HAL
open science

# User-Centric vs. System-Centric Evaluation of Recommender Systems

Paolo Cremonesi, Franca Garzotto, Roberto Turrin

## To cite this version:

Paolo Cremonesi, Franca Garzotto, Roberto Turrin. User-Centric vs. System-Centric Evaluation of Recommender Systems. 14th International Conference on Human-Computer Interaction (INTERACT), Sep 2013, Cape Town, South Africa. pp.334-351, 10.1007/978-3-642-40477-1_21 . hal-01504894

**HAL Id: hal-01504894**

**https://inria.hal.science/hal-01504894v1**

Submitted on 10 Apr 2017

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

# User-centric vs. System-centric Evaluation of Recommender Systems

Paolo Cremonesi[1], Franca Garzotto[1], Roberto Turrin[2]

[1] Politecnico di Milano, Milano, Italy
[2] ContentWise, Milano, Italy

[paolo.cremonesi, franca.garzotto]@polimi.it
roberto.turrin@contentwise.tv

**Abstract.** Recommender Systems (RSs) aim at helping users search large amounts of contents and identify more effectively the items (products or services) that are likely to be more useful or attractive. The quality of a RS can be defined from two perspectives: *system-centric*, in which quality measures (e.g., precision, recall) are evaluated using vast datasets of preferences and opinions on items *previously* collected from users that are *not* interacting with the RS under study; *user-centric*, in which user measures are collected from users interacting with the RS under study. Prior research in e-commerce has provided some empirical evidence that system-centric and user-centric quality methods may lead to inconsistent results, e.g., RSs that were "best" according to system-centric measures were not the top ones according to user-centric measures. The paper investigates if a similar mismatch also exists in the domain of *e-tourism*. We discuss two studies that have adopted a system-centric approach using data from *210000 users*, and a user-centric approach involving 240 users interacting with an *online hotel booking* service. In both studies, we considered four RSs that employ an implicit user preference elicitation technique and different baseline and state-of-the-art recommendation algorithms. In these four experimental conditions, we compared system-centric quality measures against user-centric evaluation results. System-centric quality measures were *consistent* with user-centric measures, in contrast with past studies in e-commerce. This pinpoints that the relationship between the two kinds of metrics may depend on the business sector, is more complex that we may expect, and is a challenging issues that deserves further research.

**Keywords:** Recommender systems, E-tourism, Evaluation, Decision Making.

## 1 Introduction

Recommender Systems (RSs) aim at helping users search large amounts of digital contents and identify more effectively the items that are likely to be more useful or attractive. For consumers overwhelmed by excessively wide offer of products or services, recommendations reduce information overload, facilitate the discovery of what

adfa, p. 1, 2011.
© Springer-Verlag Berlin Heidelberg 2011

they need or are interested to, help them to make choices among a vast set of alternatives, and potentially improve their decision making process. From a provider's perspective, RSs are regarded as a means to improve users' satisfaction and ultimately increase business.

Most recommender systems operate by predicting the opinion (i.e., the numerical *rating*) that a user would give to an item (such as a movie, or a hotel), using a statistical model built from the characteristics of the item (content-based approaches) or the opinions of a community of users (collaborative-based approaches).

Some research has explored the effectiveness of RSs as decision support tools in the e-tourism domain, and has investigated how they influence users' decision making processes and outcomes [7,24,31,33,13,28]. Empirical evidence suggests that RSs improve user's decision making and their influence depends on a variety of factors which are related to the quality of the recommender system.

RS quality can be defined either in terms of system-oriented metrics, which are evaluated algorithmically (e.g., precision, recall), or with user-centric experiments. [8,12,28].

- In *user-centric* evaluation, users interact with a running recommender system and receive recommendations. Measures are collected by asking the user (e.g., through interviews or surveys), observing her behavior during use, or automatically recording interactions and then subjecting system logs to various analyses (e.g., click through, conversion rate).
- With *system-centric* methods, the recommender system is evaluated against a *pre-built* ground truth dataset of opinions. Users do not interact with the system under test but the evaluation, in terms of *accuracy*, is based on the comparison between the opinion of users on items as estimated by the recommender system and the judgments *previously* collected from real users on the same items.

Although the user-centric approach is the only one able to truly measure the user's satisfaction on recommendations and the quality of the decision making process, conducting empirical tests involving real users is difficult, expensive, and resource demanding. On the contrary, system-centric evaluation has the advantage to be immediate, economical and easy to perform on several domains and with multiple algorithms.

Recently, many researchers have argued that the system-centric evaluation of RSs in e-commerce applications does not always correlate with how the users perceive the value of recommendations [2,5,6,19,22,27]. This may happen because system-centric evaluation cannot reliably measure non-accuracy metrics such as novelty – the extension to which recommendations are perceived as new – which more reflects the user and business dimensions. These works suggest that RS effectiveness in e-commerce applications should not be evaluated simply in terms of system-oriented accuracy but user-centric metrics should be adopted as well.

These contrasting results between system-centric and user-centric evaluation of RSs do not necessarily hold for e-tourism applications, because of the peculiar nature of the touristic product [11,25,26,30]:

- Touristic products lack the feature of "try-before-buy" or "return in case the quality is below expectance". Online tourist service purchasing involves a certain amount of risk taking.
- A priori comprehensive assessment of the quality of the touristic product is impossible: tourists must leave their daily environment to use it.
- The touristic product has to do with an overall emotional experience.
- In many circumstances, novelty is a weak quality attribute of touristic products. Tourists can "reuse" and buy the same product again and again if they consider the experience emotionally satisfying.

Because of these differences that might impact on users' decision making, the online selling of touristic services cannot be considered as a special case of e-commerce, and the quality characteristic of this process might differ significantly in the two domains.

This paper explores the influence of recommendations on decision making in the wide application arena of online tourism services, specifically considering hotel booking. Our research is grounded on a specific case study – the online reservations service provided by Venere.com, a subsidiary company of the Expedia group, one of the worldwide leaders in the hotel booking market, featuring more than 120,000 hotels, bed and breakfasts and vacation rentals in 30,000 destinations worldwide. Our joint work with Expedia addresses the following research question:

> *Do the algorithms which perform best in terms of system-centric quality generate recommendations that provide the best effects on decision making?*

We focus our research on the effects of recommendations on decision making in relationship to a specific design factor – the recommendation *algorithm* used. We aim at exploring the *differences* between users who use an online booking system *without recommendations* and those who use the same booking system *extended with personalized recommendations* generated by *different algorithms*.

Our research investigates the effects of recommender algorithms from both a user-centric ("subjective") point of view and a system-centric ("objective") perspective. To explore our general research question and to evaluate if system-oriented metrics are able to correctly capture the quality of the decision making process from a user perspective, we carried on two wide and articulated empirical studies:

1. a *system-centric evaluation* to measure the objective quality in terms of *accuracy* (recall and fallout); this involved 210,000 simulated users, characterized by absence or presence of personalized recommendations, the latter being generated by three different algorithms (collaborative, content-centric, and hybrid);
2. a set of *user-centric experiments* involving 240 users and measuring different decision making attributes in *four experimental conditions*, characterized by the same four recommenders adopted in system centric evaluation.

The comparison of the evaluation outcomes shows that system-centric and user-centric metrics lead to consistent results, in contrast with past studies in e-commerce [5,6], and suggests that in the online hotel booking domain system-centric accuracy

measures are good predictors of the beneficial effect of personalized recommendations on user's decision making. Our findings pinpoints that the relationship between the system-centric and user-centric metrics may depend on the business sector, is more complex that we may expect, and is a challenging issues that deserves further research.

## 2 Related Work

### 2.1 Recommender Systems in e-Tourism

The potential benefits of RSs in e-tourism have motivated some domain-specific researches. Ricci et al. in [24] present NutKing, an online system that helps the user to construct a travel plan by recommending attractive travel products or by proposing complete itineraries. The system collects information about personal and travel characteristics and provides hybrid recommendations. NutKing searches for user-centric similar items and later ranks them based on a content-centric similarity between items and user's requirements. Levi et al. in [18] describes a recommender system for online hotel booking. The system adopts a recommendation technique symmetric to technique described in [24] and adopts sentiment-analysis to estimate user's rating from their reviews. Zanker et al. [33] present an interactive travel assistant, designed for an Austrian spa-resort, where preference and requirement elicitation is explicitly performed using a sequence of question/answer forms. Delgado et al. in [7] describe the application of a collaborative attribute-centric recommender system to the Ski-Europe.com web site, specialized in winter ski vacations. Recommendations are produce by taking into account both implicit and explicit user feedbacks. Implicit feedback is inferred whenever a user prints, bookmarks, or purchases an item (positive feedback) or does nothing after viewing an item (negative feedback).

### 2.2 Evaluation of recommender systems

Several studies have investigated how to measure the effectiveness of recommenders. A systematic review of *system-centric* evaluation techniques is reported by Herlocker et al. in [12]. More recently, some researchers [3,19,20,22,27] have argued that RS effectiveness should not be evaluated simply in terms of system-centric metrics and have investigated *user-centric* evaluation methods, which focus on the human/computer interaction process (or User eXperience, UX) [18,22,29].

Swearingen and Sinha [27] were among the first studies to point out that subjective quality of a RS depends on factors that go beyond the quality of the algorithm itself. Without diminishing the importance of the recommendation algorithm, these authors claim that RS effectiveness should not be evaluated simply in terms of system-centric accuracy metrics. Other design aspects, ignored by these metrics, should be measured, and in particular those related to the acceptance of the recommender system and of its recommendations.

Along the same vein, other researchers have investigated the so called user-centric methods, which focus on how user characteristics are elicited and recommended items are presented, compared, or explained. They explore "subjective" quality of RSs and attempt to correlate it to different UX factors. They highlight that, from a user's perspective, an effective recommender system should inspire credibility and trust towards the system [22] and it should point users towards new, not-yet-experienced items [18].

Due to the intrinsic difficulty of performing user studies in the RS domain, empirical results in this field are tentative and preliminary. Celma and Herrera [4] report an experiment that studied how users judged novel recommendations provided by a CF and a CBF algorithm in the music recommendation context. Ziegler et al. [34] and Zhang et al. [32] propose diversity as a quality attribute: recommender algorithms should seek to provide optimal coverage of the entire range of user's interests. This work is an example of a combined use of automatic and user-centric quality assessment techniques. Pu et al. [22] developed a framework called ResQue, which defines a wide set of user-centric quality metrics to evaluate the perceived qualities of RSs and to predict users' behavioral intentions as a result of these qualities.

**Table 1.** Experimental conditions used in the two studies

| Study | Type | Independent variables (algorithms) | Dependent variables | | | Users |
|---|---|---|---|---|---|---|
| 1 | System-centric simulation | HotelAvg<br>PureSVD<br>DirectContent<br>Interleave | Accuracy (recall and fallout) | | | 210,000 (simulated) |
| 2 | User-centric experiment | | Choice satisfaction | Satisfaction | subjective | 240 (total) |
| | | | Choice risk | Trust | | |
| | | | Perceived time | Effort | objective | |
| | | | Elapsed time | | | |
| | | | Extent of hotel search | | | |
| | | | Menu interactions | Efficacy | | |

## 3 The Design of the Studies

The research question presented in the Introduction has been explored with *two studies* – a system-centric simulation and a user-centric experiment – summarized in Table 1. In both the studies, the effects of recommendations have been explored under 4 different experimental conditions defined by one manipulated variable: the *recommendation algorithm.* Our study considers *one* non-personalized algorithm and *three* personalized RSs representatives of three different classes of algorithms: *collaborative*, *content* and *hybrid*.

- **HotelAvg** is a non-personalized algorithm and presents hotels in decreasing order of *average user rating* [15]. This is the default ranking option adopted in our study when the user does not receive personalized recommendations. The same ranking

strategy is adopted by most online hotel booking systems such as TripAdvisor, Expedia, and Venere.

- **PureSVD** is a collaborative algorithm based on matrix-factorization; previous research shows that its accuracy is one of the best in the movie domain [6].
- **DirectContent** recommends hotels whose content is similar to the content of hotels the user has rated [18]. Content analysis takes into account the 481 features (e.g., category, price-range, facilities), the free text of the hotel description, and the free text of the hotel reviews. *DirectContent* is a simplified version of the LSA algorithm described in [1].
- **Interleave** is a hybrid algorithm that generates a list of recommended hotels alternating the results from *PureSVD* and *DirectContent*. Interleave has been proposed in [3] with the name "mixed hybridization" and, although trivial in its formulation, has been shown to improve diversity of recommendations.

### 3.1 Study 1: System-centric evaluation

The first study analyzes the accuracy of recommendations as a function of the recommender algorithm. For the evaluation, Venere.com made us available a catalog of more than 3,000 hotels and 72,000 related users' reviews. Each accommodation is provided with a set of 481 features concerning, among the others: accommodation type (e.g., residence, hotel, hostel, B&B) and service level (number of stars), location (country, region, city, and city area), booking methods, average single-room price, amenities (e.g., spa), and added values (e.g., in-room dining). User's reviews associated to each accommodation consist of numeric ratings and free-text.

We have enriched the dataset with additional reviews extracted from the TripAdvisor.com web site using a web crawling tool. Table 2 reports the detailed statistics of the dataset used in our experiments.[1]

**Table 2.** Dataset used in the two studies.

| | Total (Venere+TripAdvisor) | Venere | TripAdvisor (crawled) |
|---|---|---|---|
| **Hotels** | 3,100 | 3,100 | – |
| **Users (reviewers)** | 210,000 | 72,000 | 138,000 |
| **Reviews and ratings** | 246,000 | 81,000 | 165,000 |
| **Hotel features** | 481 | 481 | – |

**Dependent Variables.**

System-centric quality can be measured by using either accuracy metrics (e.g., precision, recall and fallout) or error metrics (e.g., RMSE and MAE). The hybrid algorithm tested in this study cannot be evaluated with error metrics since it does not compute actual ratings [15]. Hence we have considered only accuracy metrics. In particular we focused our attention on *recall* (the conditional probability of suggesting

[1] The dataset is available by contacting the authors.

a hotel given it is relevant for the user) and *fallout* (the conditional probability of suggesting a hotel given it is irrelevant for the user) as the *dependent* variables. A good algorithm should have large recall (i.e., it should be able to recommend hotels of interest to the user) and low fall-out (i.e., it should avoid to recommend hotels of no interest to the user).

The methodology adopted to measure the two variables is the same used in many works on recommender system – e.g., [6,16]. Ratings in the dataset were randomly split into two subsets: *training* set (90% of the ratings) and *test* set (10% of the ratings). In order to measure recall, we first trained the algorithms using the ratings in the training set. Then, for each user in the test set and for each hotel in the test set that was rated 10-stars or 9-stars by the user (we assumed these hotels to be relevant for the users) we followed these steps:

- We randomly selected 1,000 additional hotels 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 relevant hotel and for the additional 1,000 hotels.
- We formed a recommendation list by picking the *N* hotels with the largest predicted ratings (top-N recommendation).

For each recommendation we have a *hit* (e.g., a successful recommendation) if the relevant hotel is in the list. 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

$$\text{recall}(N) = \frac{\text{number of times the relevant hotel is in the list}}{\text{number of recommendations}}$$

A similar approach was used to measure fallout, with the only difference being that we selected non-relevant hotels – defined as the hotels rated lower than 2 out of 10 stars. The fallout was computed as

$$\text{fallout}(N) = \frac{\text{number of times the non–relevant hotel is in the list}}{\text{number of recommendations}}$$

Recall and fallout range from 0% to 100%. An ideal algorithm should be able to recommend all of the interesting hotels (i.e., recall equals to 100%) and to discard all of the uninteresting hotels (i.e., fallout equals to 0%).

### 3.2 Study 2: User-centric evaluation

The second study analyzes opinions and behavior of 240 users interacting with an online hotel booking service.

**Dependent Variables.**

We model the *effects* on decision making that can be associated to the introduction of personalized recommendations using some subjective attributes resulting from the user's perception and judgment of the decision activity (subjective variables) as well

as some objectively measurable attributes of the decision processes (objective variables). The variables, listed in Table 1, have been defined according to the model described in [30]:

- **Choice satisfaction:** the subjective evaluation of the reserved hotel in terms of quality/value for the user;
- **Choice risk:** the user's perception of uncertainty and potentially adverse consequences of booking the chosen hotel, measured in terms of the perceived degree of mismatching between the characteristics of the chosen hotel emerging from the use of the system and the real characteristic of the accommodation;
- **Perceived time:** the user's judgment on the length of the decision making process;
- **Elapsed time:** the time taken for the user to search for hotel information and make a reservation decision;
- **Extent of hotel search:** the number of hotels that have been searched, for which detailed information has been acquired;
- **Number of sorting changes:** the number of times the user *changed the ordering* of hotels in the list view: the *default* ordering is (i) descending order of average user ratings, when there are no recommendations, and (ii) descending order of estimated user relevance, when there are recommendations. Ordering change is a measure of the *efficacy* of RSs in situations where the conversion rate, i.e., the percentage of recommended items that are actually purchased by users [33] cannot be assessed. This happens, for example, when a system does not present a *separate* list of recommended items, but recommendations are rendered by sorting items in descending order of relevance as estimated by the recommender algorithm, which represents the "de facto" recommendation list.

A *questionnaire* has been used to measure *choice satisfaction*, *choice risk* and *perceive time*. Measures of *effort* and *efficacy* (*execution time, extent of hotel search, number of sorting changes)* have been obtained by analyzing the interaction data collected from the system logs.

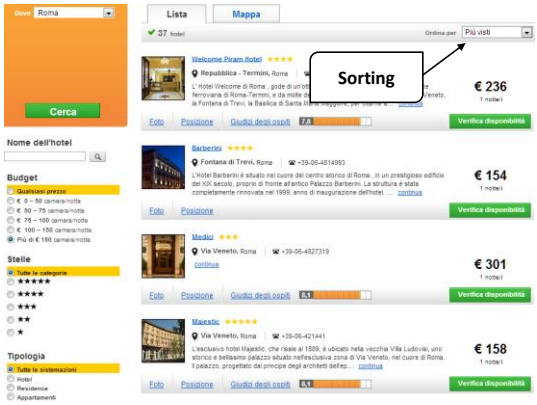


**Fig. 1.** Screenshot of the PoliVenus page showing a list of hotels

**Instruments.**

For the purpose of our study, we have developed PoliVenus, a web-centric testing framework for the hotel booking field, which can be easily configured to facilitate the execution of controlled empirical studies in e-tourism services (Figure 1). PoliVenus implements the same layout as Venere.com online portal and simulates *all* its functionality with the exception of payment functions. PoliVenus contains the same catalog of hotels and reviews used in the first study. The Polivenus framework is based on a modular architecture and can be easily customized to different datasets and types of recommendation algorithms. Its current library of algorithms comprises 20 algorithms, developed in cooperation with ContentWise[2]. PoliVenus can operate in two different configurations:

- **Baseline configuration**: it corresponds to the existing Venere.com portal. Users can filter the hotel catalogue according to hotel characteristics (e.g., budget range, stars, accommodation type, city area), and retrieve non-personalized results. They can sort the list by popularity, service level, price, or user average rating (the default sorting option). The default sorting option corresponds to the HotelAvg non-personalized recommender algorithm.
- **Personalized configuration**: it is almost identical to the baseline configuration, the only difference being the personalized recommendations. When watching a list of hotels previously filtered according to accommodation characteristics, the user is offered an additional option to sort hotels on the basis of the personalized recommendations (this is the default sorting option).

The user profile required by the algorithms to generate recommendations is based on the user's current interaction with the system (*implicit elicitation*). This choice is motivated by three specific reasons:

(i) we want to support users who have no rating history or who are not interested in logging into the system;

(ii) we are interested in exploring a smooth integration of personalized recommendations in existing online booking systems; to enable explicit elicitation would require the introduction of an intrusive add-on;

(iii) according to a large number of works, the lower effort of implicit elicitation (as compared to explicit elicitation) is related to higher perceived effectiveness of recommendations [9,10,14,23].

The implicit elicitation mechanism adopted in PoliVenus is the following: whenever a user interacts with an object on the interface, the system assigns a score to the hotel related to that object (e.g. link, button, map, picure, etc.). With all these *signals*, PoliVenus builds the user profile for the current user session: the user profile contains implicit hotel ratings, where each rating is computed as the maximum of all the signals generated for that hotel. The user profile is continuously updated with every new signal and the list of recommended hotels is updated accordingly.

Following the definition of Ricci and Del Missier in [24] we considered short-term and long-term signals. In our implementation the short-term signal is the last interac-

[2] www.contentwise.tv

tion of the user with the system. All previous interactions with the system are considered long-term signals. In order to give more importance to the most recent interaction, we have employed an exponential decay function to the ratings. Whenever a new signal enters in the user profile, all previous ratings are divided by a dumping factor. The magnitude of the damping factor controls the decay rate. In our experiments we have used a damping factor of 2.

**Participants and procedure.**

Our main research audience is represented by users aged between 20 and 40 who have some familiarity with the use of the web and had never used Venere.com before the study (to control for the potentially confounding factor of biases or misconceptions derived from previous uses of the system). The total number of recruited subjects who completed the task and filled the questionnaire by the deadline was *240*. They were equally distributed in the four experimental conditions. We recruited participants from current students and ex-alumni from the School of Engineering and the School of Industrial Design of our university. They were contacted by e-mail, using university mailing lists. The invitation included the description of the activities to be performed and the reward for taking part in the study. Users were not aware of the true goal of the experiment.

To encourage participation, and to induce participants to play for real and to take decisions as they would actually do when planning a vacation, we used a lottery incentive [21]. Participants had the chance of winning a *prize*, offered to a randomly selected person who completed the assigned decision making task and filled the final questionnaire by a given deadline. The prize consisted of a coupon of the value of 150€ to be used to stay in the hotel fictitiously reserved using PoliVenus.

All participants were given the following instructions: "*Imagine that you are planning a vacation in Rome and are looking for an accommodation during Christmas season; choose a hotel and make a reservation; dates and accommodation characteristics (stars, room type, services, and location) are at your discretion. After confirming the reservation (simulated), please complete the final questionnaire*" .

**Table 3.** Excerpt from the questionnaire

| Question | Range | Dependent Variable |
|---|---|---|
| How much are you satisfied with your final choice? | 1:not too much<br>5:very much | Choice Satisfaction |
| The time required to book the hotel is: | 1:short<br>5:overmuch | Perceived Time |
| How much will the characteristics of the reserved hotel correspond to those of the real accommodation? | 1:not much,<br>5:very much | Choice Risk |

After accessing PoliVenus, reading the instructions, and agreeing on the study conditions (lottery participation and privacy rules), each participant was automatically moved to the homepage of the PoliVenus hotel reservation system and randomly as-

signed to one of the four experimental conditions. After committing the reservation, the user was directed to the questionnaire page containing 11 questions, a subset of which is reported in Table 3.

# 4 Results

## 4.1 Study 1: System-centric evaluation

Figure 2 presents a plot of recall versus fallout, where each point on the curve corresponds to a value of the number *N* of recommended hotels. A perfect recommendation algorithm will generate a curve that goes straight to the upper left corner of the figure, until 100% of relevant hotels and 0% of non-relevant hotels have been recommended.

The figure shows that the hybrid *Interleaved* algorithm is the most accurate. The accuracy of the content and collaborative algorithms (*DirectContent* and *pureSVD*) is not significantly better than the accuracy of the non-personalized baseline *HotelAvg*.

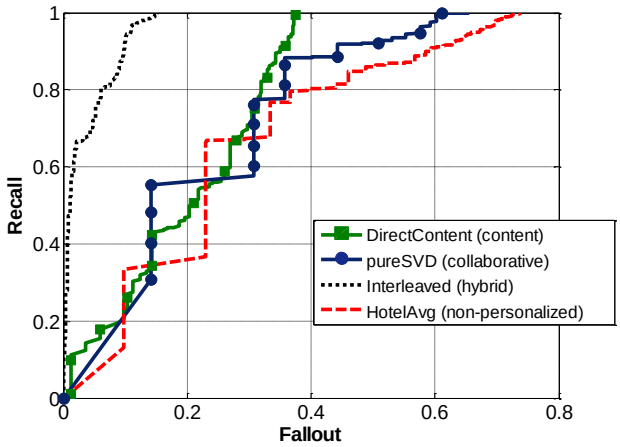


**Fig. 2.** System-centric evaluation: Recall vs.Fallout

## 4.2 Study 2: User-centric evaluation

We first polished the collected data by removing the ones referring to subjects who showed apparent evidences of gaming with the testing system (e.g., those who interacted with the system for less than 2 minutes) or left too many questions unanswered. In the end, we considered the data referring to *229* participants. They were almost equally distributed in the four experimental conditions, each one involving a number of subjects between 54 and 58.

**Subjective metrics.**

We used ANOVA to test the effects of the algorithm on the dependent variables. The tests suggest that the algorithm has a significant impact ($p<0.05$) on all of the subjective variables (Choice Satisfaction, Perceived Risk, Perceived Time). We ran multiple pair-wise comparison post-hoc tests using Tukey's method on subjective variables. The results are shown in Figures 3–5, where the mean is represented by a circle and the 95% confidence interval as a line.

The results show that the adoption of the hybrid *Interleaved* algorithm significantly increases user satisfaction and decreases both perceived risk and effort with respect to the non-personalized *HotelAvg* configuration. On the contrary, content and collaborative algorithms (*DirectContent* and *pureSVD*) do not affect user satisfaction and perceived effort with respect to the baseline scenario, although *DirectContent* reduces the perceived risk.

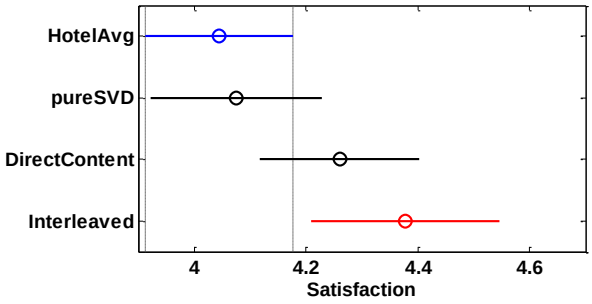


**Fig. 3.** Choice Satisfaction: how much are you satisfied with your final choice? (1: not too much – 5: very much)

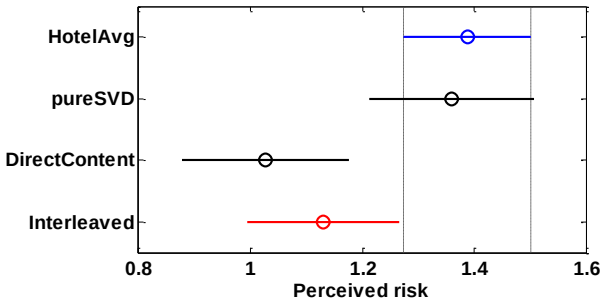


**Fig. 4. Choice Risk**: how much the characteristics of the reserved hotel will correspond to those of the real accommodation? (1: not much – 5: very much)

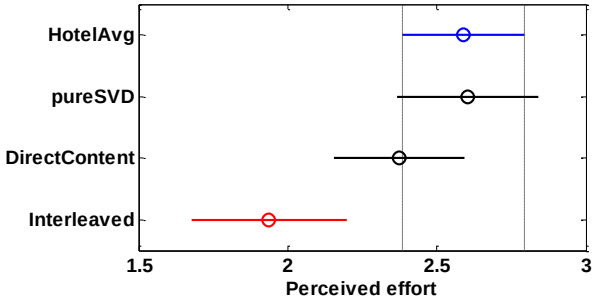


**Fig. 5. Perceived Effort**: the time required to book the hotel is … (1: short – 5: overmuch)

**Objective metrics.**

Behavioral data emerging from the tracking of users' interaction provides some interesting results. PoliVenus does not us enable to directly measure the efficacy of the RS using conversion rate metrics, because it does not explicitly present a separate list of recommended items. The hotel list view, where hotels are sorted in descending order of relevance as estimated by the recommender algorithm, is the actual recommendation list. As mentioned in the previous section, we have estimated the *efficacy* of the RS by measuring how many users changed the default sorting options (recommended hotels first) by setting other parameters (e.g., price, popularity, stars). The results show that only *37%* of the users with personalized recommendations changed the sorting of hotels compared to *54%* of the users in the baseline scenario (e.g., with non-personalized recommendations).

Surprisingly, the value of the two objective effort variables (*elapsed time* and *number of explored hotels*) is not affected by the recommender systems. Regardless the presence of personalized recommendations, the *number of explored hotels* is between 5 and 8 for almost all of the users. In other words, customers wish to compare in details at least 5 and no more than 8 alternative choices before committing to a final decision. Moreover, the time required to complete the task is between 10 and 22 minutes for 95% of the users and is not affected by the presence of personalized recommendations.

## 5 Discussion and conclusions

This section discusses the validity of our empirical study, analyses the findings in relationship to our research question, and provides some possible interpretation of results, also in light of the current state of the art.

### 5.1 Validity of our study

The internal validity of our study is supported by the accuracy of our research design and by the quality of study execution. We have carefully implemented a mechanism to randomly assign participants to the different experimental conditions. We have adopted a lottery incentive to improve the accuracy of the task's execution and offered a shared motivation to all participants. Obviously, the individuals' intrinsic characteristics and actual behavior always bring to an experiment a myriad of factors that can be hardly controlled. In terms of external validity, the results of our study are limited to those participants and conditions used in our study. The applicability of our results might not be confined to the specific online booking system used. Most services available in the market provide a user experience very similar to Venere.com, in terms of filtering criteria and information/navigation structures, and it is likely that replications of our study on other systems may lead to results consistent with our findings. Finally, the high overall number of testers (240) and the relatively high number of subjects involved in each experimental condition (between 54 and 58) may allow us to generalize our results to a wider population of users aged 20-40.

In the "with-RS" experimental conditions, there might be a potentially confounding factor which might affect the results. The list of presented hotels for a user in the "with-RS" condition changes after each interaction, as the user profile is updated with each new "signal". On the contrary, the user in the "without-RS" condition sees no updates to the list of presented hotels, unless she explicitly changes the search criteria. This phenomenon could explain why users not receiving personalized recommendations take more initiative in changing the sorting of hotels in the list. However, as the different "with-RS" conditions provided exactly the same results, we may argue that the effect of recommender systems on the users' behavior is genuine and not affected by the different "dynamics" of the list of hotels "with" and "without recommendations". To further validate this assumption, in future tests it may be worth considering changing the list of hotels in the menu entry of presented hotels after each signal also for the no-RS subjects (in some random re-sort).

### 5.2 Research question

Overall, our findings *answer our research question*: system-centric evaluation can be used to compare the effects of different recommender algorithms on the decision-making process in online hotel booking. A finer grained analysis of the statistically relevant relationships among all the different variables offers a much more articulated picture of the results, which are apparently in contrast with previous findings and suggest a number of interesting considerations for e-tourism applications:

- system-centric accuracy metrics (e.g., recall and fallout) are a good approximation of the quality perceived by the use;
- personalized recommendations do not reduce the decision-making effort; moreover, there is a mismatch between objective and perceived effort.

**System-centric vs. user-centric evaluation.**

The comparison between Figures 2 and 3 shows a strong consistency between system-centric accuracy (e.g., recall and fallout) and the users' perceived quality. We can claim that, in the e-tourism domain, system-centric quality attributes are good predictors of how the users perceive the quality of a recommender algorithm. This result is apparently in contrast with previous works [6,19,22,27] which state that user satisfaction is not correlated with accuracy of algorithms.

A possible interpretation of this result is to consider that, in traditional e-commerce applications, non-accuracy attributes such as novelty have an important role in shaping the user satisfaction. This is not necessarily the case with touristic services as they have a number of peculiar aspects that call for a rethinking of the decision making process normally assumed in traditional e-commerce and affect which RSs attributes impact on the user final choices [11,25,26,30]:

- The touristic product is complex and emotional. Booking hotel rooms requires a considerable decision effort, as the user might lack background knowledge of the characteristic of the possible accommodations. A priori comprehensive assessment of hotel accommodations is impossible. Touristic products lack the feature of "try-before-buy" or "return-if-not-satisfied". In case the quality of the accommodation is below expectance, the whole travel experience will be negatively affected and the memory of it will be persistent over time. The opportunity to fill that particular period of the user life with a positive experience will be lost.
- The touristic product is volatile. Hotel rooms are limited in number and, especially in high-seasons, the best-value accommodations are sold-out months in advance. Even optimizers (i.e., users who would choose the "best" possible hotel) could be urged to take sub-optimal decisions.

These aspects imply that (i) online hotel booking involves a certain amount of risk taking, and (ii) the user needs to minimize such risk by searching for an accommodation perfectly matching his/her largely unexpressed requirements. Accuracy, e.g., the perfect matching between user needs and solutions proposed, may become the most important attribute of recommendations as it reduces the perceived risk.

**Decision effort.**

There is *no significant variation among personalized algorithms with* respect *to objective effort*: none of them statistically differ from the baseline in terms of execution time and extent of product search, i.e., number of explored hotel pages. Our results show a *mismatch between satisfaction* and *effort*: users exposed to hybrid and content recommendations perceived the decision activity process as more satisfying than those without personalized recommendations, although they spent the same time in the process. Moreover, there is a *mismatch between objective and perceived effort*: users exposed to hybrid recommendations perceived the decision activity process as shorter (Figure 5), and more trustworthy (Figure 4), than the others, although they spent the same time and explored the same number of hotels.

This result is partially in line with some previous studies which explored subjective vs. objective effort under different conditions of RS design [24]. Works on preferences elicitation pinpoint for example that more effort-intensive sign up activities do not necessarily imply higher perceived effort [27], as if the effort perception were mitigated by the benefits of a more satisfying decision process. Other works hypothesize that, thanks to RSs, users spend less time in searching for items and more time in the more satisfactory activity of exploring information related to the choice processes [24].

### 5.3 Conclusions

In spite of some limitations, our work represents a contribution to the research and practice in RS design and evaluation, for the specific domain of online booking and from a more general perspective. Our research differs from previous work in this domain for a number of aspects:

- We smoothly *extend* a *true* online booking system (Venere.com) with personalized recommendations without creating major modifications to the "standard" interaction flow and overall user experience. Our user – both in the control group and in the treatment groups – is always in control of all information and functionality of conventional online booking services. In contrast, most previous works either are based on prototype RSs or create ad hoc user experiences for the subjects exposed to recommendations.
- We compare *three different personalized algorithms* against the baseline scenario (non-personalized) and against each other. Previous works limit their analysis to a single recommendation algorithm evaluated against a non personalized baseline. In addition, we consider recommenders involving *implicit* elicitation, while most of existing studies in the e-tourism domain address recommenders with (more intrusive) explicit elicitation
- Our results on the comparison between system-centric and user-centric evaluation are totally new for the e-tourism domain. They show that the peculiarity of the touristic products may exploit different aspects of recommender systems, and suggest the possibility to adopt system-centric evaluation techniques as a good approximation of the user experience.

In relationship to other studies in e-tourism and in other domains, the *research design* of our empirical study is per se a strength of our work, for a number of reasons: the large number of variables that have been measured, the sophisticated technological instrument used (the PoliVenus framework), the vast size of the involved subjects (240), and the lottery based incentive mechanisms adopted to motivate users and commit them to realistic and sound task execution.

Overall, our findings extend our understanding of the potential of introducing recommendations to improve decision making processes. At the same time, our results help us to identify potential weakness of current state of the art approaches in RSs and can orient future investigations in the field.

## 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. Bollen, D., Knijnenburg, B.P., Willemsen, M.C., and Graus, M., Understanding choice overload in recommender systems. In *Proc. of the fourth ACM conference on Recommender systems* (RecSys '10). ACM, New York, NY, USA, 63-70
3. Burke, R. 2007. Hybrid web recommender systems. In *The adaptive web*, Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl (Eds.). Lecture Notes In Computer Science, Vol. 4321. Springer-Verlag, Berlin, Heidelberg 377-408.
4. Celma, O. and Herrera, P., A new approach to evaluating novel recommendations. In *Proceedings of the 2008 ACM conference on Recommender systems (RecSys '08)*. ACM, New York, NY, USA, 179-186.
5. Cremonesi, P., Garzotto, F. Negro, S. Papadopoulos, A. Turrin, R., Looking for "good" recommendations: a comparative evaluation of recommender systems. In *Proc. of the 13th IFIP TC 13 international conference on Human-computer interaction - Volume Part III (INTERACT'11)*, Springer-Verlag, Berlin, Heidelberg, 152-168.
6. 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, 41 pages.
7. Delgado, J. and Davidson, R. Knowledge bases and user profiling in travel and hospitality recommender systems, 2002.
8. Deshpande, M. and Karypis, G..Item-centric top-N recommendation algorithms. *ACM Trans. Inf. Syst.* 22, 1 (January 2004), 143-177
9. 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.
10. 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.
11. Hennig-Thurau T., Gwinner K.P., Walsh G., Gremler D.D. (2004) Electronic Word-of Mouth via Consumer-Opinion Platforms: What Motivates Consumers to Articulate Themselves on the Internet? *Journal of Interactive Marketing* 18 (1), 38-52.
12. Herlocker, J.L., Konstan, J.A., Terveen, L.G., and Riedl, J.T. 2004. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.* 22, 1 (January 2004), 5-53
13. Jannach, D., Zanker, M., Jessenitschnig, M., and Seidler, O.. Developing a conversational travel advisor with advisor suite. In *Marianna Sigala, Luisa Mich, and Jamie Murphy, editors, Information and Communication Technologies in Tourism*, pages 43–52. Springer Vienna, 2007.
14. 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.
15. Konstan, J.A., Riedl, J., 2012. Recommender systems: from algorithms to user experience. *User Model. User-Adapt. Interact*. 22, 101-123.
16. Koren, Y. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '08). ACM, New York, NY, USA, 426-434.
17. Levi, A., Mokryn, O., Diot, C. and Taft, N.. Finding a needle in a haystack of reviews: cold start context-centric hotel recommender system. In *Proceedings of the sixth ACM*

*conference on Recommender systems*, RecSys ’12, pages 115–122, New York, NY, USA, 2012. ACM.

18. Lops, P., De Gemmis, M., and Semeraro, G., 2011. Content-centric 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.
19. Mahmood, T., Ricci, F., Venturini, A., and Höpken, W.. Adaptive recommender systems for travel planning. In *Proceedings of the International Conference on Information Technology and Travel & Tourism*, ENTER, pages 1–11. Springer, 2008.
20. 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.
21. Porter, S. R., and Whitcomb, M. E. “The Impact of Lottery Incentives on Survey Response Rates.” *Research in Higher Education*, 2003, *44*(4), 389–407.
22. 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.
23. 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.
24. Ricci, F. and Missier, F.. Supporting travel decision making through personalized recommendation. In *Clare-Marie Karat, Jan Blom, and John Karat, editors, Designing Personalized User Experiences in eCommerce*, volume 5 of Human–Computer Interaction Series, pages 231–251. Springer Netherlands, 2004.
25. Ricci, F. and Wietsma, R. Product Reviews in Travel Decision Making, *Proc. of the International Conference on Information and Communication Technologies in Tourism* 2006 Lausanne, Switzerland, 2006. Springer Verlag, 296-307.
26. Senecal, S., and Nantel, J. The Influence of Online Product Recommendations on Consumers' Online Choices, *Journal of Retailing* (80:2), 2004, pp. 159-169.
27. Swearingen, K. and Sinha, R. 2001. Beyond algorithms: An hci perspective on recommender systems. In *Proceedings of the ACM SIGIR Workshop on Recommender Systems*
28. Takács, G., Pilászy, I., Németh, B., and Tikk, D. Scalable Collaborative Filtering Approaches for Large Recommender Systems. *J. Mach. Learn. Res*. 10 (June 2009), 623-656.
29. Tintarev, N. 2007. Explanations of recommendations. In Proceedings of the 2007 ACM conference on Recommender systems. RecSys ’07. ACM, New York, NY, USA, 203–206.
30. Werthner, H. and Ricci, F., E-commerce and tourism. Commun. ACM 47, 12 (December 2004), 101-105
31. Xie, M., Lakshmanan, L.V.S., and Wood, P.T.. Comprec-trip: A composite recommendation system for travel planning. In *Data Engineering*, ICDE, 2011 IEEE 27th International Conference on, pages 1352 –1355, april 2011.
32. Zhang, Y., Callan, J., and Minka, T.. 2002. Novelty and redundancy detection in adaptive filtering. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '02)*. ACM, New York, NY, USA, 81-88.
33. Zanker, M., Fuchs, Matthias, W., Mario Tuta, H. and Müller, N.. Evaluating recommender systems in tourism - a case study from austria. In *Proceedings of the International Conference on Information Technology and Travel & Tourism*, ENTER, pages 24–34. Springer, 2008.
34. Ziegler, C.N., McNee, S.M., Konstan, J.A., and Lausen. G., Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web (WWW '05)*. ACM, New York, NY, USA, 22-32.