IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Service-Aware Personalized Item Recommendation

NOEMI MAURO[1], ZHONGLI FILIPPO HU[2], AND LILIANA ARDISSONO.[3]

[1]Dipartimento di Informatica, Università degli Studi di Torino, Corso Svizzera 185, I-10149 Torino, Italy (e-mail: noemi.mauro@unito.it)
[2]Dipartimento di Informatica, Università degli Studi di Torino, Corso Svizzera 185, I-10149 Torino, Italy (e-mail: zhonglifilippo.hu@unito.it)
[3]Dipartimento di Informatica, Università degli Studi di Torino, Corso Svizzera 185, I-10149 Torino, Italy (e-mail: liliana.ardissono@unito.it

Corresponding author: Noemi Mauro (e-mail: noemi.mauro@unito.it).

This work has been funded by the University of Torino under grant ARDL_RILO_20_01.

ABSTRACT Current recommender systems employ item-centric properties to estimate ratings and present the results to the user. However, recent studies highlight the fact that the stages of item fruition also involve extrinsic factors, such as the interaction with the service provider before, during and after item selection. In other words, a holistic view on consumer experience, including local properties of items, as well as consumers' perceptions of item fruition, should be adopted to enhance user awareness and decision-making. In this work, we integrate recommender systems with service models to reason about the different stages of item fruition. By exploiting the Service Journey Maps to define service-based item and user profiles, we develop a novel family of recommender systems that evaluate items by taking preference management and overall consumer experience into account. Moreover, we introduce a two-level visual model to provide users with different information about recommendation results: (i) the higher level summarizes consumer experience about items and supports the identification of promising suggestions within a possibly long list of results; (ii) the lower level enables the exploration of detailed data about the local properties of items.
In a user test instantiated in the home-booking domain, we compared our models to standard recommender systems. We found that the service-based algorithms that only use item fruition experience excel in ranking and in the minimization of the error in rating estimation. Moreover, the combination of data about item fruition experience and item properties achieves slightly lower recommendation performance; however, it enhances users' perceptions of the awareness and of the decision-making support provided by the system. These results encourage the adoption of service-based models to summarize user preferences and experience in recommender systems.

INDEX TERMS Information filtering, Recommender systems, Data visualization, Service modeling

I. INTRODUCTION

In service modeling research, Stickdorn et al. [1] point out that items are complex entities whose fruition might involve stages of interaction with multiple services and actors that jointly impact customer experience. For instance, in the services related to the circular economy, such as home-booking, the offered value goes beyond item features and includes the interaction with apartments' hosts, implying different attitudes toward renting rooms or complete homes [2]. Moreover, in online retailing, the satisfaction with products depends both on their properties and on the experience with post-sales services related to customer care.

Starting from this considerations, we point out that, when personalizing the recommendation of items, their local features and the expected experience with them should be jointly analyzed in the identification of the most relevant options, as well as in their presentation to the user.

Content-based, feature-based and collaborative recommender systems [3] base their suggestions on local item properties such as the features and aspects extracted from catalogs, and on the overall ratings received by items, which represent the only utility factors steering the recommendations. Review-based recommender systems study consumer feedback to extract data about people's experience [4], [5]. However, as they do not contextualize reviews in the stages of item fruition to which consumers are exposed, these algorithms cannot aggregate information in an effective way. To comply with these limitations, we propose to enable recommender systems to reason about consumer experience in the stages of item fruition by integrating them with service

modeling techniques. As we aim at enhancing recommendation performance and user-awareness support to improve decision-making, we pose the following research questions:

- RQ1: *Does the extension of recommendation algorithms with a service-based representation of items (which explicitly models the item fruition stages) enhance recommendation quality in Top-N recommender systems, compared to only considering local item properties and overall ratings?*
- RQ2: *Does the presentation of both item properties and service-based information about their fruition enhance users' awareness about the suggested options, and confidence in the selection decisions, compared to only presenting item properties?*

To answer these questions we propose a family of service-aware recommender systems that evaluate items based on individual user preferences and on evaluation dimensions associated with item fruition stages. These dimensions abstract from the individual details that emerge from item reviews. Thus, they can be used to provide the user with a holistic summary of the experience collected by previous consumers. To specify the item fruition stages, we employ the Service Journey Maps design model [6].

We selected the home-booking domain as a test-bed for our work because it involves the user in a rich experience regarding both the home and the interaction with its host. However, our model can be applied to the suggestion of items in other domains, such as hotel booking and e-commerce in the sharing economy. In fact, in those scenarios, users can be exposed to the interaction with amateur service providers and retailers, possibly offering a low quality of service levels. Therefore, item fruition can be impacted by exogenous risk factors [7] and a summarization of customer experience can enhance the acceptance of recommendation results [8].

We compared our service-aware recommender systems to standard algorithms in a user study involving 48 participants. We tested five recommendation models and three visualization models by retrieving data about homes and reviews from the Airbnb location-based service (https://airbnb.com). The results of this study reveal that the service-based algorithms exclusively based on item fruition experience achieve the best results on ranking and minimization of error in rating estimation. Moreover, the algorithms that combine item properties with fruition experience achieve slightly lower recommendation performance. However, they enhance users' perceived awareness support and confidence in item selection. In summary, we provide two novel contributions:

1) We define different service-aware recommendation algorithms based on item features, on evaluation dimensions associated with item fruition stages, or on both.
2) We compare the performance of these algorithms to standard recommender systems in terms of utility, rating estimation and ranking capability. Moreover, we compare users' perceived quality of suggestions, their awareness about the proposed items, and their perception of the interface adequacy during the interaction with these systems.

This work is framed within the Apartment Monitoring application that helps users in finding homes from Airbnb. We extend the work described in [9] with the introduction of service-aware recommendation and with a presentation model that supports the overview of recommendation lists.

In the following, we present the related work (Section II); then we describe our dataset and data processing method (Section III). Section IV introduces the recommendation models we define. Sections V and VI describe the user study we carried out and its results, which we discuss in Section VII. Section VIII outlines limitations and future work. Sections IX and X summarize the ethical issues of our work and conclude the paper.

II. BACKGROUND AND RELATED WORK

A. SERVICE JOURNEY MAPS

The Service Journey Maps (SJMs) [6] support the design and development of products and services by focusing on the customer's viewpoint. A SJM is a visual description of user experience with a service, such as a hotel, or an online retailer, which models the stages that customers encounter during service fruition. The graphic visualization of a SJM follows a temporal line from the start point (e.g., enter website) to the end one (e.g., customer care) to describe the stages a person engages in when using the service.

Different from standard recommender systems, we employ the Service Journey Maps to describe the process underlying the fruition of the suggested items. Specifically, we use the domain model built using SJMs to steer the analysis of item reviews by clustering feedback around the specified service stages. As a result, we define a small set of evaluation dimensions that a service-aware recommender system can exploit (possibly fusing them with information about item properties) (i) to estimate item ratings and (ii) to generate a visual overview of recommendation lists, based on a holistic summary of previous consumers' experience with items.

B. RECOMMENDATION ALGORITHMS

Most recommender systems generate personalized suggestions using item-centric data that does not reflect consumer experience. Collaborative filtering evaluates items based on the ratings provided by users [10], [11]. Multi-criteria recommender systems introduce multi-dimensional ratings [14], [15]. Moreover, to ground systems' inferences on richer types of information, content-based filtering combines pure ratings with item features extracted from catalogs [16]. Some graph-based recommenders personalize the suggestions based on the chains of relations that connect users to items [18], [19], possibly by exploiting the Linked Open Data cloud [20]. Finally, hybrid recommender systems integrate different algorithms to improve their suggestions [21]–[25].

Despite the integration of different data sources for recommendation, all these systems reason about local properties of

TABLE 1. Overview of recommender systems. As no reviewed system exploits service modeling, we omitted the table column specifying this type of information.

Algorithm	Data	Item reviews	Item Presentation	Citation
Collaborative Filtering	Item ratings	No	Rating-based: bar graphs, neighbour graphs	[10], [11], [12], [13]
Multi-criteria recommenders	Multi-dimensional item ratings	No	No focus on presentation	[14], [15]
Content-based Filtering	Item features	No	Feature-based, item-user similarity	[16], [17]
Graph-based recommenders	User-item relations	No	Relation graph	[18], [19], [20]
Hybrid recommenders	Depends on the integrated recommenders	Depends on the integrated recommenders	Stackable bars, grids Venn diagrams, scatter plots, UpSet matrix, text, ...	[21], [22], [23], [24], [25]
Review-based recommenders	Item aspects	Yes	No presentation	[26], [27], [28], [29], [30], [31]
Review-based recommenders	Rating conversion	Yes	No presentation	[32]
Review-based recommenders	Item aspects	Yes	Aspect-based: text, bar charts, tag clouds	[4], [33], [34], [35], [36], [37], [38]

items because they overlook the consumer feedback provided by online reviews.

Review-based recommender systems [4] extract item features and aspects from online reviews to build user and item models [26]–[30]. Some systems estimate item ratings from their reviews [31], [32], or analyze reviews to evaluate their helpfulness to item evaluation [39]. However, as these systems ignore service modeling, they cannot aggregate the data they extract, and recognize user preferences, with respect to the stages of item fruition.

Differently, we enrich item recommendation with a holistic evaluation of consumer experience during the stages of item fruition. As the Service Journey Maps support the identification of a small number of evaluation dimensions describing such experience, they enable us to replace the detailed item aspects mentioned in the reviews with a few factors to be evaluated in rating estimation. Compared to the research about review-based recommender systems, we analyze item aspects but we synthesize the data they bring directly into the dimensions of experience. Thus, we separate the interpretation of the sentiment emerging from consumer feedback from item evaluation.

C. PRESENTATION OF RECOMMENDATION RESULTS

Different presentation styles are applied to describe results, depending on the recommendation algorithm. In collaborative filtering, users seem to appreciate the bar graphs of neighbors' ratings [13]. Content-based recommender systems typically present suggestions by highlighting the degree of match between item features and user preferences, as in [17], [40]. Moreover, in the research about exploratory search and hybrid recommender systems, several works focus on empowering the user to tune the impact of different relevance perspectives on item recommendation [21]–[24], [41].

Product comparison is a crucial decision stage that buyers usually perform before they make a choice [33]. Some aspect-based recommender systems indirectly support this activity by presenting the features of items which match, or mismatch, the target user's preferences [34]–[36], fusing recommendation and explanation of results to enhance transparency [12], [42]. Other works group items by their properties to facilitate their comparison [33], [37]. Moreover, to support the transparency of Matrix Factorization, McAuley and Leskovec match the item features extracted from reviews to latent factors used in the presentation of results [38].

As all these systems do not model the services behind item fruition, they cannot synthesize the information about consumer experience in this respect. Therefore, they present fairly long lists of aspects that they organize using metadata [37], or which they shorten by removing the less relevant aspects [21].

Differently, our work supports the organization and interpretation of aspects and features with respect to a small number of evaluation dimensions measuring consumer experience, the same for each suggested item, regardless of how many aspects characterize it. This is the basis for the generation of visual overviews of recommendation lists that limit information overload by enabling the user to selectively inspect the details of the relevant items, regarding the evaluation dimensions (s)he cares about.

III. DATA

Our experiments are based on the home-booking domain using data provided by Airbnb. That platform supports searching for homes in a large context that covers both leisure and work time. Similar to other services, such as Booking.com (https://www.booking.com), Airbnb allows customers to write at most one review for each home after the end of the renting contract. This approach enhances the reliability of consumer feedback because it guarantees that comments and evaluations are provided by people who experienced the service.

TABLE 2. Descriptive statistics of the filtered dataset.

	Min	Max	Mean	Standard Deviation
Words per review	1	1002	47.00	46.41
Reviews per listing	1	648	20.80	35.96
Amenities per listing	0	66	20.98	7.85

A. DATASET

For our experiments, we used a public dataset of Airbnb reviews concerning London city.[1] The dataset contains information about homes (denoted as "listings"), their hosts, and the offered amenities, i.e., item features such as Wi-Fi and washing machine. The dataset also stores the reviews about homes uploaded by their renters ("guests") but it does not report the associated ratings. From this dataset, we selected the reviews written in English and we removed the listings that did not receive any comments during the last three years. The filtered dataset contains 764,958 guests, 43,604 listings, and 906,967 reviews. Table 2 provides some descriptive statistics of the filtered dataset.

It can be noticed that several reviews of this dataset are very long and mention a wide spectrum of aspects of homes, hosts, and the surrounding environment. For example: *"The flat was bright, comfortable and clean and Adriano was pleasant and gracious about accommodating us at the last minute. The Brixton tube was a very short walk away and there were plenty of buses. There are lots of fast food restaurants, banks, and shops along the main street."*

B. EVALUATION DIMENSIONS

The service-aware recommender systems we propose build on Mauro et al.'s work [9], which we outline to keep the paper self-contained. Mauro et al. defined a Service Journey Map (SJM) for home-booking by taking inspiration from existing maps developed for hotel booking [43], and from previous analyses about home-booking services [44]. The SJM, shown in the upper portion of Figure 1, focuses on the guest's renting experience, from the search for homes on the Airbnb website to the check-out phase. As it is aimed at describing consumer experience when entering the homes, it overlooks the interaction between the user and backstage services for reservation and payment, and it only models the guest and host roles.

The SJM includes four service stages corresponding to the main activities the guest engages in: `Visit website`, `Check-in`, `Stay in apartment`, `Check-out`. In the present work, we overlook `Visit website` because we are not interested in evaluating the user experience with the Airbnb platform.

In [9], the authors derived from the SJM five evaluation dimensions summarizing guests' renting experience. Moreover, they mapped the stages of the map to these evaluation dimensions. See the lower portion of Figure 1. In the present work we consider four dimensions:

1) `Host appreciation` represents guests' perceptions of the host and of the interaction with her/him at any time of service fruition.
2) `Check-in/Check-out` summarizes guests' experience at check-in and check-out times. It concerns aspects such as timeliness.
3) `In-apartment experience` represents guests' perceptions within the apartment. It covers aspects such as its cleanliness and comfort.
4) `Surroundings` describes the perception of the area where the home is located, in terms of aspects such as available services and quietness.

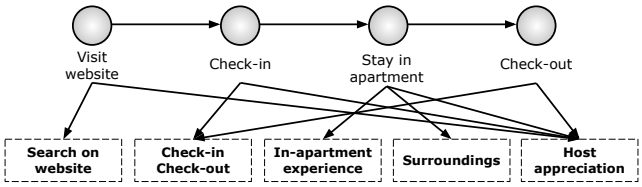

FIGURE 1. This figure is taken from [9]. The upper portion shows the stages of the Service Journey Map describing the home-booking process. Each stage is connected to the associated experience evaluation dimensions.

C. ANALYSIS OF REVIEWS ABOUT HOMES

We organize the opinions emerging from the reviews around the previously described evaluation dimensions. For each home h, we analyze its reviews in three steps that we present in the following subsections.

1) Extraction of aspects from the reviews of h and computation of sentiment

We extract the aspects and corresponding adjectives from the reviews by applying an extension of the Double Propagation algorithm [45] after having analyzed sentences through dependency parsing. After that, we count the number of occurrences ($frequency$) of each $< aspect, adjective >$ pair to measure how frequently people express the corresponding opinion. Moreover, we compute the polarity of the aspect as the mean value returned by the TextBlob [46] and Vader [47] opinion mining libraries and we normalize this value to obtain an $evaluation$ in [0, 1]. The output of this step is a list of $< aspect, adjective, evaluation, frequency >$ tuples, one for each aspect-adjective pair that appears in the reviews of h. See the first four columns of Table 3, which concerns a sample home of our dataset.

2) Classification of aspects in evaluation dimensions

Similar to [9], we group the aspects extracted from the reviews of h by experience evaluation dimension; see the fifth column of Table 3. For this task we use four dictionaries that specify the terms typically used by people to refer to such dimensions. For instance, the `In-apartment`

[1]The dataset is periodically updated and can be downloaded from http://insideairbnb.com/get-the-data.html. We downloaded the dataset for this study in January 2021.

TABLE 3. A subset of the aspects extracted from the reviews of a sample Airbnb home.

Aspect	Adjective	Evaluation	Frequency	Evaluation Dimension
host	wonderful	0.8929	7	`Host appreciation`
host	friendly	0.7172	9	`Host appreciation`
communication	lovely	0.7715	4	`Check-in/Check-out`
check-in	easy	0.7184	5	`Check-in/Check-out`
apartment	nice	0.7554	14	`In-apartment experience`
balcony	sunny	0.6054	2	`In-apartment experience`
bed	comfortable	0.7277	6	`In-apartment experience`
restaurant	good	0.7851	11	`Surroundings`
neighborhood	nice	0.7554	8	`Surroundings`

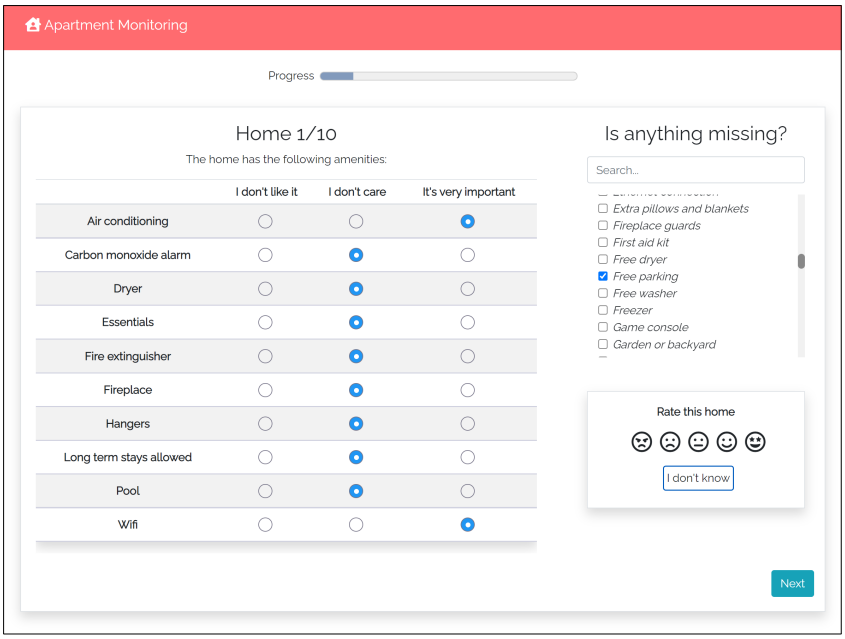

FIGURE 2. User interface to collect the user's preferences in the `FEATURES` and `CBF` recommender systems.

`experience` dictionary includes words like "kitchen", "bed" and "bathroom".

3) Computation of the values of the experience evaluation dimensions of h

Given a home h, let's consider an evaluation dimension d (e.g., `Host appreciation`) and the set AA_{dh} of $< aspect, adjective >$ pairs extracted from the reviews of h that are classified in d. We compute the value of d in h ($value_{dh}$) as the weighted mean of the evaluations of the pairs $p \in AA_{dh}$. For each pair, we use as weight its frequency in the reviews of h to tune its influence based on how many people share the same opinion:

$$value_{dh} = \frac{\sum_{p \in AA_{dh}} frequency_p * evaluation_p}{\sum_{p \in AA_{dh}} frequency_p} \quad (1)$$

where $frequency_p$ is the frequency of pair p in the reviews of h, and $evaluation_p$ is the evaluation of p derived from the polarity of the aspect included in p. For instance, referring to Table 3, for the `Host appreciation` dimension we compute the weighted mean of the evaluation and frequency values of <host, wonderful> and <host, friendly>.

In a preliminary user study, we found that people perceive the lack of information about a home as a negative evaluation factor [48]. Thus, if the reviews of h do not mention any aspects related to a dimension d, or the home has no associated reviews, we set d to 0.1.

IV. RECOMMENDATION MODELS

This section describes the service-aware recommendation models we define and the baselines we use to evaluate them. For each model we present both the algorithm underlying it and the user interface for its evaluation with users. We first present the baselines, which some of our service-aware recommenders integrate into a hybrid system. We adopt the following notation:

- I is the set of items (homes) and U is the set of users (guests).
- For each $i \in I$ and $u \in U$, the **i** and **u** vectors represent the item and user profile, respectively.
- Given i and u, we denote the rating of i estimated by the

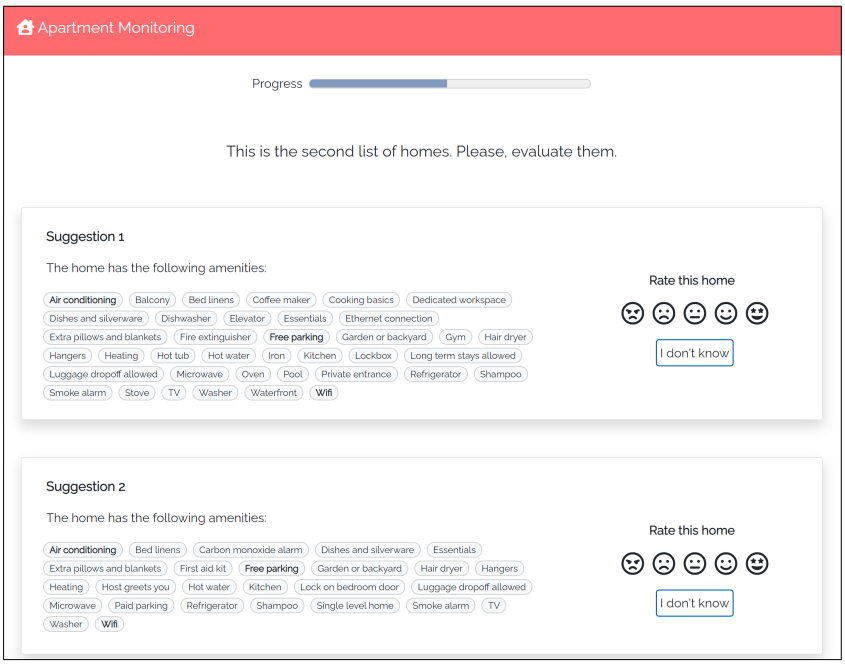

FIGURE 3. User interface for the presentation of suggestions in FEATURES and CBF recommender systems.

recommender system as $\hat{r}_{ui}$.

- D represents the set of experience evaluation dimensions we consider.

A. FEATURES (BASELINE)

1) Model

FEATURES is a feature-based recommender system. In this work, we map features to the overall set of amenities ($\{f_1, \ldots, f_z\}$) offered by the homes:

- **Item profile: i** = $< f_1, \ldots, f_z >$ stores the values of the item features. For $j \in \{1, \ldots, f_z\}$, $f_j = 1$ if i offers the corresponding amenity, 0 otherwise.
- **User profile: u** = $< p_1, \ldots, p_z >$ stores the user's preferences for the item features. For $j \in \{1, \ldots, z\}$, u_j has value 1 ("It's very important"), 0 ("I don't like it"), or 0.5 ("I don't care", default value).

This algorithm focuses on the features that u likes or dislikes. It estimates $\hat{r}_{ui}$ by normalizing in [1, 5] the cosine similarity between the projections of **u** and **i** vectors (denoted as $\vec{\mathbf{u}}$ and $\vec{\mathbf{i}}$) on the components whose value is 0 or 1:

$$\hat{r}_{ui} = 1 + 4 * \frac{\vec{\mathbf{i}} \cdot \vec{\mathbf{u}}}{\|\vec{\mathbf{i}}\|_F \ * \|\vec{\mathbf{u}}\|_F} \tag{2}$$

where $\cdot$ is the scalar vector product, $\| \cdot \|_F$ is the Frobenius Norm and * is the decimal product.

If $\vec{\mathbf{u}}$ is empty, $\hat{r}_{ui}$ is computed by applying a standard popularity-based recommendation algorithm (POP) that suggests the items which received the highest number of reviews.

2) User interface

Acquisition of user preferences (Figure 2). The system shows the amenities offered by the visualized home and enables the user u to declare the importance of the corresponding preferences. The right sidebar enables u to select the amenities that the home lacks but other homes offer; for each selected amenity, the system sets u's preference to "It's very important". However, if u marks the same amenity both as preferred and as disliked when viewing different homes, the ambiguity in the user's behavior is interpreted by setting to "I don't care" the preference in **u**.

The rating elicitation component at the bottom of the page is not relevant to FEATURES but the interface includes it because it is used in CBF; see Section IV-B. This widget shows a list of smilies mapped to the [1, 5] scale, and the "I don't know" button enabling users to opt-out if they are not able to evaluate a home. We omit details that could influence the item evaluation, such as name, price, number of accepted guests, and picture [42].

Presentation of suggestions. Figure 3 shows the user interface supporting the visualization of the recommendation list and the evaluation of items. The amenities (features) that the user has marked either as liked (e.g., Air conditioning), or disliked (none), are in boldface.

B. CBF (BASELINE)

1) Model

This is a content-based recommendation algorithm [16]:

- **Item profile: i** = $< f_1, \ldots, f_z >$ stores the values of the item features. For $j \in \{1, \ldots, z\}$, $f_j = 1$ if i offers the corresponding amenity, 0 otherwise.
- **User profile: u** = $< p_1, \ldots, p_z >$ stores the user's preferences for the item features. Similar to [49], each component of **u** has value 1 if the user has positively

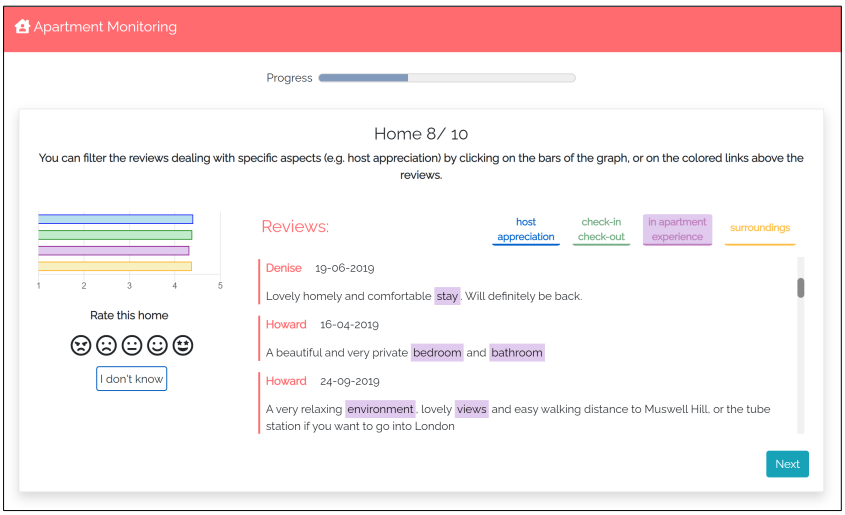

FIGURE 4. User interface to collect user preferences about the experience evaluation dimensions in the STAGES recommender system.

rated (in [4, 5]) at least one item that offers the corresponding feature, 0 otherwise.

If u has positively rated at least one item, CBF evaluates i by computing the cosine similarity between **u** and **i**, normalized in the [1, 5] interval. Otherwise, it uses POP.

2) User interface

For the **acquisition of the user's preferences** this system exploits the user interface shown in Figure 2. The **presentation of the recommendation list** is similar to the one of Figure 3 but does not visualize any features in bold because it does not work on explicit user preferences.

C. STAGES (SERVICE-AWARE)

1) Model

This model employs the information about consumer experience with items in the various fruition stages to generate personalized suggestions and present them to the user. It evaluates items based on a set of experience evaluation dimensions $D = \{d_1, \ldots, d_m\}$ and on their estimated importance to the user. In this work, D = {Host appreciation, Check-in/Check-out, In-apartment experience, Surroundings}. Given $u \in U$ and $i \in I$:

- **Item profile: i** = $< value_1, \ldots, value_m >$ stores the values of the evaluation dimensions extracted from the reviews of i by applying Equation 1. For $j \in \{1, \ldots, m\}$, $value_j$ is the value of d_j; see Table 3.
- **User profile: u** = $< importance_1, \ldots, importance_m >$ stores the estimated importance of $d_1, \ldots, d_m$ to u, i.e., how strongly each of them impacts item selection.
  For $j \in \{1, \ldots, m\}$, we infer $importance_j$ by normalizing in [0, 1] the Pearson correlation between the overall item ratings provided by u, and the values of the evaluation dimension d_j in the respective items. In the computation of the correlation, we ignore the "I don't know" ratings because they are not informative.
  Intuitively, if u evaluates positively the items having high values in d_j, and negatively the items having low values in the same dimension, we hypothesize that d_j is important to her or him. Conversely, if u's ratings are inconsistent with respect to the values of d_j, it is likely that the interest in d_j is low.

Given **u** and **i**, we compute the rating of i as follows:

$$\hat{r}_{ui} = 1 + 4 * \prod_{j=1}^{m} (imp_{ju} * value_{ji} + 1 - imp_{ju}) \quad (3)$$

where imp_{ju} is the importance of dimension d_j in **u** and $value_{ji}$, is the evaluation of dimension d_j in **i**. The $(imp * value + 1 - imp)$ expression tunes the terms of the product (i) by smoothing the impact of low values if they refer to dimensions that u does not care about, and (ii) by maintaining the value of important dimensions thanks to the "$1 - imp$" addendum.

If u has not evaluated any items in the user preferences acquisition phase, STAGES estimates ratings by using POP.

2) User interface

Acquisition of user preferences. Figure 4 shows the user interface to elicit the importance of evaluation dimensions from the user. For each home h, the system shows:

- A bar graph that summarizes the consumer experience with h extracted from its reviews (one colored bar for each evaluation dimension). Even though these values are in [0, 1], the bars are displayed in [1, 5] for coherency with the five-point scale used to rate homes.
- The rating elicitation component to evaluate the home.
- The reviews of h. To support information filtering, the system enables the user to select one or more evaluation dimensions by clicking on the respective bars, or on the list of dimensions located above the reviews. In both

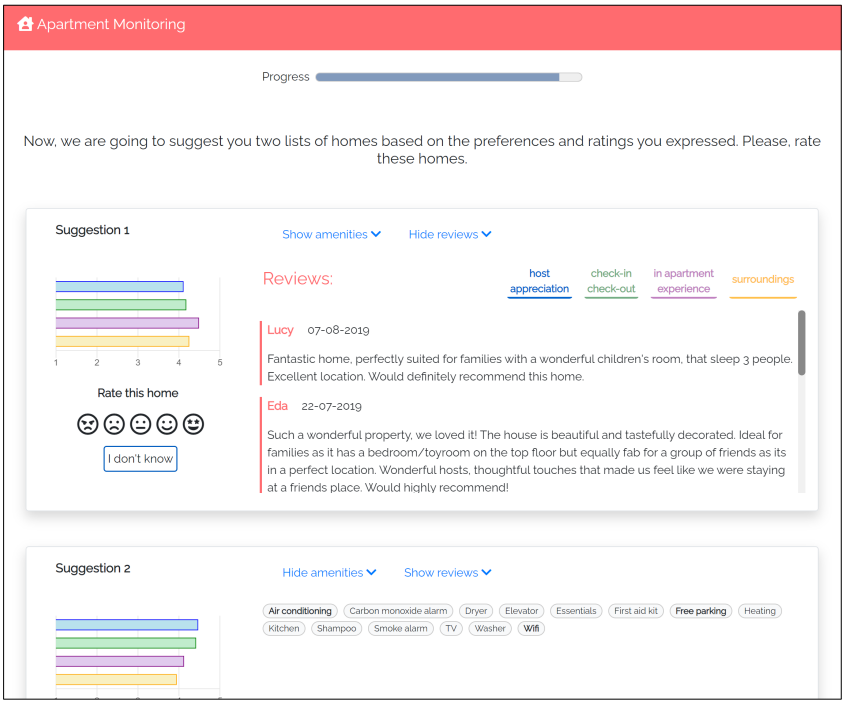

FIGURE 5. User interface for the presentation of suggestions in FEATURES−STAGES and CBF−STAGES recommender systems.

cases, the system shows the comments including at least one aspect that refers to the selected dimension(s). We use color-coding to highlight the corresponding terms in the comments. For instance, the figure shows a selection of reviews related to In−apartment experience. As explained in Section III-C, we use dictionaries to group aspects by dimension.

Presentation of suggestions. This user interface is very similar to Figure 4 but specifies that it shows the personalized suggestions proposed by the system. The bar graph provides the user with a summary of consumer experience with items. Moreover, the user can retrieve detailed comments by inspecting the reviews in a selective way. The amenities offered by the home are hidden.

D. FEATURES−STAGES (SERVICE-AWARE)

1) Model

This algorithm combines the information about item features with the service-based perspective on consumers' experience to offer the user a complete view of items. It integrates feature-based and service-based recommendation by computing item ratings as the arithmetic mean of the ratings estimated by FEATURES and STAGES.

2) User interface

We omit the user interface for the **acquisition of user preferences** because we tested this model on the user profiles built using the user interfaces of FEATURES and STAGES, which provide the preference data to feed it.

In the **presentation of suggestions** we combine the user interfaces of FEATURES and STAGES by using tabs to support the exploration of both types of information. See the two homes visualized in Figure 5.

E. CBF−STAGES (SERVICE-AWARE)

1) Model

This algorithm integrates content-based filtering with service-based recommendation. It computes item ratings as the arithmetic mean of the ratings estimated by CBF and STAGES.

2) User interface

CBF−STAGES uses the same user interface as FEATURES−STAGES to elicit user preferences and present the recommendations to the user.

V. STUDY DESIGN

We aim at testing the recommendation performance and the level of decision-making support provided by the five models described in Section IV.

A. CONTEXT

We carried out the user study by exploiting an interactive test application that we developed to guide participants through the phases of the experiment without our intervention. Section IV has described a portion of the user interface of that system; see Figures 2 to 5.

People joined the study on a voluntary basis, without any compensation, and they gave their informed consent to participate in it. We recruited ($\geq$ 18 years old) participants using social networks and mailing lists. In the message presenting

TABLE 4. Post-task questionnaire. Statements are grouped by user experience construct. Participants answered in the {Strongly disagree, Disagree, Neither agree nor disagree, Agree, Strongly agree} scale.

Construct	Factor	Statement
Perceived Quality of Recommendations (Q)	Q1	The items recommended to me matched my interests.
	Q2	This system gave me good suggestions.
	Q3	The items recommended to me are similar to each other.
Perceived User-Awareness Support (U)	U1	This system explains why the products are recommended to me.
	U2	I understood why the items were recommended to me.
	U3	This recommender system made me more confident about my decision.
Interface Adequacy (I)	I1	The labels of this recommender system interface are clear.
	I2	Finding an item to book with the help of this recommender system is easy.
	I3	The information provided for the recommended items is sufficient for me to make a booking decision.

TABLE 5. Post-test questionnaire. Participants answered the questions in the {Very little, Little, I don't care, Important, Very important} scale.

#	Question
1	How much is the host appreciation important in your choices?
2	How much are check-in/check-out important in your choices?
3	How much is the in-apartment experience important in your choices?
4	How much are the surroundings important in your choices?
5	How much is the visualization of the amenities offered by the home (e.g. WiFi, washing machine, etc.) important in your choices?
6	How much is the visualization of the bar graph characterizing the home (e.g. host appreciation, surroundings, etc.) important in your choices?

the experiment, we specified that we were looking for people who had previously used a home or hotel booking system.

B. METHOD

We applied the within-subjects design to the user study. We considered each treatment condition as an independent variable and every participant received all the treatments. In the test application, we counterbalanced the order of tasks to minimize the impact of result biases and the effects of practice and fatigue. The experiment took on average 36.79 minutes with a Standard Deviation = 19.83. To comply with diverse users' backgrounds and levels of confidence with technology, we did not impose any time limits to complete the study, which was organized in three phases:

1) The test application asked users to declare whether they were $\geq$ 18 years old or not; moreover, it asked them to express their consent to participate in the study. People could continue the test only if they positively answered the first question and they accepted the consent.[2]
   Then, the application proposed participants to fill in a questionnaire that inquires basic demographic information, cultural background, familiarity with booking platforms, and whether they tend to trust a person or thing, even though they have little knowledge about it. The questionnaire is an adaptation of the ResQue one for recommender systems [50].
2) The application acquired participants' preferences and built their user profiles.[3] For this purpose, it asked people twice, in different moments of the experiment, to rate ten homes; see Figures 2 and 4. The application also asked to rate the homes presented in one suggestion list for each tested algorithm (Figures 3 and 5). Each list contained five homes to be evaluated according to their suitability as candidates for rent, using the star-based rating elicitation component.
   After the evaluation of each recommendation list, the test application proposed a post-task questionnaire in which users declared their degree of agreement with the statements reported in Table 4. The questionnaire is a subset of ResQue. In the table, statements are grouped in three constructs: *Perceived Quality of Recommendations* (Q); *Perceived User-Awareness Support* (U), and *Interface Adequacy* (I).
3) Home-booking is a high-investment domain: similar to [51], the definition of "investment" rests on the concept of price. Thus, we hypothesized that people need detailed information and feedback about items to make a renting decision. To check this hypothesis, before closing the experiment, our application asked participants to answer the post-test questionnaire of Table 5. This questionnaire is aimed at understanding to what extent they considered the visualization of amenities and the summarization of consumer experience important in the evaluation of the system's suggestions.

[2]The text of the consent is available here: https://bit.ly/3jjYlEa.

[3]As some recommender models share the user interface for the acquisition of the user profiles, the first model selected for execution acquired the user preferences and propagated them to the other models.

TABLE 6. Recommendation performance of algorithms. The best results are in boldface. For each evaluation metric, (*) denotes different levels of the statistical significance of the difference between the best performing algorithm, and the other ones. The last column shows the number of "I don't know" evaluations provided by participants when using the algorithms.

Algorithm	RMSE	MAE	NDCG	Utility	#Opting out
FEATURES	**0.6919**	**0.5170**	0.9792	5.1805	6
CBF	0.8857	0.7219	0.9669*	3.9482*	10
STAGES	0.8561	0.7393	**0.9847**	**5.3388**	9
FEATURES−STAGES	0.7225	0.5883	0.9736	4.7379	0
CBF−STAGES	0.9829	0.7900	0.9612*	3.4969*	0

VI. EXPERIMENTAL RESULTS

A. DEMOGRAPHIC DATA AND BACKGROUND

We conducted a power analysis to determine the minimum number of participants to obtain statistically significant results. A calculation of power analysis involves the following four parameters: *Alpha (α = 0.05):* a p value that indicates the probability threshold for rejecting the null hypothesis when there is no significant effect (Type I error rate). *Power = 0.80:* the probability of accepting the alternative hypothesis if it is true (Type II error rate). *Effect size = 0.40:* the expected effect size, i.e., the quantified magnitude of a result present in the population; our goal was to find medium-sized effects. *Sample size N:* the required size of the sample of participants to maintain statistical power. The estimation of the sample size resulted in N = 42 that supports the actual statistical power of 80%.

Having set the minimum sample size to N = 42, we recruited 48 participants (N = 48) for the user study from May 15 to June 15, 2021. The subjects are 20 female, 28 male, 0 non-binary, 0 did not answer. Their age is distributed as follows: 1 person in range 18-20; 30 in 21-30; 11 in 31-40; 1 in 41-50; 4 in 51-60; 1 older than 60. Regarding the education level, 4 subjects attended the high school, 34 the university and 10 have a Ph.D. 17 people have a technical background, 22 a scientific one, 5 humanities and languages, 3 economics and 1 other background. 37 participants declared that they are advanced computer users, 9 average ones and 2 beginners. 15 people declared that they use e-commerce platforms or online booking services few times a month, 8 use them 1-3 times a week, 11 daily, and 14 a few times a year. Finally, 4 participants declared that they very probably would trust a person or thing, even though they had little knowledge about it, 15 probably would trust it, 23 probably would not trust it, and 6 very probably would not trust it.

B. RECOMMENDATION QUALITY

We evaluated the recommendation performance of the algorithms by focusing on ranking because the placement of good solutions at the top of a recommendation list is important to support their identification. Moreover, we considered the minimization of rating estimation errors as an accuracy measure. We computed the following metrics:

- NDCG (Normalized Discounted Cumulative Gain). It measures the ranking quality. The gain of items is accumulated from the top of the result list to the bottom and it is discounted logarithmically at lower ranks.
- RMSE (Root-Mean-Square Error) and MAE (Mean Absolute Error). They are used to compute the error between the ratings predicted by the algorithm, and the real rating given by the participants of the user study.
- Utility. This accuracy metric computes a score for the whole list (rather than individual items) based on user ratings. The worth of the suggested items declines for the lower positions of the list. The formula for a list of five suggestions is the following:

$$Ut_u = \sum_{j=1}^{5} \frac{max(r_{ui_j} - n), 0}{2^{\frac{j-1}{\alpha-1}}} \quad (4)$$

where r_{ui_j} is the rating given by a user u to the item in the j^{th} position; n represents the neutral vote (we set it to 3); α is a half-life parameter that corresponds to the position of the item in the list with 50% chance of being inspected and rated by the user. In our experiments, users rated all the five items of the list, thus $\alpha = 5$.

Table 6 shows the evaluation results. We conducted a one-way ANOVA analysis to compare the performance of the algorithms. We only computed it on NDCG and Utility because RMSE and MAE are not computed per user, but on the overall set of ratings. We found significance on both metrics: NDCG [$F(232,4) = 4.31$; $p < 0.003$], and Utility [$F(232,4) = 7.58$; $p < 0.001$].

We then conducted a *post-hoc* comparison using a Tukey HSD test. We found that STAGES has the best NDCG, with significant results compared to CBF ($p < 0.05$), and CBF−STAGES ($p < 0.003$). Regarding the Utility, the best performing model is again STAGES that obtained significant results compared to CBF ($p < 0.01$), and CBF−STAGES ($p < 0.001$). As far as the minimization of error in rating estimation is concerned, the best performing algorithm is FEATURES.

The last column of Table 6 reports the number of opting-outs ("I don't know" ratings) in the evaluation of homes. This phenomenon was more frequent when using CBF (10 occurrences), STAGES (9 occurrences), and FEATURES (6 cases). Differently, CBF−STAGES and FEATURES−STAGES, which also show item reviews, did not get any "I don't know" evaluations.

TABLE 7. Post-task questionnaire results for each recommender system. For each mean value, the asterisks denote the statistical significance of the difference between the best-performing algorithm and the other ones. Significance levels: (***)$p < 0.001$, (**)$p < 0.05$.

Construct	Factor	Recommendation Algorithm				
		FEATURES Mean(SD)	CBF Mean(SD)	STAGES Mean(SD)	FEATURES−STAGES Mean(SD)	CBF−STAGES Mean(SD)
Perceived Quality of Recommendations (Q)	Q1	**4.31(0.72)**	3.83(0.97)	4.15(0.68)	4.27(0.71)	3.54(1.13)
	Q2	**4.29(0.62)**	3.85(0.99)	4.15(0.82)	4.25(0.84)	3.58(1.05)
	Q3	3.92(0.79)	3.88(0.87)	**3.94(0.76)**	3.85(0.74)	3.56(0.85)
	Mean	**4.17(0.73)**	3.85(0.94)	4.08(0.76)	4.12(0.78)	3.56(1.01)***
Perceived User-Awareness Support (U)	U1	3.29(1.20)	3.07(1.21)	3.05(1.17)	**3.62(1.09)**	3.49(0.92)
	U2	3.94(0.79)	3.94(0.92)	3.52(0.82)	**4.19(0.68)**	3.72(0.88)
	U3	3.60(1.05)	3.35(1.14)	3.50(0.90)	**3.81(0.89)**	3.54(1.07)
	Mean	3.61(1.04)	3.46(1.14)	3.36(0.99)**	**3.88(0.93)**	3.59(0.96)
Interface Adequacy (I)	I1	3.81(0.92)	3.74(1.03)	3.87(1.01)	3.96(0.81)	**3.98(0.82)**
	I2	3.60(1.09)	3.38(1.12)	3.69(0.78)	**3.92(0.87)**	3.69(0.99)
	I3	3.27(1.16)	3.02(1.34)	3.15(0.95)	**3.71(0.97)**	3.63(1.06)
	Mean	3.56(1.08)	3.38(1.20)**	3.57(0.96)	**3.86(0.88)**	3.76(0.97)

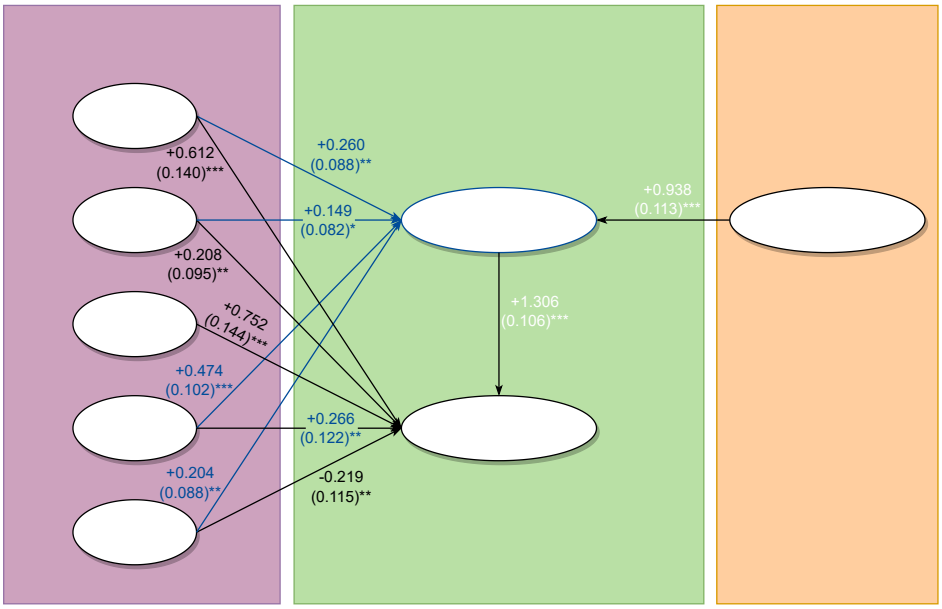

FIGURE 6. Structural Equation Model. Significance levels: (***)$p < 0.001$, (**)$p < 0.05$, (*)$p < 0.1$. The numbers on the arrows represent the β-coefficients (and standard error) of the effect.

C. USER FEEDBACK ANALYSIS

1) User experience with the recommender systems

Table 7 shows the results of the post-task questionnaire for each of the tested algorithms, grouped by user experience constructs. A one-way ANOVA analysis comparing user experience in the recommendation algorithms showed significance on all the constructs:

- *Perceived Quality of Recommendations* (Q) [$F(235,4) = 7.34$; $p < 0.001$];
- *Perceived User-Awareness Support* (U) [$F(235,4) = 2.69$; $p < 0.05$];
- *Interface Adequacy* (I) [$F(235,4) = 2.53$; $p < 0.05$].

Moreover, according to a *post-hoc* comparison based on a Tukey HSD test:

- Concerning the *Perceived Quality of Recommendations* (Q), FEATURES (M=4.17, SD=0.73) is the best algorithm for quality of suggestions. FEATURES is only significantly higher compared to CBF−STAGES. However, FEATURES−STAGES (M=4.12, SD=0.78) is the second-best algorithm and obtains the best results in the other constructs.
- Regarding the *Perceived User-Awareness Support* (U), FEATURES−STAGES (M=3.88, SD=0.93) is the best algorithm. FEATURES−STAGES obtains significantly higher results compared to STAGES ($p < 0.05$).
- Looking at the *Interface Adequacy* (I), FEATURES−STAGES (M=3.86, SD=0.88) is the best algorithm and has significantly higher results than CBF ($p < 0.05$).

2) Structured Equation Model analysis

We performed the Structured Equation Model analysis [52] to gain a deep understanding of the user experience with the five recommenders. This analysis is useful to find the relationship between unobserved constructs (latent variables) by leveraging observable variables. It is difficult to define measures that perfectly represent user experience with an intelligent system that includes a recommendation algorithm. However, we can use different elements defined by statements to measure user experience and group them into constructs to find the relations with the algorithms.

Based on the post-task questionnaire 4, we associated two constructs (*Perceived User-Awareness Support* and *Perceived Quality of Recommendations*) to Decision-making Support (DS) aspects; one construct (*Interface Adequacy*) to User Interfaces aspects, and we tested five Algorithms (ALG) represented as dummy variables (`CBF`, `FEATURES`, `STAGES`, `CBF-STAGES` and `FEATURES-STAGES` in Figure 6). These constructs are good candidates for a Structured Equation Model because they include at least three statements each.

We performed the Confirmatory Factor Analysis to check the validity of the constructs. This analysis requires:

1) The computation of the convergent validity to check that the statements of the constructs are related. For this purpose, we examined the Average Variance Extracted (AVE) of each construct, which must be over 0.50 to be respected.
2) The computation of the discriminant validity to check that the statements belonging to different constructs are not related. In this case, the squared root of the AVE value must be less than the correlation value.

All the constructs we defined respected the required constraints:

- *Perceived User-Awareness Support*: $AVE = 0.5463$, $\sqrt{AVE(0.5463)} = 0.7391$, largest correlation = 0.410.
- *Perceived Quality of Recommendations*: $AVE = 0.5913$, $\sqrt{AVE(0.5913)} = 0.7690$, largest correlation = 0.410.
- *Interface Adequacy*: $AVE = 0.5983$, $\sqrt{AVE(0.5913)}$ = 0.7735, largest correlation = 0.337.

Figure 6 shows the Structural Equation Model with dependencies and β-coefficients and standard error that indicate the correlations between the constructs. The *Interface Adequacy* has a positive effect (+0.938; $p < 0.001$) on the *Perceived User-Awareness Support*. This can be explained by the fact that the user-awareness support given by the system is influenced by how items are presented. Moreover, there is a positive correlation (+1.306; $p < 0.001$) between the *Perceived User-Awareness Support* and the *Perceived Quality of Recommendations*. This suggests that, when users feel that they have enough information about items, they perceive that the suggestions have higher quality.

All the algorithms, except for `STAGES`, positively affect the *Perceived User-Awareness Support*. This suggests that consumer feedback alone is not enough to choose a home for rent. Indeed, consumer feedback does not guarantee that the home has the amenities that the user needs. It is worth noticing that `FEATURES-STAGES` shows the largest correlation value with *Perceived User-Awareness Support* (+0.474; $p < 0.001$). We can explain this with the fact that, by explicitly listing the offered amenities, the overview of consumer feedback (bar graph), and the reviews, the algorithm supports decision-making in a complete way. Looking at the *Perceived Quality of Recommendations*, we observe that all the algorithms, except for `CBF-STAGES`, have a positive correlation with this aspect. We believe that `CBF-STAGES` has a negative correlation (-0.219; $p < 0.05$) because of its low evaluation performance in accuracy, ranking and error estimation; see Section VI-B. Finally, `STAGES` has the largest correlation (+0.752; $p < 0.001$) with *Perceived Quality of Recommendations*. We explain this finding with the fact that consumer feedback is a very useful information source to generate good predictions.

3) Post-test results

Table 8 shows the results of the post-test questionnaire. `In-apartment experience` and `Surroundings` emerge as the most important dimensions to decision-making. The situation of `Host` and `Check-in/Check-out` is mixed because several participants consider them as important or very important but a few ones declare that they are unimportant or little important. As far as the visualization of information is concerned, the amenities are considered as important more frequently than the bar graphs. This is probably due to the fact that people want to be sure that the selected homes offer the features they care about.

VII. DISCUSSION

A. DISCUSSION OF RESULTS

The user study provides interesting findings about the objective and perceived performance of the models we tested regarding both the recommendation of items, and the visualization of results.

The recommendation performance measures show that `STAGES`, that relies exclusively on the evaluation of user experience in the stages of item fruition, achieves the best results concerning the ranking of items. This finding suggests that the experience evaluation dimensions are a precious summary for the identification of relevant items. `FEATURES` achieves the best results regarding the minimization of error in rating estimation. However, this is a secondary finding because our first goal is that of promoting good items in the recommendation lists.

Notice that the recommender systems that use a single type of information, i.e., either user experience data (`STAGES`), or features (`CBF`, and `FEATURES`), received some opting outs from participants. Differently, when users interacted with the systems that combine these types of information (`CBF-STAGES`, `FEATURES-STAGES`), they were able to

TABLE 8. Post-test questionnaire results.

Importance of dimensions in users' choices (number of users)					
	Very little	Little	I don't care	Important	Very important
Host	9	9	4	16	10
Check-in/Check-out	9	10	4	22	3
In-apartment experience	1	3	0	17	27
Surroundings	8	2	0	21	17
Importance of visualization of information in users' choices (number of users)					
Amenities	1	3	2	16	26
Bar graphs	3	6	4	24	11

evaluate all the suggested items. This is a first indication that the joint presentation of data about item features and user experience enhances users' confidence in item evaluation. FEATURES−STAGES, which combines these two types of information, is the second-best algorithm for rating estimation and obtains fairly good NDCG results.

As far as our first research question (RQ1) is concerned, these findings support the hypothesis that the integration of a service-based representation of items with data about their features improves recommendation accuracy. Indeed, we obtain the best results by only relying on service-based information about items (STAGES); however, in that case, some users do not feel confident in decision-making. Thus, a good compromise between recommendation quality and coverage is the integration of consumer experience and item features in the presentation of the suggestions, as done in FEATURES−STAGES. That system achieved the second-best ranking performance and did not get any opting-outs.

To answer RQ2, we analyze participants' perceptions after having interacted with the systems. Regarding the *Perceived Quality of Recommendations* (Q), people perceived FEATURES as the model that generates the best suggestions. This is probably due to its coherence with respect to the user's requirements. In fact, that system recommends the homes that reflect the amenities marked as important during preference elicitation and it highlights them in bold in the presentation of results. However, FEATURES−STAGES is perceived as the best system regarding both *Perceived User-Awareness Support* (U) and *Interface Adequacy* (I), which describe users' comprehension of the rationale behind the recommendations, their awareness about the suggestions, and their confidence in decision-making. We explain this finding with the fact that by showing amenities, bar graphs, and item reviews, the system helps users analyze and compare candidate homes in a more efficacious way than by only presenting amenity data.

The Structural Equation Model confirms these results. The *Interface Adequacy* (I), has a positive effect on the *Perceived User-Awareness Support* (U) because by providing more data about items the system makes the user more confident about the available options to choose from. Moreover, the *Perceived User-Awareness Support* (U) positively influences the *Perceived Quality of Recommendations* (Q) because, to perceive the suggestions as good ones, the user needs a sufficient amount of data about items.

The results of the post-test questionnaire show that participants considered the visualization of data about the amenities offered by the homes as more important than the bar graphs summarizing consumer experience. However, by jointly considering these results, and the fact that FEATURES−STAGES is recognized as the algorithm providing the highest user-awareness support, we conclude that both offered amenities, and data extracted from consumer feedback, are key to decision-making.

Given all these findings, we can positively answer research question RQ2: if a recommender system presents both item features and service-based data in the suggestion lists, it enhances users' awareness about the available options, as well as their confidence in decision-making. The reason is that it provides people with complete information to evaluate items from the viewpoint of their features and of the other aspects concerning item fruition.

B. THEORETICAL IMPLICATIONS

This work advances the state of the art in recommender systems and in particular of review-based and aspect-based ones ([13], [20], [28], [31], [33], [42], [53]) by integrating service models in rating estimation and in the presentation of results. Review-based recommender systems use consumer experience about items to integrate metadata with aspects extracted from online reviews. However, they extract item-centric data which fail to overview the expected experience at fruition time. Differently, we model these stages and we group the aspects extracted from consumer feedback in evaluation dimensions aimed at separately measuring user experience. This approach makes it possible to weight aspects in different ways, depending on the importance of the individual evaluation dimensions to the user. Moreover, it supports the summarization of previous consumers' experience to enhance item evaluation and presentation within a recommendation list. The user study we carried out showed that our approach enhances recommendation performance, user-awareness about the suggested options, and users' confidence in item-selection decisions.

VIII. LIMITATIONS AND FUTURE WORK

The first limitation of our work concerns the number of participants involved in the user test. Even though the power analysis that we conducted suggests that this number is enough to obtain a robust statistical evaluation, we plan to test our systems with a larger number of users to increase the statistical power of the experiments. With a larger number of participants, we could develop, and test, service-based recommendation algorithms based on Collaborative Filtering such as the multi-criteria ones presented in [14] and [54].

The second limitation concerns the extraction of the aspects from the reviews. In this work, we leveraged the method described in [45] that uses dependency parsing to analyze textual information. This is a non-supervised opinion mining technique and does not require a large annotated dataset for training. However, it bases the match between aspects and dimensions on *ad hoc* dictionaries. We plan to extend our model with semantic Natural Language Processing techniques to extract aspects, and their synonyms, using standard language resources.

We also plan to investigate other models to define the service fruition stages and the dimensions for the evaluation of experience that underlie recommendation. So far, we leveraged the largely used Service Journey Maps. However, other approaches, such as the Service Blueprints [55], can be used to develop finer-grained service models. We also plan to test our recommender systems on the sales of experience products to assess their applicability to heterogeneous items. The specification of a new application domain is supported by the existence of service models that can be adapted to the peculiarities of the selected domain.

IX. ETHICAL ISSUES

In planning the user study we complied with literature guidelines on controlled experiments[4] [56]. Through the user interface of our test application, participants were informed about their rights:

- the right to stop participating in the experiment, possibly without giving a reason;
- the right to obtain further information about the purpose, and the outcomes of the experiment;
- the right to have their data anonymized.

As described in Section V, before starting the experiment, participants were asked to: (i) read a consent form, stating the nature of the experiment and their rights, (ii) confirm that they had read and understood their rights by clicking on the user interface of the test application, and (iii) confirm that they were 18 years old or over. Every participant was given the same instructions before the experimental tasks.

We did not store participants' names. During the user study, and the analysis of its results, we worked with anonymous codes.

[4]https://www.tech.cam.ac.uk/research-ethics/school-technology-research-ethics-guidance/controlled-experiments

X. CONCLUSIONS

In this paper, we pointed out that current recommender systems use item-centric data to estimate ratings and to present their results. Even though review-based recommender systems extract aspects from consumer feedback, they overlook the user experience during all the stages of item fruition, which is key to decision-making.

In order to address this limitation, we investigated the integration of recommender systems with service modeling to explicitly represent the evaluation dimensions of consumer experience during the stages of item fruition. Building on existing analyses of user experience with items, we developed different recommendation models that employ item features, experience evaluation dimensions and their combination to recommend and holistically present items to the user. The novelty of our approach is that we group the aspects of items extracted from reviews based on these evaluation dimensions, around which we organize preference modeling, recommendation, and information presentation. This enables us to steer the suggestions to the user's preferences for all such dimensions, and to summarize the user experience with items enhancing the identification of relevant options within the recommendation lists. In a user study, we found that, compared to state of the art recommender systems, our approach enhances recommendation performance, user-awareness about items and confidence in decision-making. These findings encourage the adoption of service-based models in recommender systems research.

XI. ACKNOWLEDGMENTS

We thank Michele Colombino and Gianmarco Izzi for having helped us with the development of the user interfaces of the recommended systems. This work has been funded by the University of Torino under grant ARDL_RILO_2019.

REFERENCES

[1] Marc Stickdorn, Jakob Schneider, and Kate Andrews. This is service design thinking: Basics, tools, cases. Wiley, 2011.
[2] Carmen Kar Hang Lee. How guest-host interactions affect consumer experiences in the sharing economy: New evidence from a configurational analysis based on consumer reviews. Decision Support Systems, 152:113634, 2022.
[3] Francesco Ricci, Lior Rokach, and Bracha Shapira. Recommender Systems: Introduction and Challenges, pages 1–34. Springer US, Boston, MA, 2015.
[4] Li Chen, Guanliang Chen, and Feng Wang. Recommender systems based on user reviews: the state of the art. User Modeling and User-Adapted Interaction, 25(2):99–154, 2015.
[5] Anindya Ghose and Panagiotis G. Ipeirotis. Estimating the helpfulness and economic impact of product reviews: mining text and reviewer characteristics. IEEE Transactions on on Knowledge and Data Engineering, 23(10):1498–1512, 2011.
[6] Adam Richardson. Using Customer Journey Maps to improve customer experience. Harvard Business Review, 2015.
[7] Jisu Yi, Gao Yuan, and Changsok Yoo. The effect of the perceived risk on the adoption of the sharing economy in the tourism industry: The case of airbnb. Information Processing & Management, 57(1):102108, 2020.
[8] Shini Renjith, A. Sreekumar, and M. Jathavedan. An extensive study on the evolution of context-aware personalized travel recommender systems. Information Processing & Management, 57(1):102078, 2020.
[9] Noemi Mauro, Liliana Ardissono, Sara Capecchi, and Rosario Galioto. Service-aware interactive presentation of items for decision-making. Ap-

plied Sciences, Special Issue Implicit and Explicit Human-Computer Interaction, 10(16):5599, 2020.

[10] Xia Ning, Christian Desrosiers, and George Karypis. A Comprehensive Survey of Neighborhood-Based Recommendation Methods, pages 37–76. Springer US, Boston, MA, 2015.

[11] Yehuda Koren and Robert Bell. Advances in Collaborative Filtering, pages 77–118. Springer US, Boston, MA, 2015.

[12] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. How should i explain? a comparison of different explanation types for recommender systems. International Journal of Human-Computer Studies, 72(4):367–382, 2014.

[13] Ingrid Nunes and Dietmar Jannach. A systematic review and taxonomy of explanations in decision support and recommender systems. User Modeling and User-Adapted Interaction, 27(3–5):393–444, December 2017.

[14] Gediminas Adomavicius and YoungOk Kwon. Multi-Criteria Recommender Systems, pages 847–880. Springer US, Boston, MA, 2015.

[15] Pan Li and Alexander Tuzhilin. Latent multi-criteria ratings for recommendations. In Proceedings of the 13th ACM Conference on Recommender Systems, RecSys '19, pages 428–431, New York, NY, USA, 2019. ACM.

[16] Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. Content-based recommender systems: state of the art and trends, pages 73–105. Springer US, Boston, MA, 2011.

[17] Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, and Katrien Verbert. What's in a user? towards personalising transparency for music recommender interfaces. In Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, UMAP '20, page 173–182, New York, NY, USA, 2020. Association for Computing Machinery.

[18] Saeed Amal, Chun-Hua Tsai, Peter Brusilovsky, Tsvi Kuflik, and Einat Minkov. Relational social recommendation: application to the academic domain. Expert Systems with Applications, 124:182 – 195, 2019.

[19] Hongwei Wang, Fuzheng Zhang, Jialin Wang, Miao Zhao, Wenjie Li, Xing Xie, and Minyi Guo. Ripplenet: propagating user preferences on the knowledge graph for recommender systems. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM '18, page 417–426, New York, NY, USA, 2018. Association for Computing Machinery.

[20] Cataldo Musto, Fedelucio Narducci, Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. Linked open data-based explanations for transparent recommender systems. International Journal of Human-Computer Studies, 121:93 – 107, 2019. Advances in Computer-Human Interaction for Recommender Systems.

[21] Cecilia di Sciascio, Peter Brusilovsky, Christoph Trattner, and Eduardo Veas. A roadmap to user-controllable social exploratory search. ACM Transaction on Interactive Intelligent Systems, 10(1), aug 2019.

[22] Bruno Cardoso, Gayane Sedrakyan, Francisco Gutiérrez, Denis Parra, Peter Brusilovsky, and Katrien Verbert. IntersectionExplorer, a multi-perspective approach for exploring recommendations. International Journal of Human-Computer Studies, 121:73 – 92, 2019.

[23] Chun-Hua Tsai and Peter Brusilovsky. Exploring social recommendations with visual diversity-promoting interfaces. ACM Transactions on Interactive Intelligent Systems, 10(1):5:1–5:34, August 2019.

[24] Katrien Verbert, Denis Parra, and Peter Brusilovsky. Agents vs. users: visual recommendation of research talks with multiple dimension of relevance. ACM Transactions on Interactive Intelligent Systems, 6(2), July 2016.

[25] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. Personalized explanations for hybrid recommender systems. In Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI '19, page 379–390, New York, NY, USA, 2019. Association for Computing Machinery.

[26] Tong Zhao, Julian McAuley, and Irwin King. Improving latent factor models via personalized feature projection for one class recommendation. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM '15, pages 821–830, New York, NY, USA, 2015. ACM.

[27] Claudiu-Cristian Musat and Boi Faltings. Personalizing product rankings using collaborative filtering on opinion-derived topic profiles. In Proceedings of the 24th International Conference on Artificial Intelligence, IJCAI'15, page 830–836. AAAI Press, 2015.

[28] María Hernández-Rubio, Iván Cantador, and Alejandro Bellogín. A comparative analysis of recommender systems based on item aspect opinions extracted from user reviews. User Modeling and User-Adapted Interaction, 29(2):381–441, 2019.

[29] Heng-Ru Shen, Rong-Ping andZhang, Hong Yu, and Fan Min. Sentiment based matrix factorization with reliability for recommendation. Expert Systems with Applications, 2019.

[30] Guang-Neng Hu, Xin-Yu Dai, Feng-Yu Qiu, Rui Xia, Tao Li, Shu-Jian Huang, and Jia-Jun Chen. Collaborative filtering with topic and social latent factors incorporating implicit feedback. ACM Transactions on Knowledge Discovery from Data, 12(2):23:1–23:30, January 2018.

[31] Xiaojiang Lei, Xueming Qian, and Guoshuai Zhao. Rating prediction based on social sentiment from textual reviews. IEEE Transactions on Multimedia, 18(9):1910–1921, 2016.

[32] Dionisis Margaris, Costas Vassilakis, and Dimitris Spiliotopoulos. What makes a review a reliable rating in recommender systems? Information Processing & Management, 57(6):102304, 2020.

[33] Li Chen, Feng Wang, Luole Qi, and Fengfeng Liang. Experiment on sentiment embedded comparison interface. Knowledge-Based Systems, 64:44–58, 2014.

[34] Khalil Ibrahim Muhammad, Aonghus Lawlor, and Barry Smyth. A live-user study of opinionated explanations for recommender systems. In Proceedings of the 21st International Conference on Intelligent User Interfaces, IUI '16, page 256–260, New York, NY, USA, 2016. Association for Computing Machinery.

[35] Jianmo Ni and Julian McAuley. Personalized review generation by expanding phrases and attending on aspect-aware representations. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 706–711, Melbourne, Australia, 2018. Association for Computational Linguistics.

[36] Yichao Lu, Ruihai Dong, and Barry Smyth. Coevolutionary recommendation model: Mutual learning between ratings and reviews. In Proceedings of the 2018 World Wide Web Conference, WWW '18, page 773–782, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.

[37] Li Chen and Feng Wang. Explaining recommendations based on feature sentiments in product reviews. In Proceedings of the 22nd International Conference on Intelligent User Interfaces, IUI '17, page 17–28, New York, NY, USA, 2017. Association for Computing Machinery.

[38] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In Proceedings of the 7th ACM Conference on Recommender Systems, RecSys '13, page 165–172, New York, NY, USA, 2013. Association for Computing Machinery.

[39] Noemi Mauro, Liliana Ardissono, and Giovanna Petrone. User and item-aware estimation of review helpfulness. Information Processing & Management, 58(1):102434, 2021.

[40] Martijn Millecamp, Cristina Conati, and Katrien Verbert. "knowing me, knowing you": personalized explanations for a music recommender system. User Modeling and User-Adapted Interaction, 2022.

[41] Benedikt Loepp, Katja Herrmanny, and Jürgen Ziegler. Blended recommending: integrating interactive information filtering and algorithmic recommender techniques. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI '15, pages 975–984, New York, NY, USA, 2015. ACM.

[42] Nava Tintarev and Judith Masthoff. Explaining recommendations: design and evaluation, pages 353–382. Springer US, Boston, MA, 2015.

[43] Lianping Ren, Hanqin Qiu, Peilai Wang, and Pearl M.C. Lin. Exploring customer experience with budget hotels: Dimensionality and satisfaction. International Journal of Hospitality Management, 52:13 – 23, 2016.

[44] Mingming Cheng and Xin Jin. What do airbnb users care about? an analysis of online review comments. International Journal of Hospitality Management, 76:58 – 70, 2019.

[45] Guang Qiu, Bing Liu, Jiajun Bu, and C. Chen. Opinion word expansion and target extraction through double propagation. Computational Linguistics, 37:9–27, 03 2011.

[46] Steven Loria. TextBlob: Simplified text processing. 2020. https://textblob.readthedocs.io/en/dev/index.html.

[47] C.J. Hutto and Gilbert Eric. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the 8th International AAAI Conference on Weblogs and Social Media, pages 216–225, New York, NY, USA, 2014. AAAI.

[48] Noemi Mauro, Zhongli Filippo Hu, Liliana Ardissono, and Gianmarco Izzi. A service-oriented perspective on the summarization of recommendations. In Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, Utrecht, The Netherlands, 2021. ACM.

[49] Roberto Saia, Ludovico Boratto, and Salvatore Carta. A class-based strategy to user behavior modeling in recommender systems. In Liming

Chen, Supriya Kapoor, and Rahul Bhatia, editors, Emerging Trends and Advanced Technologies for Computational Intelligence. Springer International Publishing, Cham, Switzerland, 2016.

[50] Pearl Pu, Li Chen, and Rong Hu. A user-centric evaluation framework for recommender systems. In Proceedings of the Fifth ACM Conference on Recommender Systems, RecSys '11, page 157–164, New York, NY, USA, 2011. Association for Computing Machinery.

[51] Nava Tintarev and Judith Masthoff. Evaluating the effectiveness of explanations for recommender systems. User Modeling and User-Adapted Interaction, 22(4-5):399–439, 2012.

[52] Jodie B. Ullman and Peter M. Bentler. Structural equation modeling. Handbook of Psychology, Second Edition, 2, 2012.

[53] Cataldo Musto, Marco de Gemmis, Pasquale Lops, and Giovanni Semeraro. Generating post hoc review-based natural language justifications for recommender systems. User-Modeling and User-Adapted Interaction, 27, 2020.

[54] Yukuo Cen, Jianwei Zhang, Xu Zou, Chang Zhou, Hongxia Yang, and Jie Tang. Controllable Multi-Interest Framework for Recommendation, page 2942–2951. Association for Computing Machinery, New York, NY, USA, 2020.

[55] Mary Jo Bitner, Amy L. Ostrom, and Felicia N. Morgan. Service blueprinting: A practical technique for service innovation. California Management Review, 50(3):66–94, 2008.

[56] R. E. Kirk. Experimental design: Procedures for the behavioral sciences. SAGE Publications, Inc., 2013. https://www.doi.org/10.4135/9781483384733.

NOEMI MAURO is an Assistant Professor at the Computer Science Department of the University of Torino where she obtained a Ph.D. in Computer Science with Honors. Her research interests concern information filtering, information visualization, user modeling, and recommender systems. She has been a visiting researcher at the Alpen-Adria-Universität Klagenfurt and at the University College Dublin where she worked on topics related to recommender systems. She is the author of several research articles related to intelligent systems and information filtering and a program committee member of the top conferences in her research areas. She has been co-chair of three editions of the Workshop on Personalized Access to Cultural Heritage (PATCH) and she has been a co-guest editor of the special issue "AI and HCI Methods and Techniques for Cultural Heritage Curation, Exploration and Fruition" in the Applied Sciences journal.

ZHONGLI FILIPPO HU received a B.Sc. degree in Computer Science with Honors from the University of Torino in 2018, and an M.Sc. degree in Computer Science with option Artificial Intelligence and IT Systems with Honors and Special Mention from the University of Torino in 2020. He is currently a Ph.D. student in Computer Science at the Computer Science Department of the University of Torino. His research interest includes recommender systems and information visualization. He is co-author of some research papers related to information filtering and recommender systems. During the M.S. degree, he worked for the thesis in a research group at Alpen-Adria-Universität Klagenfurt.

LILIANA ARDISSONO (Ph.D. in CS) is a Full Professor at the Computer Science Department of the University of Torino, Italy, where she teaches Object-Oriented and Web-based programming. Her research interests are User Modeling, Recommender Systems, and Information exploration support, with specific attention to geographic information search. On these topics, she has published more than 100 articles in international scientific conferences and journals. She is a member of the Editorial Board of the international journal User Modeling and User-Adapted Interaction (www.umuai.org/boards.html, Springer) and a member of the Advisory Board of User Modeling Inc.. She has been Co-Chair of several editions of the Workshop on Personalized Access to Cultural Heritage (PATCH). Moreover, she was Guest Editor of two special Issues of the User Modeling and User-Adaptive Interaction Journal, and she was Program Co-Chair of the International Conference on User Modeling 2005. She will be Program Co-Chair of UMAP 2022, the reference conference for user modeling and user-adaptive interaction.