to provide their personal information (age, gender, education, nationality, and how many movies they watched per month). Afterwards, they were invited to browse the movie catalog using PoliRec, rating their degree of appreciation or interest for the movies encountered at any point during navigation (Fig. 1a), using a 1 to 5 scale (1 = low interest for or appreciation of the movie; 5 = high). Recommendations were generated once X ratings (X = profile length) were collected. Then users were invited to explore the recommendations, to score perceived relevance for each recommended item on a 1 to 5 scale (Fig. 1b) and finally to reply to questions regarding global satisfaction. Each user session lasted between 15 and 20 minutes, and took place in informal environments, such as university, interviewer's place, or interviewee's place. Test results did not present significant differences that can be referred to the execution context. Recruitment and data collection was carried out by a team of 60 master students, organized in 6 groups (2 groups per each study). They were selected among the best students attending the "Interactive TV" course at our School of Information Engineering. They 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. They were motivated to perform the evaluation to the best of their capabilities, as the work was constantly monitored and accounted for 20% of their grade in the courses.

3. RESULTS

Figure 2 shows the box plot of the perceived relevance for each algorithm. Upper and lower ends represent 25th and 75th percentiles. Whiskers extend to the most extreme data point, which is no more than 1.5 times the interquartile range. The median is depicted with a solid line, while the mean with a dot. Outliers are represented with empty circles. All the algorithms have an average relevance between 3 and 4. This shows that, on average, users were satisfied with the quality of the recommendations (the median for all the algorithms is greater than or equal to 3). Moreover, with PureSVD, 75% of the users received relevant recommendations.

The first notable result is that *there is no significant difference in the perceived relevance between users with short profile and users with long profile.* This finding is surprising, at least for PureSVD and DirectContent, because long profiles are considered to be correlated with increased accuracy of recommendations, which in turn *should* improve *perceived* relevance. In contrast, according to our study, *users with a 20 ratings profile receive recommendations that are perceived to have the same degree of relevance as the users with a 5 ratings profile.*

According to Fig. 3, similar results hold for *global satisfaction in the two personalized algorithms*: *there is no remarkable difference between users with short profile and users with long profile.* For the non-personalized algorithm TopPop, we can notice a slightly different behavior: among users with a *short* profile, only 25% were *not* satisfied with the experience (score lower than 2), a percentage that rises to 50% for the users with a long profile. This difference, although not statistically significant (as highlighted by Table 1), provides a rough estimate of the burden which the additional ratings have on the perceived quality

Table 1. Comparison tests between short and long profiles

	Perceived relevance short profile – long profile			
Algorithm	**Difference**	**Lower**	**Upper**	**P-Value**
PureSVD	-0.0352	-0.1765	0.1059	0.6233
DirectContent	-0.0865	-0.2369	0.0639	0.2589
TopPop	0.0470	-0.1017	0.1958	0.5341
	Global satisfaction short profile – long profile			
Algorithm	**Difference**	**Lower**	**Upper**	**P-Value**
PureSVD	0.0447	-0.1385	0.2280	0.6314
DirectContent	-0.1558	-0.3482	0.0365	0.1120
TopPop	0.04914	-0.1522	0.2504	0.6315

of the UX. In order to compare the results more analytically, we ran pair-wise comparison tests using Tukey's method. All tests were run using a significance level α = 0.05. The findings, described in Table 1, confirm what emerges from fig. 2 and 3: *having a short or long profile does not provide "better" recommendations in terms of any of the UX quality dimensions (i.e., perceived relevance and global satisfaction).*

4. DISCUSSION AND CONCLUSIONS

The results presented in the previous section confute both our research hypothesis in all the three replicated studies. They provide empirical evidence that, when increasing the user's profile length from 5 to 20 ratings, perceived relevance and global satisfaction do not increase. We can metaphorically say that the two contrasting forces generated by profile length on the quality of the UX mutually compensate: the potentially positive effects of increased accuracy resulting from a longer profile are eroded by the burden of a more demanding rating process. The validity of our findings is restricted to the actual experimental conditions considered. In addition, a weakness of our study is the limited number of user-centric attributes considered for UX quality. Still, in a field where empirical work is particularly complex and resource demanding, our research represents one of the widest and most articulated studies, and provides contributions for both RS design practice and the emerging research in user-centric evaluations of RSs. Our work can help designers to prioritize design decisions, suggesting that there is no real need of building systems that collect extremely long profiles (e.g., more than 5 ratings).From a theoretical perspective, our results provide an answer to our initial research question, but also highlight the discrepancy between what we know and what we need to know, pinpointing directions for future research. On the one hand, the study presented in the paper confirms our previous results [3][4][4] that show that accuracy metrics are "weak forces", less crucial than we may expect in improving user's perception of RS's quality. On the other hand, they suggest that also the burden of the rating process is, in absolute terms, a weak force: in the context of a non personalized algorithm (TopPop - study 1), as the profile length increases without any increment of accuracy, the overall satisfaction slightly decreases, but at a much less degree that we may expect. This may suggest that other "forces" exist that intervene in the complex trajectory from the experience of the system to quality perception [13]. Our work pinpoints the need for more research that integrates traditional and user-centric approaches to study the whole gamut of design characteristics of recommender systems that are likely to influence the quality of the UX. These factors include both pragmatic attributes, such as interface usability, and more hedonic attributes such as attractiveness, aesthetics, or engagement.

5. REFERENCES

[1] Bambini, R., Cremonesi, P. and Turrin, R. 2011. A Recommender System for an IPTV Service Provider: a Real Large-Scale Production Environment. *Recommender Systems Handbook* 2011: 299-331

[2] Pu, P., Chen, L., Hu, R.: A user-centric evaluation framework for recommender systems. In *Proc. RecSys'11,* ACM, New York, NY, 23–27

[3] Cremonesi, P., Garzotto, F., and Turrin, 2012. Investigating the Persuasion Potential of Recommender Systems from a Quality Perspective: an Empirical Study, *ACM Trans. on Interactive Intelligent Systems* (to appear)

[4] Cremonesi, P., Garzotto, F. Negro, S. Papadopoulos, A. Turrin, R. 2011. Comparative evaluation of recommender system quality. In *Proc. CHI 2011 Extended Abstracts*. ACM, New York, NY, USA, 1927–1932.

[5] Cremonesi, P., Garzotto, F. Negro, S. Papadopoulos, A. Turrin, R. 2011. Looking for "good" recommendations: a comparative evaluation of recommender systems. In *Proc. INTERACT 2011*. Springer-Verlag, , 152-168.

[6] Cremonesi, P., Koren, Y. and Turrin, R. 2010. Performance of recommender algorithms on top-n recommendation tasks. In *Proc. RecSys '10*. ACM, New York, NY, USA, 39–46.

[7] Cremonesi, P., Lentini, E., Matteucci, M., Turrin, R. 2008. An evaluation methodology for recommender systems. In *Proc. 4th Int. Conf. on Automated Solutions for Cross Media Content and Multi-channel Distribution*, 224–231.

[8] Jonathan, L., Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl, 2004. Evaluating collaborative filtering recommender systems. *ACM Trans. on Information Systems* 22 (2004), no. 1, 5–53.

[9] Jonathan L. Herlocker, Joseph A. Konstan, Al Borchers, and John Riedl, 2009. An algorithmic framework for performing collaborative filtering, In *Proc. SIGIR 2009*, ACM, New York, NY, USA, 230–237.

[10] Lekakos, G., Giaglis, G. M.. A hybrid approach for improving predictive accuracy of collaborative filtering algorithms. *User Modeling and User-Adapted Interaction* 17, 2007, 5–40

[11] Lops, P., De Gemmis, M., and Semeraro, G. 2011. Content-based recommender systems: State of the art and trends. In *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor (Eds). Springer US, 73–105

[12] Mcnee A., Lam S., Konstan J. , Riedl J. 2003. Interfaces for eliciting new user preferences in recommender systems. In Proc. *UM 2003,* Springer-Verlag Berlin, Heidelberg. 178–188.

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

[14] Rashid, A. M., Karypis, G., and Riedl, J. 2008. Learning preferences of new users in recommender systems: an information theoretic approach. *SIGKDD Explorer Newsletter 10*, 90–100.

[15] Sarwar, B., Karypis, G., Konstan, J. and Reidl, J.. 2001. Item-based collaborative filtering recommendation algorithms. In *Proc, WWW 2001*. ACM, 285–294

[16] Shani, G. and Gunawardana, A. 2011. Evaluating recommendation systems. In *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Springer US, 2011

[17] Berkovsky, S., Eytani, Y., Kuflik, T., Ricci, F.. 2007. Enhancing privacy and preserving accuracy of a distributed collaborative filtering. In *Proc. RecSys '07*. ACM, New York, NY, USA, 9-16.

[18] Weng, Li T. , Xu Y., Li Y., Nayak, R., 2008. An Efficient Neighbourhood Estimation Technique for Making Recommendations. *ICEIS LNBIP*, 19, 5, Springer 253-264

[19] Ziegler, C-N., McNee, S.M., Konstan, J.A., Lausen, G., 2005. Improving recommendation lists through topic diversification. *Proc. WWW '05,* ACM, New York, 22–32