User Effort vs. Accuracy in Rating-based Elicitation

Paolo Cremonesi
Politecnico di Milano
p.zza L.da Vinci 32, Milano
Italy
paolo.cremonesi@polimi.it

Franca Garzotto
Politecnico di Milano
p.zza L.da Vinci 32, Milano
Italy
franca.garzotto@polimi.it

Roberto Turrin
Moviri
via Schiaffino 11, Milano
Italy
roberto.turrin@moviri.com

ABSTRACT
One of the unresolved issues when designing a recommender system is the number of ratings – i.e., the profile length – that should be collected from a new user before providing recommendations. A design tension exists, induced by two conflicting requirements. On the one hand, the system must collect "enough" ratings from the user in order to learn her/his preferences and improve the accuracy of recommendations. On the other hand, gathering more ratings adds a burden on the user, which may negatively affect the user experience. Our research investigates the effects of profile length from both a subjective (user-centric) point of view and an objective (accuracy-based) perspective. We carried on an offline simulation with three algorithms, and a set of online experiments involving overall 960 users and four recommender algorithms, to measure which of the two contrasting forces influenced by the number of collected ratings – recommendations relevance and burden of the rating process – has stronger effects on the perceived quality of the user experience. Moreover, our study identifies the potentially optimal profile length for an explicit, rating based, and human controlled elicitation strategy.

Categories and Subject Descriptors
H.5.2 [**Information Interfaces and Presentation**]: User Interfaces; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms
Design, Experimentation, Human Factors

Keywords
Elicitation, Profile Length, New User Problem, User-Centric Evaluation, Perceived Relevance, Perceived Quality, Accuracy

1. INTRODUCTION
Whenever a *new user* joins a Recommender System (RS), the system tries first to learn her preferences in order to provide personalized recommendations as soon as possible. This *preference elicitation process* is fundamental both at cold-start time (i.e., when bootstrapping a new RS) and during the normal operational life of the system, and has effects along multiple dimensions. The preference elicitation strategy can affect the "new user utility" (how well the system can make good recommendations to the new user who is undergoing the elicitation process) and the "system or community utility" (how well the system can provide good recommendations to *all* users, given what it learns from the new user)[12][18][27]. In addition, the elicitation process represents the user's initial experience with the recommender and is crucial to shape her attitude towards the system and her decision process and behavior (i.e., what she will do with the recommendations).

A wide amount of studies have explored different techniques for preferences elicitation (considering, for example, which questions to ask a new user, which and how many items to propose, in which form and order). A number of design criteria have been identified for making this process more effective in terms of both (new user and community) utility and the quality of the user interaction with the RS. Maximizing both utility and quality of use are somehow conflicting requirements. Obviously, the system needs to learn from new users and to collect enough preferences to generate good and satisfying recommendations; not gathering enough information can result in a poor user model, which may lead to limited accuracy of recommendations and in turn may negatively affect the quality of the user interaction with the RS. Still, requiring users to spend too much time and energy with the system before they receive any recommendation can be annoying, and cause some users to give up the sign up process. Hence the developers of elicitation strategies must face a potential *design tension*: to raise utility by increasing the amount of information gathered from new users, and to make the elicitation process smooth from a user interaction perspective, limiting complexity and user effort during sign-up tasks.

Finding a compromise that solves this tension is not obvious, and represents an unsolved issue in current research on RSs. This paper investigates this challenge for a specific category of elicitation strategies, which can be referred to as explicit, rating based, and human controlled. An *explicit* elicitation process means that the system learns from specific facts provided by the new users about their taste and preferences. In explicit *rating based* elicitation processes, such facts are user's opinions, i.e., binary or multi-scale ratings, on a set of items that, in *human controlled* methods, are selected by the users themselves. In the context of this kind of strategies, one possible measure of the user effort during the elicitation process is the *profile length* - the number of ratings the new user must provide to the system before starting receiving recommendations. Hence we focus on the trade-off that exists between *maximizing* the *user utility* and *minimizing* the rating effort. More precisely, we explore the following research question: "*Which of the two potentially contrasting "forces" that depend on profile length – user utility and user effort – have stronger effects on the perceived quality of the user interaction?"*

There are two implicit assumptions in the above research question, which are intuitive but not always confirmed by prior studies. The first assumption is that profiles length positively affects user utility. Some works show that profile length of new

RecSys'12, September 9–13, 2012, Dublin, Ireland, UK.

users is positively correlated to the accuracy of recommendations in term of user utility [12][13][4]. However, this result cannot be easily generalized, as its supporting experiments are limited to item-based collaborative algorithms, and accuracy is measured only in terms of error metrics: RMSE [13] and MAE [4]. Moreover, [28] finds that the correlation between *profile length* and utility is not always present, but it depends on the elicitation strategy adopted. These studies instill some doubts on the general assumption that a longer profile corresponds to more accurate recommendations. We may wonder, for example, to what extent we can claim that the fallout of a content-based recommender algorithm improves with the profile length.

The second assumption in the above research question is that profile length negatively affects the perceived quality of the user interaction because of the rating burden. As we discuss in the next section, several studies have explored the relationship between the characteristics of the preferences elicitation strategy, the user effort during sign-up, and the user interaction with the RS. Still, existing empirical findings reveal some discrepancies and provide different results in different experimental contexts. According to some authors, there is a negative force induced by the user effort which seems to dominate in the user interaction with respect to the force originated from increased utility. In contrast, other studies suggest that as users become aware of the better quality of the recommendations that result from a richer amount of preferences, they somehow feel that the system best understand their taste, and tend to not perceive the extra burden.

For all these reasons, before exploring the general question we need to address two *preliminary* research questions:

"Does the accuracy of a recommender algorithm increase with the profile length?"

"*Does the increased burden of ratings collection affect perceived quality of user interaction?"*

The three research questions have been addressed by carrying on *three main studies*, involving off-line and on-line experiments that have been replicated in different experimental conditions, involving overall *four recommender algorithms* and *960* users.

The rest of the paper is organized as it follows. The next section provides an overview of the state of the art which is more relevant for the scope of our research, pinpointing the contrasting results that emerge from the current literature. Section 3 describes the design of the three studies. Section 4 presents the findings and discusses the key results. Section 5 draws the conclusions and outlines directions for future work.

2. RELATED WORK

It is generally acknowledged that the strategy and the interface designed to elicit new users' preferences influences the perceived quality of the user interaction with the RS, and has impact on users' decision accuracy and the intention to return [5]. Hence a wide amount of studies have explored the elicitation process, trying to understand how the construction of preferences process takes place [26], which questions to ask a new user [27], which items to propose [26], in which form and order [26]. For more exhaustive reviews of these issues, the reader is referred to [5]. In this section, we shortly outline the works that are more relevant for the context of our research.

In their review [24], the authors discuss the tradeoff between accuracy versus effort, and suggest "minimizing preferences elicitation in the profile initialization". The arguments for this design guideline are both theoretical and empirical. According to behavioral decision theories [14], users are likely to settle on the immediate benefit of saving effort over the delayed gratification of higher accuracy. A number of works discussed in [28] support this principle (e.g., [12][16]), which is also confirmed by a more recent online study presented in [13], which pinpoints that, in a content-based recommender, a higher perceived system effectiveness is related to reduced effort in the elicitation activity, measured in terms of amount of browsing before receiving recommendations.

Still, not all studies confirm the above guideline. While some works show that the risk in requiring users to provide too much information is to annoy them [19], or to have them give up the sign up process [22][28], other researches show that users are willing to face a more complex elicitation if they feel they are rewarded with useful recommendations [28].

Most authors, such as [27] and [28], consider *system controlled* elicitation methods and explore different measures to select items for the user to rate (e.g., popularity, entropy). These authors conducted a set of off line experiments to evaluate, in terms of accuracy, the strength and weakness of the different item selection strategies. They then compared these results with those ones emerging from online studies where they collected users' opinion about the perceived effort of the signup process in the different experimental conditions. Their findings show that, even if different item selection measures caused an objective variation of effort (measured by the number of pages the user must see before starting to get recommendations), users seemed not to notice the extra burden; hence the authors suggest that the initial recommendation quality (i.e., accuracy) and not the user effort should be considered as the deciding factor to judge (and choose) the desired elicitation strategy.

Similar results are outlined in [29], where experiments show that more elicited ratings do not necessarily imply more perceived effort. However, these findings have a different motivation with respect to [27] and [28]. Users in [29] perceive a low effort with poor quality recommender algorithms, even in the case of a very long elicitation process, as they feel the need to provide more ratings to the RS in order to improve quality.

The experiments described in [22] explore the design tradeoff between user effort and the benefits it brings either to the system (who needs to learn about the user) and the user (who needs to receive useful or convincing recommendations) in different conditions of user control during the elicitation process. The authors compare three interfaces to elicit information from new users, respectively using a system-control method, where the system proposes the lists of items to evaluate, a user-control method, where the user herself selects the items to rate, and a hybrid, mixed-initiative method (a combination of the other two methods). For each interface, they measured the quality of the user models, using a common measure of recommendation accuracy (MAE), and, through a survey, users' perception of the complexity and burden (time/effort) of the sign-up process. They found that the two "pure" interfaces both provide accurate user models. Still, users in the user controlled elicitation group who completed the sign up process were 8-10% less than the other groups and spent twice more time, which indicates that the extra burden had a significantly negative effect on these subjects. On the other hand, the persons who completed the process thought that the system best understood their taste, felt more motivated being in charge of the process, and did not feel the extra effort. This result is confirmed by a study reported in [13], which tests the effects on the user interaction of explicit vs. implicit elicitation methods. These findings show that explicit control over preferences elicitation, in spite of the extra burden on the user, leads to a slightly higher perceived recommendation quality. In addition, in the same study the system effectiveness is judged

Table 1.Studies at-a-glance

Study	Type	Research scope	Metric (dependent variable)	Research question	Algorithms	Users	Profile lengths
1	Off-line simulation	Profile length vs. new user utility	Accuracy (recall and fallout)	*does the accuracy of a recommender algorithm increase with the profile length?*	PureSVD AsySVD DirectContent	60,000 (simulated)	5 – 40
2	On-line user experiment	Burden of the rating process	Global satisfaction	*does the increased burden of ratings collection affect perceived quality of user interaction?*	TopPop	60 (total)	5,10
3	On-line user experiment	Design tension between utility and burden	Global satisfaction Perceived relevance	*which of the two potentially contrasting "forces" that are created by the profile length – user utility and user effort – have stronger effects on the perceived quality of the user interaction?*	PureSVD DirectContent	900 (total)	5,10,20

higher by participants who rate more items (as they noticed an increase of accuracy due to a wider system's knowledge about them).

A number of works have analyzed and compared the cognitive effort related to non-rating-based sign-up processes (tagging items [12], elicitation of user preferences on product features [2], personality quiz [15], and affective feedback [26]). In the experiments described in [26], for example, cognitively less demanding elicitation methods were perceived low in effort and high in liking. Still, follow-up studies reported in the same paper, which explored the trade-off between giving detailed preference feedback and effort, show that users are willing to spend more effort if the feedback mechanism enables them to be more expressive. This provides some insights on the intrinsic motivational factors that lead people to spend more effort to give more detail about a preference.

Few works have studied the impact of the *profile length*, i.e., the number of collected ratings, on the accuracy of recommendations, and results are sometimes contrasting. Some authors highlight that *profile length* is positively correlated to the accuracy of recommendations, both in term of:

(i) *new user* utility [13][4], measured on collaborative RSs with error metrics such as MAE and RMSE;
(ii) *community* utility [11][12], such as eliciting ratings for movies that don't have many, or committing users to do more valuable work for the community (e.g., tagging content and posting comments).

Surprisingly, [28] finds that the correlation between *profile length* is not always present, but depends on the elicitation strategy adopted.

3. THE DESIGN OF THE STUDIES

The three research questions presented in Section 1 (Introduction) have been explored in three main (sub)studies - one off-line simulation and two on-line experiments – summarized in Table 1.

3.1 Study 1: Accuracy

The first study analyzes the accuracy of three recommender algorithms as a function of the new user profile length.

For the evaluation, we used a subset of the Netflix dataset. Our subset consisted of 6,500 items and about 8.8 million ratings given by 250,000 users. In addition, for the purpose of using a content-based algorithm, the dataset was integrated with metadata collected online (e.g., genre, actors, director). The subset was created by extracting movies for which we were able to find the complementary data. The data was added automatically, yet their quality was manually checked to cleanup any possible redundancy[1].

As recent research founds that improvements in MAE and RMSE are not necessarily the path to improvements in the user experience [18], accuracy has been measured by using information retrieval metrics that are wider adopted in the evaluation of commercial RSs [10]. In particular we focused our attention on *recall* (the percentage of relevant items that are recommended to a user) and *fallout* (the percentage of non-relevant items that are recommended to a user). We did not include precision as accuracy metric because it cannot be estimated in a reliable way unless all ratings are known for all users and all items. Most datasets contains a large number of unrated items: as these are considered irrelevant, they miss a fraction of unknown positive relevance, and lead to precision underestimation [3].

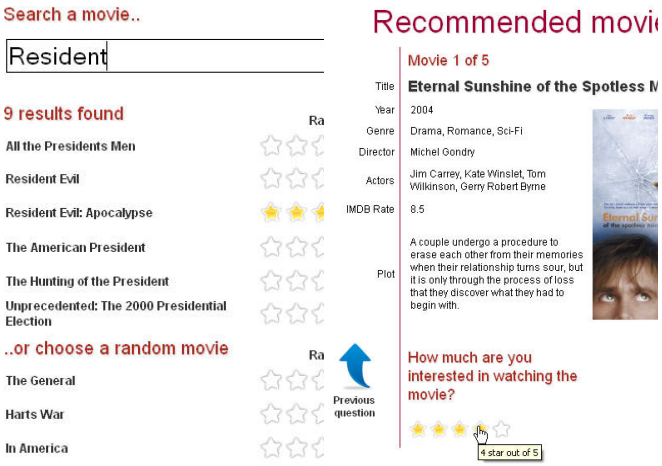
The study considers accuracy vs. profile length for *three algorithms*: two collaborative algorithms (PureSVD and AsySVD) and one content-based algorithm (DirectContent). PureSVD and AsySVD are based on matrix-factorization and previous research shown that their accuracy is one of the best [7][10]. DirectContent recommends items whose content is similar to the content of items the user has positively rated in the past [21]. 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].

The testing methodology adopted in this study is a modified version of the technique described in [10]. Users in the dataset are randomly split into two subsets: *training* set (70% of the users) and *test* set (30% of the users). The test set is further modified by randomly removing 30% of ratings from each user's profile. The removed ratings are the *probe* set, while the modified test set is used to simulate new users. The test set contains 75,000 users and 60,000 of them have a user profile with less than 40 ratings (after removing the probe set).

In order to measure recall, we first trained the algorithm using the ratings in the training set. Then, for each user in the test set and for each item in the probe set that was rated 5-stars by the user, we followed these steps:

- We randomly selected 1,000 additional items that were not rated by the user. We assumed that the user was not interested in most of them.
- We predicted the ratings for the 5-stars rated item and for the additional 1,000 items.

[1] The dataset is available for free download at the following address:
http://home.dei.polimi.it/cremones/recsys/Enriched_Netflix.zip
When using the dataset, please cite this paper.



(a) Providing explicit ratings during navigation

(b) Providing relevance scores for a recommended item

Figure 1. PoliRec framework

- We formed a top-5 recommendation list by picking the 5 items with the largest predicted ratings.

Overall, we generated a number of recommendation lists equal to the number of 5-stars ratings in the probe set. For each list we had a *hit* (e.g., a successful recommendation) if the 5-star rated item was in the list, because we can reasonably state that the item was relevant to the user. Therefore, the overall recall was computed by counting the number of hits (i.e., the number of successful recommendations) over the total number of recommendations

$$\text{recall} = \frac{\#\text{ times the removed 5 stars item is in the list}}{\#\text{ recommendation lists}}$$

Recall is defined as the percentage of items interesting for the user that have been effectively recommended by the system.

A similar approach was used to measure fallout, with the only difference being that we selected 1-stars ratings from the probe set, as we can reasonably state that these ratings refer to items not relevant to the users. The fallout was computed as

$$\text{fallout} = \frac{\#\text{ times the removed 1 stars item is in the list}}{\#\text{ recommendation lists}}$$

Fallout is defined as the percentage of items uninteresting for the user that have been erroneously recommended. Recall and fallout range from 0% to 100%. An ideal algorithm should be able to recommend all interesting items (i.e., recall equals 100%) and to discard all uninteresting items (i.e., fallout equals 0%).

3.2 Study 2 (Burden)

The second experiment investigates the impact of the profile length on the *perceived quality* of the interaction with a RS, measured in term of *global satisfaction*. Global satisfaction is an indicator of how users feel about the overall experience with the system, and represents an important quality factors in user centric approaches to RS evaluation [6].

We measured 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, the same algorithm, and the same user interface, but a *different rating process*, asking users to rate a *different* number of movies. In other words, the two experimental conditions were characterized by different profile lengths (*independent variable*), respectively 5 and 10. As we wanted the only difference in the two experimental conditions to be the objective user effort (measured by the profile length), we considered a *non-personalized* algorithm, which recommends the same predefined list of items to everybody, regardless his or her user profile, hence not sensitive to profile length. Specifically, we used a simple, non-personalized algorithm (*TopPop*), which recommends top-N items with the highest popularity (largest number of ratings) [7]. As the accuracy of recommendations generated by TopPop does not depend on the user profile, the only measured force playing in the two experimental conditions – long and short profile – is the rating burden. Hence, in this study, we expect the perceived quality to decrease with the profile length.

3.2.1 Instruments

We used the web-based recommender and evaluation framework PoliRec, 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 and Lovefilm. 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 select and apply a specific recommender algorithm among the three that we considered, and to 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.

3.2.2 Participants

This empirical research involved 60 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. None of them had been previously exposed to the system of our study, and none of them had any technical knowledge about RSs.

3.2.3 Procedure

Each participant was initially asked to provide his/her personal information (age, gender, education, nationality, and how many movies they watched per month). Afterwards, users were invited to browse the movie catalog using PoliRec, rating his/her 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.

All the users were told they were receiving personalized recommendations on the basis of their input ratings, although all users were receiving exactly the same list of 5 top popular movies. Finally, each user was invited to explore the recommendations, to score perceived relevance for each recommended item on a 1 to 5 scale (Fig. 1b) and to reply to questions regarding global satisfaction.

Each user session lasted between 15 and 20 minutes, and took place in informal environments, such as the university, the interviewer's place, and the 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 PhD student of the Computer Science Engineering Department at Politecnico di Milano, as part of his PhD research.

[2]www.polirec.org – www.contentwise.tv

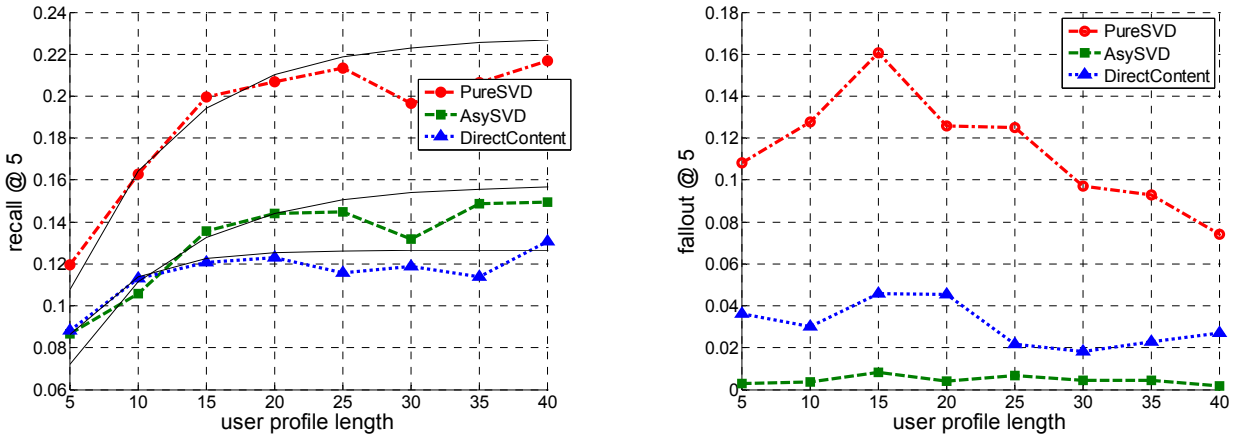


Figure2. Study 1: Off-line simulation: Recall and Fallout of new users on the Netflix dataset as a function of the profile length

3.3 Study 3 (Tension)

In this final study we analyze the *combined* effect on the user interaction of variations in utility (accuracy) and rating burden.

The third research was designed as *two replicated* between-subjects studies. In each study, we measured user's *perceived quality* of a RS in the movie domain in *three* different experimental conditions. Similarly to the previous study, the three experimental conditions were characterized by different profile lengths (*independent variable*), respectively 5, 10 and 20.

User's perceived quality has been operationalized in terms of two measurable factors (*dependent variables*): *perceived relevance* and *global satisfaction* (defined in the previous section). Perceived relevance measures how well the user believes that recommendations match his or her interests, preferences, and taste, and, similarly to global satisfaction, is acknowledged as an important quality factors in user centric evaluation framework [6].

3.3.1 Participants

The overall empirical research involved 900 subjects over a period of two months (December 2011 - January 2012). In *each* of the two replicated studies, we involved 450 subjects who were split in three groups of the same size, and randomly assigned to either experimental condition 1 (profile length = 5 ratings), experimental condition 2 (profile length = 10 ratings) or experimental condition 3 (profile length = 20 ratings). 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 any technical knowledge about RSs.

3.3.2 Procedure

The third study was executed using the same system used in Study 2 (PoliRec and ContentWise), but two personalized recommender algorithms were used: PureSVD and DirectContent. PureSVD has been chosen because, according to Study 1, it exhibits the best accuracy in term of recall. DirectContent has been chosen as representative of content-based algorithms.

The procedure adopted for this study was identical to the one adopted for Study 2 (Burden), but recruitment and data collection were carried out by a team of 45 master students, organized in 6 groups (2 groups per each experimental condition). 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 course.

4. DISCUSSION ON RESULTS

In this section we analyze and discuss the results of the three studies.

4.1 Study 1: Accuracy

From the analysis of Figure 2, we can observe that accuracy in terms of recall clearly improves with the number of ratings in the new user profile for the three tested algorithms. On the contrary, accuracy in terms of fallout does not change significantly, with the exception of PureSVD, where we observe an improvement (i.e., a decrease) in the fallout values.

We can also observe that the increase in recall seems to be bounded by an asymptotic limit, this limit being different for different algorithms. In order to find this asymptotic value, for each of the tested algorithms we have fitted recall with the exponential function below

$$r(l) = a(1 - e^{-bl})$$

where r is the recall, l is the number of ratings in the new user's profile, and a and b are two unknown parameters. Parameter a represents the asymptotic value of the recall, i.e., the maximum recall achievable by an algorithm in the hypothesis of having a very large number of ratings in the user's profile. Parameter b represents the speed at which recall increases towards its maximum limit.

By using least squares, we have fitted the exponential model to the recall data. The resulting parameters are listed in Table 2 and the corresponding exponential functions are plotted in Fig. 2 as continuous lines. The same table shows the profile length values for which recall reaches 80% of its maximum value. *This maximum value is close to 10 ratings for all of the three algorithms*. Hence we should expect that *profile lengths longer than 10 ratings do not increase user perceived relevance in the recommendations* – a hypothesis that we will further analyze in light of the results of Study 3.

Table 2.Study 1: Exponential fitting of recall

	a	*b*	number of ratings at 80% max recall
PureSVD	0.2281	0.1275	13
AsySVD	0.1579	0.1223	13
DirectContent	0.1265	0.2291	7

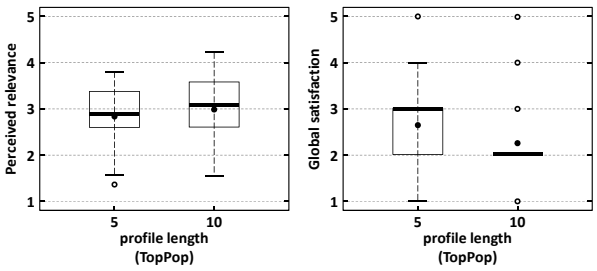


Figure 3.Study 2: Burden of the rating process (non personalized algorithm).

4.2 Study 2: Burden

Figure 3 shows the box plot of perceived relevance and global satisfaction in the two experimental conditions (respectively asking users to rate 5 and 10 movies, but proposing them exactly the same list of 5 top popular movies). 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.

As highlighted by Figure 3, there is *no significant difference* in *perceived relevance* between users with short profile and users with long profile. Moreover, in both cases only 50% of users are moderately satisfied with the quality of the recommendations (perceived relevance greater than 3) and almost no user is greatly satisfied (perceived relevance greater than 4). These findings are somehow expected, as all users were receiving the same list of recommended movies, without any attempt to match their preferences.

Still, our findings show that there is a *significant difference* in *global satisfaction* between users with short profile and users with long profile ($p < 0.1$). This finding is not surprising, and answers to research question 2: in the absence of extra benefits, an increased burden of ratings collection *negatively* affects perceived quality of user interaction, as users with a longer profile have a longer sign-up process with respect to the short profile users but no increase of quality benefits in terms of better recommendations.

4.3 Study 3: Tension

Figure 4 highlights the key results of our investigation on the combined effect on the user interaction of the positive force induced by better accuracy (as confirmed by Study 1) and the negative force induced by increased rating burden (as confirmed by Study 2). Figure 4 shows the box plot of the perceived relevance for both algorithms considered in Study 3. Both of them have a mean relevance between 3 and 4 (the median for all algorithms is greater than or equal to 3). This shows that, on average, users were satisfied with the quality of the recommendations generated by both algorithms in all the three experimental conditions (profile length 5, 10, and 20).

The first notable result is that *perceived relevance changes with the profile length, reaching its maximum when users rated 10 items.* This results is true for both the collaborative (PureSVD and content (DirectContent) algorithms. According to Figure 5, a similar result does not hold for *global satisfaction,* which is substantially unchanged for all algorithms and all profile lengths.

In order to compare the results on perceived relevance more analytically, we ran pair-wise comparison tests using Tukey's method. All tests were run using a significance level $\alpha = 0.1$. When looking at the PureSVD algorithm, *the perceive relevance for new users with 10 ratings in their profile is significantly better than the relevance perceived by users with either 5 or 20 ratings* ($p < 0.01$). The same applies for the DirectContent algorithm ($p < 0.1$).

4.4 Discussion

The results of the first two studies provide a somehow expected answer to their respective research questions: both accuracy of recommendations and perceived user effort do increase with the number of ratings elicited during the sign-up of new users.

More specifically, the study on accuracy (Study 1) partially supports previous results that investigated the relationship between accuracy and profile length in terms of error metrics (MAE and RMSE) for two item-based recommender systems [12][13]. We have provided empirical evidence that recall improves with the profile length, and this correlation exists for different algorithms. However, the same does not happen for *fallout, which is not correlated with the profile length*.

It is interesting to compare the results of Study 2 with the findings of Study 3. According to Study 2, an increased user effort during the sign-up process negatively affects the perceived quality of user interaction, in terms of global satisfaction, if the extra burden is not compensated by an increased utility (i.e., improved relevance of recommendations). This effect is visible in our experiment because of the low quality (relevance) of the recommendations provided by the non-personalized algorithm we adopted. Users were little rewarded by useful recommendations, regardless of the higher number of ratings, and they were more susceptible to feel the additional burden.

The same phenomenon does not occur in Study 3, where users receive good-quality (relevant) recommendations. If a more demanding rating process is balanced by significantly better recommendations, the global satisfaction is not affected negatively by the increased effort. It is as if the two contrasting forces (accuracy and user effort) generated by profile length on user interaction quality mutually compensate. The potentially positive effects of increased accuracy resulting from a longer profile is eroded by the burden of a more demanding rating process, but this effect is not strong enough to decrease the global opinion of the users towards the recommender system.

Still, when comparing the values of a different indicator of perceived user interaction quality – perceived relevance – in the different profile length conditions, a different phenomenon can be noticed (see Figure 5). With 5 and 10 ratings, we can observe that perceived relevance *increases* with the number of ratings – this was expected, because of the more accurate recommendations. However, when comparing profile lengths with *10 and 20 ratings*, we observe a somehow surprising behavior: *perceived relevance decreases*. This finding confirms our intuition that, as the relevance of recommendations does not increase indefinitely with the profile length, there should be a maximum number of ratings that can be elicited from a new user without having the negative force induced by increased burden overcome the positive force of relevance. The result is also coherent with the findings of Study 1, which pinpoint (see Table 2) that the maximum value for recall is close to 10 ratings for the algorithms considered in Study 3, and therefore suggest that profile lengths longer than 10 do not increase perceived relevance of the recommendations. Still, the motivation of this result from a user interaction perspective is not obvious. A possible interpretation is that the two different quality factors – global satisfaction and perceived relevance – concern different spheres of the user experience. Global satisfaction is a form of "perception", which denotes, in the terminology of [17], whether certain objective aspects of the interaction with a system register with the user at all; perceived relevance is a form of

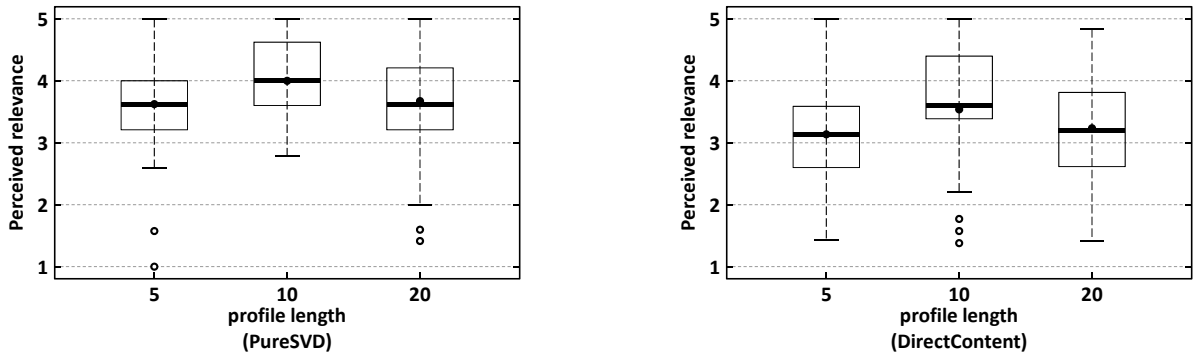


Figure 4.Study 3: Perceived relevance per algorithm and experimental condition (profile length 5-10-20).

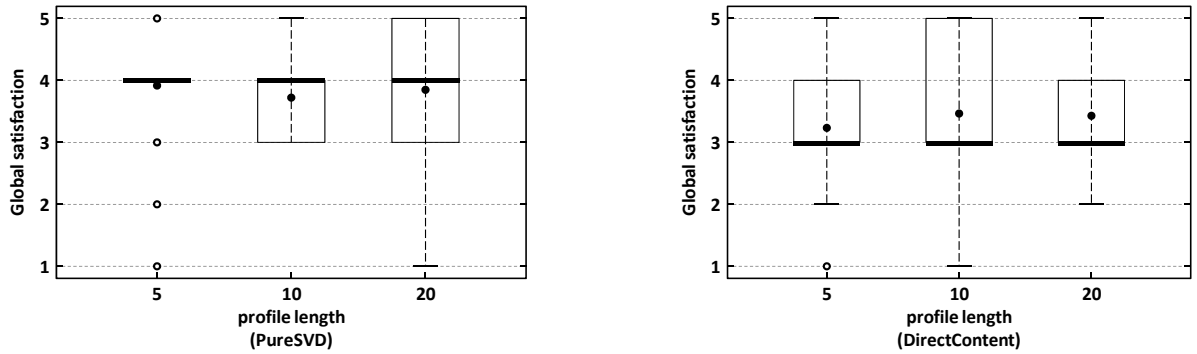


Figure 5.Study 3: Global satisfaction per algorithm and experimental condition (profile length 5-10-20).

"evaluation", which denotes whether a perceived aspect has any personal value for the user. An hypothesis could be that when the users' judgment is related to perception - represented by global satisfaction - all perceived aspects of the experience with a system are integrated into a more holistic perspective, and mutually compensate, unless some forces are significantly stronger or weaker than others, as it happens in Study 2. When the user's judgment moves towards a more value-oriented reflection – as expressed by perceived relevance – the mechanism pinpointed by [24] comes into play, and users are likely to focus on the immediate benefit of saving effort over the less (for them) measurable benefit of higher accuracy.

5. CONCLUSIONS

The studies reported in this paper shed a light on some unresolved issues in RS design: how can the system collect "enough" information (ratings) from the user, in order to learn her/his preferences and improve the quality of recommendations without adding an excess of burden on the user, which may dissolve the perceived benefits of good personalized recommendations? Our work has investigated this problem from multiple perspectives, in the context of rating based user controlled elicitation techniques, thru a set of vast offline and online studies. The validity of our findings is restricted to the actual algorithms and experimental conditions considered. In addition, a weakness of our work is the limited number of user-centric attributes considered for RS quality with respect of the spectrum of user interaction factors proposed by emerging frameworks for the evaluation of RS user experiences [6][17]. Still, in a field where empirical work is particularly complex and resource demanding, our research represents a wide and articulated study that bridges user-centric online evaluation with offline evaluation, and provides contributions for both RS research and design practice.

From a design perspective, our findings (Studies 1 and 3) suggest that the optimal number of ratings in the movie domain is between 5 and 20 ratings (more likely 10 ratings): it is within this range that the contrasting forces induced by profile length achieve a better balance, from a user interaction quality perspective. This result can be distilled into the heuristic: "10 ratings are enough", which can help designers to prioritize design decisions and suggests that there is no real need of building systems that collect extremely long profiles.

From a theoretical perspective, our findings (Study 1) show that it cannot be given for granted that the objective (i.e., statistically measured) quality of recommendations improves with the increase of profile length. We have provided empirical evidence, for different algorithms, that a positive correlation exists between recall and profile length. Still, in our experiments the same phenomenon does not happen for fallout, which is not correlated with the profile length; this infers that recommendations errors depend on the algorithm only, and are largely independent of the number of ratings collected from new users. Finally, the findings of Study 3 confirm and extend some of our previous results [7][8][9]: accuracy metrics measure "weak" user interaction forces, which are less crucial than we may expect in improving the user's perception of a recommender quality. In addition, our experiments indicate that also the burden of the rating process is, in absolute terms, a weak force, not able to strongly decrease the overall satisfaction of users if they are provided with useful recommendations. This may suggest that other "forces" exist, yet to be fully investigated, that intervene in the complex trajectory from the experience of the system to users' quality perception and evaluation of RS interaction [23]. This whole topic deserves a wider and more systematic exploration and calls for the definition of appropriate conceptual frameworks to help us building deeper and more coherent interpretations of empirical results and increasing our understanding of RS elicitation.

6. 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*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Springer US, 299-331.

[2] Adomavicius, G. and Tuzhilin, A., 2005. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Trans. on Knowl. and Data Eng.* 17(6) June 2005, 734-749.

[3] Bellogn, A., Castells, P., and Cantador, I., 2011. Precision-oriented evaluation of recommender systems: An algorithmic comparison. In *Proc. of RecSys'11*. ACM, New York, NY, USA, 333–336.

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

[5] Chen,L. and Pu, P., 2009. Interaction design guidelines on critiquing-based recommender systems. *User Modeling and User-Adapted Interaction* 19(3) August 2009, 167-206.

[6] Hu, R., and Pu, P., 2011. Enhancing collaborative filtering systems with personality information. In *Proc. of RecSys'11*. ACM, New York, NY, USA, 197-204.

[7] Cremonesi, P., Garzotto, F., and Turrin, R., 2012. Investigating the Persuasion Potential of Recommender Systems from a Quality Perspective: an Empirical Study, *ACM Transactions on Interactive Intelligent Systems*, 2(2) June 2012, 1-41.

[8] Cremonesi, P., Garzotto, F. Negro, S. Papadopoulos, A. Turrin, R., 2011. Looking for "good" recommendations: a comparative evaluation of recommender systems. In *Proc. of Human-computer interaction (INTERACT'11*). Vol. Part III. Springer-Verlag, 152-168.

[9] Cremonesi, P., Garzotto, F. Negro, S. Papadopoulos, A. Turrin, R., 2011. Comparative evaluation of recommender system quality. In *Proc. of annual conf. extended abstracts on Human factors in computing systems*. CHI EA '11. ACM, New York, NY, USA, 1927–1932.

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

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

[12] Drenner, S., Sen, S., and Terveen, L., 2008. Crafting the initial user experience to achieve community goals. In *Proc. of RecSys'08*. ACM, New York, NY, USA, 187-194.

[13] Golbandi, N., Koren, Y., and Lempel, R., 2010. On bootstrapping recommender systems. In *Proc. of the 19th ACM Int. Conf. on Information and knowledge management*. ACM, New York, NY, USA, 1805-1808.

[14] Haubl, G., Trifts, V., 2000. Consumer decision making in online shopping environments: the effects of interactive decision aids. *Mark. Sci.*19, 4–21.

[15] Hu, R. and Pu., P., 2009. A comparative user study on rating vs. personality quiz based preference elicitation methods. In *Proc. of Int.l Conf. on Intelligent user interfaces* (IUI '09). ACM, New York, NY, USA, 367-372.

[16] Jones, N., Pu, P., 2007. User technology adoption issues in recommender systems. In *Proc. of Networking and Electronic Commerce Research Conf.* (NAEC '07), 379–394.

[17] Knijnenburg, B.P., Willemsen, M.C., Gantner, Z., Soncu, H. and Newell, C., 2012. Explaining the user experience of recommender systems, *User Modeling and User-Adapted Interaction*, 22(4-5) March 2012, 441-504.

[18] Konstan, J.A., Riedl, J., 2012. Recommender systems: from algorithms to user experience. *User Model. User-Adapt. Interact*. 22, 101-123.

[19] Lekakos, G., Giaglis, G. M., 2007. A hybrid approach for improving predictive accuracy of collaborative filtering algorithms. *User Modeling and User-Adapted Interaction* 17(1-2) March 2007, 5–40.

[20] Liu, N.N., Meng, X, Liu, C., and Yang, Q., 2011. Wisdom of the better few: cold start recommendation via representative based rating elicitation. In *Proc. of RecSys'11*. ACM, New York, NY, USA, 37-44.

[21] 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.

[22] Mcnee, S., Lam, S., Konstan, J., and Riedl, J., 2003. Interfaces for eliciting new user preferences in recommender systems. In *Proc. of Int. Conf. on User Modeling*, 178–188.

[23] 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 on Human factors in computing systems*, ACM, New York, NY, USA, 1097–1101.

[24] Pu,.P, Chen, L., Hu, R., 2012. Evaluating recommender systems from the user's perspective: survey of the state of the art. *User Modeling and User-Adapted Interaction*, 22(4), 317-355.

[25] Pu P. Chen L, Hu R., 2011. A User-Centric Evaluation Framework for Recommender Systems. In *Proc. of RecSys'11*, ACM, New York, NY, USA, 157-164.

[26] Pommeranz, A., Broekens, J., Wiggers, P., Brinkman, W.P.,Jonker, C.M., 2012. Designing interfaces for explicit preference elicitation: a user-centered investigation of preference representation and elicitation process, *User Modeling and User-Adapted Interaction*, 22(4-5) 15 March 2012, 357-397.

[27] Rashid, A.M., Albert, I., Cosley, D., Lam, S.K., McNee, S.M., Konstan, J.A., and Riedl, J., 2002. Getting to know you: learning new user preferences in recommender systems. In *Proc. of Int. conf. on Intelligent user interfaces* (IUI '02). ACM, New York, NY, USA, 127-134.

[28] 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.

[29] Swearingen, K. and Sinha, R., 2000. Interaction design for recommender systems. *Designing Interactive Systems.* London, 25-28.