Justification of recommender systems results: a service-based approach

Noemi Mauro[1] · Zhongli Filippo Hu[1] · Liliana Ardissono[1]

© The Author(s) 2022

Abstract

With the increasing demand for predictable and accountable Artificial Intelligence, the ability to explain or justify recommender systems results by specifying how items are suggested, or why they are relevant, has become a primary goal. However, current models do not explicitly represent the services and actors that the user might encounter during the overall interaction with an item, from its selection to its usage. Thus, they cannot assess their impact on the user's experience. To address this issue, we propose a novel justification approach that uses service models to (i) extract experience data from reviews concerning all the stages of interaction with items, at different granularity levels, and (ii) organize the justification of recommendations around those stages. In a user study, we compared our approach with baselines reflecting the state of the art in the justification of recommender systems results. The participants evaluated the *Perceived User Awareness Support* provided by our service-based justification models higher than the one offered by the baselines. Moreover, our models received higher *Interface Adequacy* and *Satisfaction* evaluations by users having different levels of Curiosity or low Need for Cognition (NfC). Differently, high NfC participants preferred a direct inspection of item reviews. These findings encourage the adoption of service models to justify recommender systems results but suggest the investigation of personalization strategies to suit diverse interaction needs.

Keywords Justification of recommender systems results · Service models · Service Blueprints

✉ Noemi Mauro
noemi.mauro@unito.it

Zhongli Filippo Hu
zhonglifilippo.hu@unito.it

Liliana Ardissono
liliana.ardissono@unito.it

[1] Computer Science Department, University of Torino, Corso Svizzera 185, 10149 Turin, Italy

Published online: 29 October 2022

🙎 Springer

1 Introduction

The demand for predictable and accountable Artificial Intelligence (AI) grows as tasks with higher sensitivity and social impact are more commonly entrusted to AI services (Mohseni et al. 2021) and important decisions are delegated to them (Springer and Whittaker 2019). Moreover, with the introduction of the European General Data Protection Regulation (European Commission 2018), which prescribes the user's "right to obtain meaningful information about the logic involved" (right to explanation), transparency has become a mandatory condition for all intelligent systems.

Recommender systems (Ricci et al. 2022) are the mainstream AI technique that supports information filtering. In several application domains, such as information exploration and e-commerce, they prevent overloading users with the plethora of available options to choose from. Thus, the ability to properly explain or justify the recommendations has become a primary goal (Di Noia et al. 2022). A large amount of work focuses on shading light on the internal system behavior to increase recommender systems transparency by explaining *how* they suggest items (Nunes and Jannach 2017; Tintarev and Masthoff 2012, 2022; Jannach et al. 2019). This aspect has been positively associated with the acceptance of results because it manifests the logic behind them (Cramer et al. 2008; Pu and Chen 2007; Tintarev and Masthoff 2022). Moreover, some justification models have been developed to face the challenges of black-box models, whose internal behavior is difficult to interpret, by specifying *why* the system provides certain results (Musto et al. 2021; Ni et al. 2019). However, both types of approaches are unaware of the service model that determines the interaction with an item, from its selection to its usage. Thus, they cannot take the overall consumer experience into account in the presentation of recommendations.

In service modeling research, Stickdorn et al. (2011) 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 the characteristics of the homes and includes getting in contact, or sharing spaces with their hosts. This implies different attitudes toward renting rooms or complete apartments (Lee 2022). Moreover, people can be exposed to amateur providers who might offer a low quality of service level (Yi et al. 2020). As exogenous risk factors can impact the overall interaction with items, recommender systems should explain, or justify, their own suggestions by providing users with a holistic view of items.

Review-based recommender systems (Chen et al. 2015; Hernández-Rubio et al. 2019) recognize the importance of consumer feedback to extract experience data about items but they overlook the structure of the underlying service. Therefore, they provide users with item-centric information that partially supports decision-making.

We investigate a service-based information presentation approach to make users aware of the overall experience they should expect when selecting items. In this context, we pursue the justification of recommender systems results because it is agnostic with respect to the applied algorithms but can be exploited to enhance users' awareness of the pros and cons of the items suggested by the recommender. Mauro et al. (2022) tested the recommendation performance of a few service-aware recommender systems that leverage consumer feedback to extract coarse-grained experience eval-

uation dimensions of items. Those dimensions guide (i) the rating estimation, (ii) a visual summarization of the sentiment emerging from the reviews, and (iii) the indexing of reviews by evaluation dimension. However, that work does not support the presentation of item aspects, nor the justification of recommendation results, which we investigate in the present paper. Specifically, here we advance that work in different ways. First, we use the Service Blueprints (Bitner et al. 2008) to define a more detailed service model that describes the stages of interaction with items. Second, we use that model to extract experience data from reviews concerning the stages of interaction with items at two granularity levels related to coarse-grained and fine-grained evaluation dimensions. Third, we use the service model to organize the item aspects in the justification of recommendations around these two types of evaluation dimensions and we associate aspects to the service stages that the user is expected to engage in. With respect to the explanatory aims that Tintarev and Masthoff (2012) identified, we are interested in evaluating the impact of service-based models on effectiveness and satisfaction. Specifically, we pose the following research questions:

- RQ1: *How does a service-based justification of recommendations impact the user's awareness about items and his/her confidence in evaluating them?*
- RQ2: *How does a service-based justification of recommendations impact the user's satisfaction with the presentation of information about items?*

We developed two service-based justification models that use coarse-grained and fine-grained evaluation dimensions of the experience with items to present the key aspects emerging from consumer feedback. These models organize the access to item aspects differently. However, both of them support an incremental information exploration to enable the inspection of data depending on diverse interests in the evaluation dimensions. We defined these dimensions by applying the Service Blueprint model (Bitner et al. 2008).

In a user study involving 59 participants, we compared our justification models with an approach similar to (Musto et al. 2021), an aspect-based comparison of items like (Chen et al. 2014), and a feature-based presentation of items and reviews inspired by standard e-commerce web sites. We implemented all these models in a test application that supported the user study. All the participants evaluated the *Perceived User Awareness Support* provided by our models higher than the one offered by the baselines. Moreover, the people having high or low values of the curiosity trait (Kashdan et al. 2009), and those having low Need for Cognition (NfC) (Coelho et al. 2020), evaluated the *Interface Adequacy* and *Satisfaction* of our models higher than the baselines. Differently, high NfC participants preferred the direct inspection of item reviews. These findings encourage the adoption of service models in the justification of recommender systems results but suggest to investigate personalization to suit diverse interaction styles.

In the following, we introduce the Service Blueprints and the related work (Sects. 2 and 3). Then, we describe the dataset we used and the justification models (Sects. 4 and 5). In Sects. 6, 7 and 8, we present some preliminary findings, the user study and its results. Section 9 discusses the results and Sect. 10 reports limitations and future work. Finally, we discuss the ethical issues and we conclude the article (Sects. 11 and 12).

2 Background on service blueprints

Service Blueprints (Bitner et al. 2008) are visual models that support the design and development of products and services by focusing on customers' viewpoint during the stages a person engages in, from the start point (e.g., enter website or shop) to the end one (customer care). They are largely used in service and product modeling and there are many examples, especially for e-commerce. Bitner et al. (2008) provides a sample blueprint that describes consumer experience in the hotel booking domain; we used that example to specify the home-renting process with Airbnb (2022), shown in Fig. 1:

- The *Physical evidence* includes the tangibles that the customer comes in contact with. For instance, in a home-booking service, this component represents the website of the home-booking platform, the check-in tangibles (e.g., the presence of key lock-boxes, keypad or smart locks), the services and the amenities concerning the rooms, the surroundings, and so forth.
- The *Customer actions* include the actions that the guest carries out during service fruition. For instance, the reservation, the arrival at the home, the activities related with personal care and home management.
- The *Onstage/visible contact employee actions* are the actions that service providers perform while they interact with the customer, such as the processing of the registration in the home at check-in time.
- Other layers represent backstage contact employee actions and support processes but we omit them because they are not relevant to the customer's direct experience.

We use the domain model built through Service Blueprints (i) to steer the analysis of item reviews by organizing feedback around the specified service stages, and (ii) to structure the presentation of item aspects according to the expected user experience during such stages.

3 Related work

3.1 Premises

The increasingly common deployment of intelligent systems in services has shown that "when decisions are taken or suggested by automated systems, it is essential for practical, social, and—with increasing frequency—legal reasons that an explanation can be provided to users, developers, and regulators" (Confalonieri et al. 2021). Early recommender systems were black-boxes focused on algorithmic performance. However, Herlocker et al. (2000) recognized the importance of explaining their results to enhance users' acceptance and trust. Since then, work has been done to improve the understanding of personalized recommendations (Nunes and Jannach 2017; Tintarev and Masthoff 2012, 2022; Jannach et al. 2019). The approaches developed so far can be classified in (i) those that explain the system's behavior (Conati et al. 2021; Kouki et al. 2020), (ii) those that fuse recommendation and explanation in the same process (Dong and Smyth 2017; Lu et al. 2018; Rana et al. 2022), and (iii) those that provide
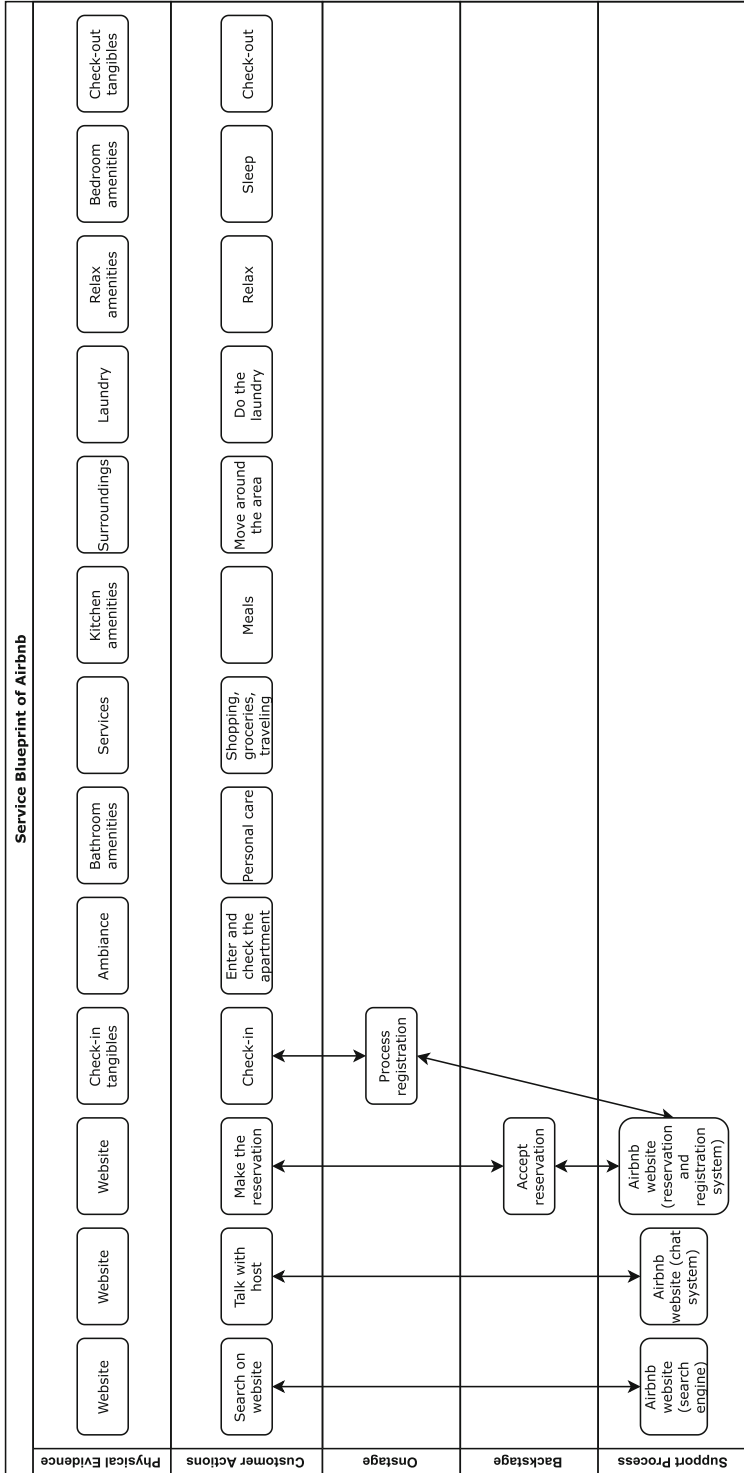
Fig. 1 Service Blueprint we defined to describe the home-booking domain (Airbnb)

post-hoc justifications of the suggestions (Musto et al. 2021; Ni et al. 2019). While the last approach is agnostic to the recommendation algorithm, the first two are tightly coupled to it. We aim at advancing the justification approach because it can be implemented on top of different recommendation algorithms to enhance users' awareness of the suggested items.

3.2 Explanation and justification techniques

Different techniques support the understanding of recommendations:

- The systems based on a single algorithm typically explain their results in terms of inference traces (Herlocker et al. 2000; Nunes and Jannach 2017). Some aspect-based recommender systems highlight the features of items which match, or mismatch, the user's preferences (Muhammad et al. 2016). Other ones group items on the basis of their pros and cons for the user, for easy comparison (Pu and Chen 2007; Chen et al. 2014; Chen and Wang 2017). Some information exploration support systems explain the suggestions by visualizing the relevance of items to the keywords of the submitted search queries (Chang et al. 2019; Di Sciascio et al. 2016). Graph-based recommenders use the connections between users and items as explanations of the suggestions (Amal et al. 2019; Wang et al. 2018; Musto et al. 2019). Some researchers assume that item recommendation and explanation should coincide because they rely on the same logic. Thus, they fuse the two processes (Dong and Smyth 2017; Lu et al. 2018), or use the strength of the possible explanations to steer the recommendation (Rana et al. 2022).
- Hybrid recommender systems explain the suggestions under multiple perspectives of relevance to represent the relative impact of the integrated systems on item rankings. For example, MyMovieFinder (Loepp et al. 2015) separately shows the recommenders that support a suggested item and RelevanceTuner (Tsai and Brusilovsky 2019) uses stackable bars to visually integrate them. TalkExplorer (Verbert et al. 2016) and IntersectionExplorer (Cardoso et al. 2019) show multiple dimensions of relevance through bidimensional graphs and grid layouts, respectively. Kouki et al. (2017) present the suggestions by means of Venn diagrams and Parra and Brusilovsky (2015) combine Venn diagrams with color bars to distinguish the contribution to item evaluation provided by the integrated recommenders.

All these models are unaware of the service underlying items. Thus, they present item-centric data that limits users' awareness of the experience they should expect from their own selections. For instance, the quality of the post-sales customer care depends on the retailer and introduces a further dimension of choice on top of the selection of the product that fits specific needs. However, a service-agnostic analysis of item reviews might hardly reveal this type of information. We use service blueprinting to characterize the stages of interaction with items and the involved entities. Then, we use the resulting service model to steer the extraction, organization and presentation of item aspects around coarse-grained and fine-grained experience evaluation dimensions that the user can consider in the analysis of items.

Commercial platforms like Airbnb (2022) and TripAdvisor (2017) present a summary of consumer feedback that reports the overall evaluation expressed by previous

customers, a set of values associated with the main aspects of the items (e.g., cleanliness and communication for Airbnb, or cuisine and atmosphere for TripAdvisor), and the reviews. However, they do not show the connections between aspects values and reviews, leaving the user the burden of summarizing the opinions that emerge from them.

Chen et al. (2014) extract item aspects from reviews and present them in quantitative and qualitative format. However, they fail to explicitly model the stages of interaction with the service behind items. Therefore, they present a flat overview of item-centric aspects that users have to interpret in terms of experience evaluation dimensions.

Some researchers discovered that users' perceptions of the explanations generated by recommender systems depend on their cognitive style (Millecamp et al. 2019, 2020), personality (Kouki et al. 2019; Millecamp et al. 2020), and domain expertise (Kouki et al. 2019). Moreover, Millecamp et al. (2022) suggested to tailor the explanations to users' personal characteristics. The adaptation of the justification model to the individual user is an interesting future development and is supported by our experimental results, where we noticed that users' curiosity trait and Need for Cognition impact the perception of the justification models. See Sect. 8.

4 Data

For this study, we used a public dataset of Airbnb reviews concerning the homes of London city. We downloaded this dataset in January 2021 from http://insideairbnb.com/get-the-data.html.

The dataset contains information about homes (denoted as "listings"), their administrators ("hosts") and features ("amenities"). Moreover, it includes the reviews uploaded by the people who rented the homes ("guests"), starting from December 21st, 2009. We noticed that, in January 2020, people rented very few homes, probably because of the COVID-19 pandemic. We thus decided to filter out the reviews uploaded after the first day of that month. Then, we selected the reviews written in English and we removed the listings that did not receive any comments since 2018 to work on recently rented homes.

The filtered dataset contains 764,958 guests, 906,967 reviews and 43,604 listings, out of which we selected the homes used in our user study. Table 1 reports the descriptive statistics of that dataset.

An analysis of the reviews shows that, while some of them are extremely concise (e.g., *"Amazing location!"*), other ones are fairly articulated. Reviews typically discuss features and aspects of homes such as the offered services and amenities or

Table 1 Descriptive statistics of the filtered dataset

	Min	Max	Mean	SD
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

their cleanliness. However, they frequently also include evaluations of the hosts and surroundings, providing a broad picture of guests' renting experience. For example:

"A warm and private place ideal for exploring London. Location was perfect and felt very safe. We stayed with our young children and they had space to stretch out with their toys, the lift was convenient and check-in was a breeze! Very clean and comfortable, we would stay here again!"

5 Methodology

The methodology we applied to develop our service-based justification models is general but we present it by referring to the home booking domain. We first describe how we defined the evaluation dimensions of experience with items that support the organization of aspects to justify recommendations. Then, we outline the extraction and analysis of aspects from item reviews. Finally, we present our justification models and the baselines for comparison.

Notice that the extraction and analysis of aspects is an offline task. It should be performed once in the dataset of reviews and possibly periodically updated to take new entries into account.

5.1 Evaluation dimensions of the experience with items

The first step to identify the evaluation dimensions of experience is the definition of a Service Blueprint that represents users' experience with items. Figure 1 shows the one we developed by extending Bitner et al. (2008)'s hotel booking one with other representations of customers experience in hotel booking (Ren et al. 2016) and a detailed analysis of Airbnb customer's needs and preferences (Cheng and Jin 2019).

As we are interested in building a justification model for recommendations, rather than designing the complete home-booking service, we focus on the Customer Actions and Physical Evidence layers of the blueprint. These layers describe the typical sequence of steps the user can carry out and the tangibles (s)he can encounter. As specified in Fig. 1, the first tangible is the Airbnb web site that supports the interaction with the host of the home and the reservation. When reaching the home, the guest checks in and this activity might involve meeting the host (or a referent that we consider the host) to receive the keys. During the stay, the guest might be involved in various activities, such as personal care, shopping, managing meals, moving around in the neighborhood, doing the laundry and relaxing or sleeping. The last step is the check-out that, again, might involve the host.

The Physical Evidence layer does not model human actors but the activities specified in the Customer Actions layer might concern different entities, including the host. As both tangibles and human actors can impact the guest's satisfaction, we add a layer to map Customer Actions to evaluation dimensions that represent the experience with the involved entities. To support both the summarization and a detailed organization of information about items, we define two types of experience evaluation dimensions (see Table 2):

Table 2 Coarse-grained and fine-grained evaluation dimensions for home-booking

Coarse-grained dimensions	Fine-grained dimensions	Physical evidence
Host appreciation	Host	–
Search on website	Website	Website
Check-in/check-out	Check-in	Check-in tangibles
	Check-out	Check-out tangibles
In apartment experience	Ambiance	Ambiance
	Bathroom	Bathroom amenities
	Kitchen	Kitchen amenities
	Laundry	Laundry
	Relax	Relax amenities
	Bedroom	Bedroom amenities
Surroundings	Surroundings services	Surroundings services

- The **fine-grained evaluation dimensions** describe consumer experience with the tangibles and actors involved in the individual customer actions. In our domain, they are the perception of the host, website of the home-booking platform, ambiance of the home, rooms, and so forth. Fine-grained dimensions also include check-in and check-out to represent the experience during those activities. For example, a guest might have a bad experience because the host shows up late at check-in.

- The **coarse-grained evaluation dimensions** summarize consumer experience by abstracting from individual customer actions. A coarse-grained dimension is mapped to multiple fine-grained ones. For instance, a generic "In apartment experience" dimension can summarize the guest's experience within a home (ambiance, rooms, etc.). Similarly, check-in and check-out can be combined into a single coarse-grained dimension.

Table 2 shows the evaluation dimensions we defined starting from the Service Blueprint of Fig. 1, mapped to the elements of the Physical Evidence layer.

5.2 Extraction and organization of item aspects

We apply an extension of the approach described in (Mauro et al. 2021a) to extract the aspects of items emerging from the reviews. Specifically, we use the distinction between coarse-grained and fine-grained evaluation dimensions of experience to classify aspects at different granularity levels. In the following, we shortly describe how we extract and organize the aspects of an item i (an individual home) starting from its reviews REV_i:

1. For each *aspect-adjective* pair[1] occurring REV_i, we produce an $< aspect, asp\#rev, adjective, asp\_adj\#rev, evaluation >$ tuple where:

   - $asp\#rev$ denotes the number of reviews $r \in REV_i$ that mention *aspect*;
   - $asp\_adj\#rev$ denotes the number of reviews $r \in REV_i$ that mention the *aspect-adjective* pair;
   - *evaluation* is the normalization in [1, 5] of the polarity of the *aspect-adjective* pair. We compute the polarity as the mean value returned by the TextBlob (Loria 2020) and Vader (Hutto and Eric 2014) libraries.

   These tuples summarize previous consumers' opinions about the item in a structured way by measuring how many guests mention them, and with which degree of appreciation. This differs from counting the frequency of terms and is robust to the occurrence of long reviews that repeat the same concepts several times.

2. We classify the aspects extracted from REV_i by fine-grained experience evaluation dimension using entity recognition (to identify references to people and places) and a set of dictionaries that collect the terms frequently used to refer to our coarse-grained dimensions. The dictionaries derive from the thesauri of (Mauro et al. 2020), which we split into subsets, each one including the terms semantically related to an individual fine-grained dimension. For example, the `Kitchen` dictionary contains "kitchen", "oven", "table" and other similar keywords. By classifying aspects, we can associate them to the stages of interaction with items by taking the fine-grained and coarse-grained evaluation dimensions into account. This is the basis to organize information at different granularity levels, and to quantitatively summarize the related consumer feedback.

3. We compute the value of each coarse-grained evaluation dimension d as the weighted mean of the evaluations received by the *aspect-adjective* pairs such that *aspect* is classified in D. For each pair, we use the number of reviews that mention it ($asp\_adj\#rev$) as weight to tune its impact on the evaluation of D coherently with the number of people who mention it. If there is no information about a dimension, its value is set to 0 that, being out of the [1, 5] range, means "zero knowledge".

Table 3 shows the type of information that this analysis produces and aggregates data by fine-grained evaluation dimension. Notice that, while we carry out the analysis, we index the sentences of the reviews by *aspect-adjective* pair to support their retrieval for justification purposes.

5.3 Service-based justification models

Our models summarize the main features and properties of items in the user interface and make additional data available on demand. Thus, users are free to expand the information they care about. Figure 2 shows a portion of the user interface of the

---

[1] The extraction of aspects and adjectives can be done by applying standard NLP techniques. In our work, we exploit lemmatization, dependency parsing and the Double Propagation algorithm (Qiu et al. 2011) that is suitable to deal with datasets that have not been previously annotated because it is unsupervised.

Table 3 Sample aspects extracted from the reviews of a sample Airbnb home

| Aspect	asp#rev	Adjective	asp_adj#rev	Evaluation	Dimension
Location	23	Great	6	4.42	Ambiance
Location	23	Excellent	2	4.57	Ambiance
Location	23	Good	2	4.14	Ambiance
Location	23	Convenient	1	3.00	Ambiance
Host	22	Great	7	4.42	Host-prop
Host	22	Friendly	4	3.87	Host-prop
Host	22	Excellent	2	4.57	Host-prop
Host	22	Lovely	2	4.09	Host-prop
Place	9	Lovely	3	4.09	Ambiance
Place	9	Great	2	4.42	Ambiance
Place	9	Airy	1	3.00	Ambiance
Bed	4	Comfortable	2	3.91	Bedroom
Bed	4	Superb	1	4.62	Bedroom
Restaurant	4	Cool	1	3.67	Surroundings
Restaurant	4	Lovely	1	4.09	Surroundings
Restaurant	4	Nice	1	4.02	Surroundings
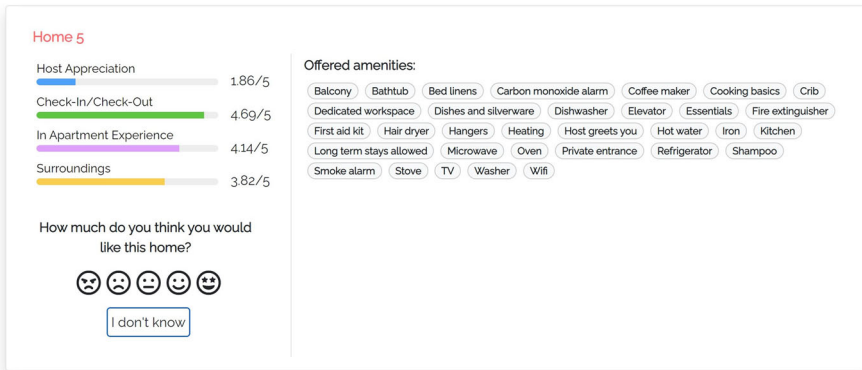
Fig. 2 Portion of the user interface shared by the proposed justification models

justification models we propose (Appendix includes a larger version of this figure and of the following ones). We focus on the component presenting individual items:

- The central area shows the features of the visualized item. In our case, these are the amenities offered by the home, which typically represent binary features and thus can be shown or hidden, depending on their value. As suggested by Tintarev and Masthoff (2022), we omit data such as the price, number of rooms, pictures and name because they might influence users' evaluation of the home or their perception of the data provided by the system. For instance, some homes have

long names that mention their location or the view; e.g., *"Beautiful Flat - London Bridge. SE1"*, or *"Clerkenwell penthouse, huge terrace"*.

- The left area shows a set of colored bar graphs that summarize previous consumer experience with the item. Each bar corresponds to a coarse-grained experience evaluation dimension D, that is, `Host Appreciation`, `Check-in/Check-Out`, `In Apartment Experience` or `Surroundings` (we omit `Search on website` because we are not interested in the user's experience with the Airbnb platform). The bar shows the value of D, which represents the evaluation that the item has received in step 3 of Sect. 5.2. If $D = 0$, the name of the bar graph is displayed in light gray to denote that there is no feedback about it and distinguish the "zero knowledge" situation from an extremely bad evaluation. The user can click on the bar graphs to receive more details about previous consumer experience with the item; see models M- THUMBS and M- ASPECTS.

- The left area also includes a rating component through which the user can evaluate the item in the [1, 5] scale, represented as a list of smilies. In this work, we do not evaluate recommendation performance but we included this component for two reasons: Firstly, we wanted to attract the user's attention to the presented data. Second, we aimed to collect some implicit feedback about her or his confidence in the evaluation of the items. The rating component includes an "I don't know" button to skip the evaluation.

We now describe the peculiarities of our justification models, which provide different information when the user clicks the bar graphs.

5.3.1 M-THUMBS

In this model (Fig. 3), when the user clicks the bar of a coarse-grained dimension D, the user interface opens the "What travelers are saying" component that shows the fine-grained dimensions $d \in D$. For example, "AMBIANCE" of "In Apartment
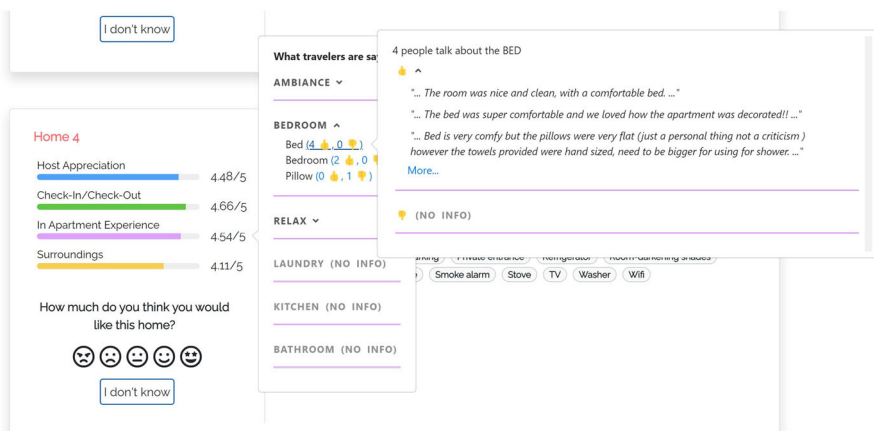

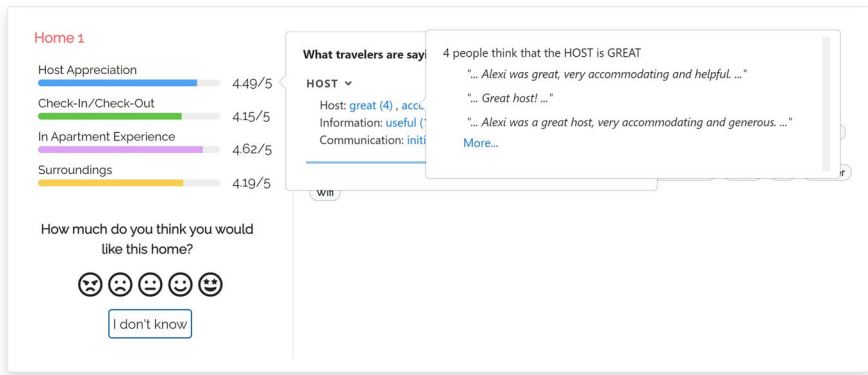**Fig. 3** User interface of justification model M- THUMBS

Fig. 4 User interface of justification model M- ASPECTS

Experience". Within the component, dimensions are sorted by the number of aspects mentioned in the reviews of the item. A dimension d can be expanded to view the most relevant aspects classified in it, sorted by decreasing relevance. For clarity, we show the dimensions that have no aspects in light gray, with a ("NO INFO") tag, and they cannot be expanded.

Here and in the following models, we compute the relevance of aspects as the number of reviews of the item that mention them; see the $asp\#rev$ field in Table 3. When the user expands a fine-grained dimension, the user interface shows a maximum of three aspects and a button to view the complete list. For each aspect, a thumb up/down reports the number of reviews expressing a positive/negative opinion about it, derived from the data of Table 3. Thumbs and numbers can be clicked to view the quotes from the reviews mentioning the aspect.

5.3.2 M-ASPECTS

In this model (Fig. 4), the "What travelers are saying" component is organized as in M- THUMBS. However, for each fine-grained dimension, aspects report the list of the most relevant adjectives referring to them in the reviews of the item. The relevance of an adjective corresponds to the $asp\_adj\#rev$ field of its $aspect\text{-}adjective$ pair in Table 3. For each adjective, the user interface reports that value and, by clicking on the term, or on the associated number, the user can view the related quotes from the reviews.

5.4 Baseline justification models

5.4.1 M-SUMMARY

In this model (Fig. 5), below the item features, the user interface shows a summary that describes the most relevant aspects ($asp\#rev$) and adjectives ($asp\_adj\#rev$) of the item extracted from its reviews.
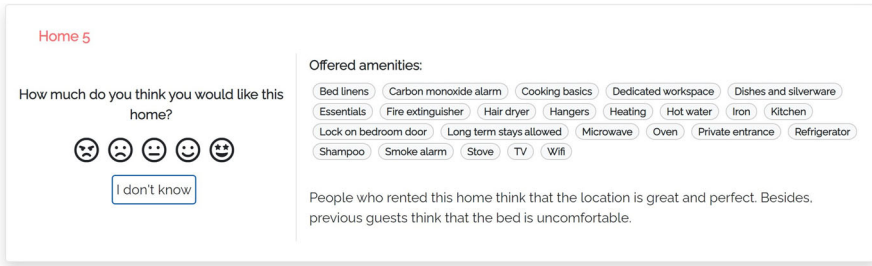
Fig. 5 User interface of justification model M- SUMMARY

Similar to (Musto et al. 2021), we dynamically compose the textual paragraphs by exploiting a Backus–Naur Form (BNF) grammar that generates different types of sentences to support variability in the summaries. Moreover, we select the aspects to be included by decreasing relevance, and similar for the adjectives to be mentioned. However, we compute the relevance of aspects and adjectives in terms of how many reviews mention them ($asp\#rev$ and $asp\_adj\#rev$) rather than through the Kullback–Leibler divergence (KL). The reason is that KL uses a dictionary that does not suit our needs because it misses (and thus is unable to evaluate) several words that guests frequently use to express their opinions about homes.

5.4.2 M-OPINIONS

In this model (Fig. 6), alongside the item features, the user interface shows the evaluations of the most relevant aspects of the item extracted from its reviews ($asp\#rev$), sorted by decreasing relevance. Each aspect evaluation is visualized as a gray bar coupled with a numeric value in [1, 5]. By clicking the bar, the user can view a component that shows the number of guests who mentioned the aspect and the adjectives they used to qualify it.
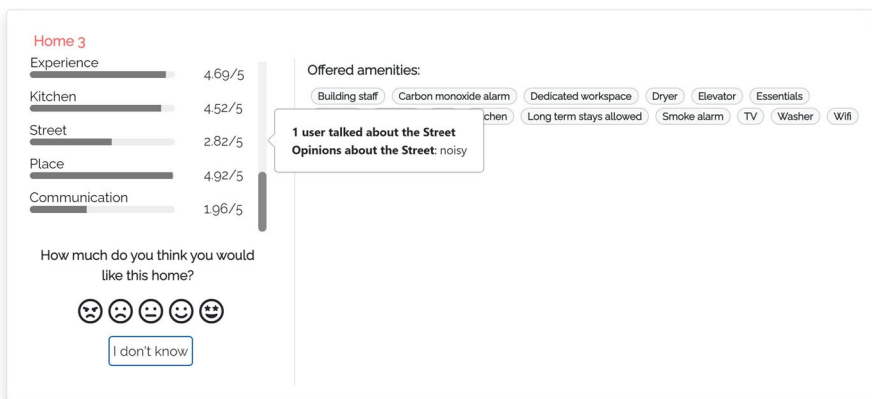

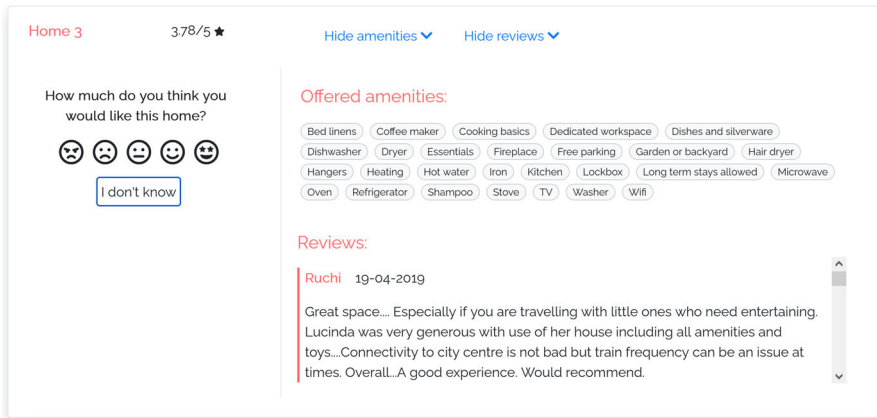**Fig. 6** User interface of justification model M- OPINIONS

Fig. 7 User interface of justification model M- REVIEWS

The value of a bar is the weighted mean of the evaluations of the related *aspect-adjective* pairs (*evaluation* in Table 3); the weight is their *asp_adj#rev*.

Overall, the user interface provides similar information about item aspects as Chen et al. (2014)'s opinion bar chart. However, we use simple bar charts to make the visualization comparable with our service-based justification models.

5.4.3 M-REVIEWS

In this model (Fig. 7), the user interface shows the typical data provided by e-commerce platforms, and in particular by home-booking services; i.e., item features, mean rating received from consumers, and reviews. The model enables the user to hide or show the features and the reviews by clicking on the tabs available at the top of the page.

6 Preliminary experiment

We conducted a preliminary experiment to test a first mock-up of service-based justification model described in (Mauro et al. 2021b); see Fig. 8. In the study we wanted to investigate the impact on decision making of the visualization of quantitative data that describes coarse-grained experience evaluation dimensions (bar graphs), combined with on-demand qualitative data concerning such dimensions (aspects manually crafted from item reviews). The model did not include the fine-grained evaluation dimensions of experience. In the following, we briefly describe this experiment.

11 participants, aged between 19 and 57, and having diverse background and familiarity with technology, joined in the study. They performed two tasks, each one requiring to evaluate 5 Airbnb homes presented by the mock-up:

- In the first task (T1), for each presented home the user could only view the bar graphs of the coarse-grained evaluation dimensions.
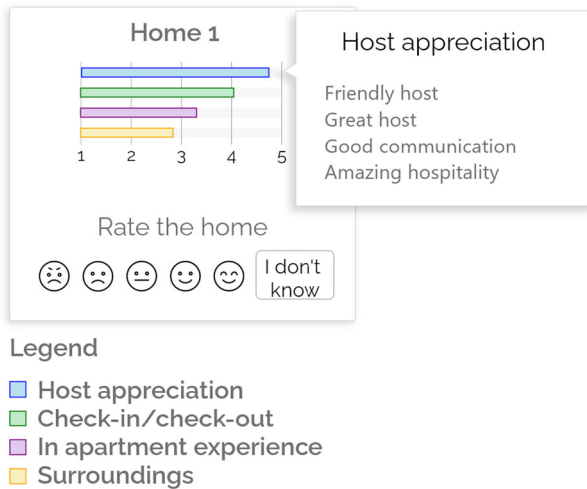
Fig. 8 Portion of the mock-up user interface used in the preliminary experiment

- In the second one (T2), the user could also inspect the aspects and adjectives concerning the coarse-grained dimensions by clicking the respective bars.

In the post-task questionnaire of T1, 54.54% of people declared that the given information was not enough to rate the homes. In T2, 27.27% said that they would have preferred to receive more data about the homes while the other ones evaluated them. This suggests that bar graphs alone are not sufficient for decision making because they do not enable the user to access any qualitative data to justify the values of the experience evaluation dimensions. When showing qualitative data the situation improves. However, more information about items is needed to help users make their selections.

The participants' comments provided insights about how they used the data provided by the mock-up. We report the two most interesting ones:

- In a real-word application, the bar graph is useful to filter out a home that does not deserve to be analyzed because it performs badly on an evaluation dimension that the user particularly cares about. For instance, a participant declared to have discarded a home having low Host appreciation to avoid interacting with difficult people.
- Some people said that the qualitative data about evaluation dimensions (i.e., the aspects presented on-demand) are useful to implicitly "tune" the values of the bars. For instance, suppose that a home h receives a low evaluation regarding a dimension d, and that the justification of d depends on aspects that are irrelevant to the user. Then, he or she might implicitly increase the evaluation of h.

In summary, our preliminary experiment suggested that the bar graphs describing coarse-grained experience evaluation dimensions present relevant quantitative data about items but have to be coupled with qualitative data to help users make their decisions. This finding drove us to the development of the novel justification models proposed in the present paper.

7 Study design

We carried out a user study to evaluate users' experience with the presented justification models and how the service-based ones perform with respect to Tintarev and Masthoff's explanatory aims of effectiveness (help users make good decisions), and satisfaction (increase the ease of use or enjoyment).

As observed by Tsai and Brusilovsky (2021), when users are free to explore a recommendation list, they tend to inspect both items placed at the top of the list and lower ranked ones. This means that the system should enable them to obtain an accurate impression of all the presented items, regardless of their suitability. To be sure that our models offer this type of support, similar to (Tintarev and Masthoff 2022), in the study we showed both high and low-quality homes.

7.1 Context

We recruited adult participants by spreading an invitation message in public mailing lists and social networks. In that message, we specified that we had a preference for people who had previously used an online booking or e-commerce platform. Users joined the experiment voluntarily, without any compensation.

We carried out the user study by exploiting an interactive web-based test application linked in the invitation message. The application guided participants in all the steps of the study. To guarantee users' privacy, the application did not collect their names or any other identifying data. At the beginning of the interaction with the user, it generated a numerical identifier to tag the anonymous data it acquired during the interaction session.

We used a power analysis to figure out how many people we needed to obtain statistically significant experimental results. The following four parameters are used for this analysis: *Alpha* ($\alpha = 0.05$): 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 = 0.35:* 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 $N = 55$ that supports the actual statistical power of 80%. We thus planned to collect at least 55 user tests.

7.2 Method

In the study, we applied a within-subjects approach. We managed each treatment condition as an independent variable and each participant received all the treatments. The test application counterbalanced the order of tasks to reduce the effects of practice and fatigue, as well as the impacts of result biases. We did not impose any time limits to the execution of the test to let people free to explore the information provided by the application. Overall, the user study was organized as follows:

1. First, the test application prompted users to read the informed consent[2] and declare that they were 18 years old or over. Moreover, it asked them to give their explicit agreement to participate in the study. Only those who positively answered both questions could take part in the experiment.

2. Next, the application asked users to fill in a questionnaire with their demographic information, cultural background, familiarity with booking and e-commerce platforms. The questionnaire is based on the ResQue recommender system questionnaire (Pu et al. 2011) and it is useful to understand the basic characteristics of the sample of users. The application also collected a set of constructs of Personal Characteristics (PC). We defined two constructs regarding the Trust in Booking Systems and the General Trust in Technology (Tsai and Brusilovsky 2021); see Table 4. Participants answered the statements in the {Strongly disagree, Disagree, Neither agree nor disagree, Agree, Strongly agree} scale, which we mapped to [1, 5]. The last column of the table shows the mean values of the participants' answers.

3. Then, the application applied (in counterbalanced order) the justification models described in Sects. 5.3 and 5.4 to interact with users. For each model, it asked participants to explore and rank five homes; then, it asked them to declare their level of agreement with the statements of the post-task questionnaire shown in Table 5, which measures the user experience with the justification models. The statements are taken from (Pu et al. 2011; Di Sciascio et al. 2019; Lewis and Sauro 2009) and capture users' perceptions of the user interface provided by each model. Participants answered in the {Strongly disagree, …, Strongly agree} scale, mapped to [1, 5]. Table 5 groups statements in three user experience constructs: *Perceived User Awareness Support*, *Interface Adequacy* and *Satisfaction* that we use to gain a deeper view of the user experience with the justification models through Structural Equation Model analysis; see Sect. 8.4.

During the test, the application also administered the statements of the Curiosity and Exploration Inventory-II (CEI-II) questionnaire (Kashdan et al. 2009) and of the Need for Cognition one (Coelho et al. 2020). CEI-II allows to understand participants' motivation to seek out knowledge and new experiences (Stretching) and their willingness to embrace the novel, uncertain, and unpredictable nature of everyday life (Embracing). Need for Cognition investigates people's tendency to engage in and enjoy thinking.

8 Results

66 people joined in the user study from November 15 to December 15, 2021 but we excluded 7 of them because they did not pass the attention checks. On average, the experiment lasted 35.45 min.

---

[2] The consent text can be found at this link: https://bit.ly/3LypcZp.

Table 4 Questionnaire about personal characteristics and mean values of the answers

Construct	Factor	Statement	M(SD)
Trust in Booking Systems (PC1)	1	I tend to trust the suggestions generated by booking systems	3.10 (0.82)
	2	I think that the ratings given by other users are enough to book homes	3.19 (0.86)
	3	I need to inspect the reviews given by other users to book homes	4.12 (0.74)
	4	I need to inspect the description of the home to book it	4.31 (0.70)
Trust in Technology (PC2)	1	I feel technology never works	1.66 (0.58)
	2	I'm less confident in doing things when I use supporting technology	1.80 (0.94)
	3	The usefulness of technology is highly overrated	1.92 (0.86)
	4	I tend to trust a person/thing, even though I have little knowledge of it	2.83 (0.89)

Table 5 Post-task questionnaire

Construct	Factor	Statement
Perceived User Awareness Support (U)	U1	The information provided was sufficient for me to understand what previous users think about the homes
	U3	The information about the homes was easy to interpret and understand
	U4	I quickly found the information about the homes
Interface Adequacy (I)	I1	It was easy to understand why some homes were good and others not
	I2	I found the user interface very intuitive
	I3	The user interface was sufficiently informative
Satisfaction (S)	S1	I think that I would like to frequently use this system to evaluate homes
	S3	I thought this system to evaluate homes was easy to use
	S4	I felt very confident using this system to evaluate homes

Statements are grouped by user experience construct

8.1 User data

- *Gender and age.* The 59 participants we considered for the test included 24 females, 33 males, 0 not-binary, 2 not declared, with the following age distribution: ≤ 20 (2 people), 21–30 (43), 31–40 (7), 41–50 (2), 51–60 (4) and > 60 (1 person).
- *Education level and background.* 13 subjects declared that they had attended the high school, 40 the university, and 6 stated that they had a PhD. 17 participants specified that they had a technical background, 31 a scientific one, 6 humanities and languages, 2 economics, and 4 other background. Furthermore, 46 people classified themselves as advanced computer users, 10 average ones, and 3 beginners.
- *Familiarity with online booking or e-commerce platforms.* 18 people declared that they used those platforms a few times in a week, 26 a few times in a month, 14 a few times in a year and 1 person never used one before.
- *Trust in Booking Systems (PC1).* As shown in the last column of Table 4, participants moderately agreed with trusting the suggestions generated by booking systems (statement 1: $M = 3.10$, SD = 0.82). They concurred that, to book a home, they needed to inspect its reviews (statement 3: $M = 4.12$, SD = 0.74) and description (statement 4: $M = 4.31$, SD = 0.70). However, they only partially agreed that the ratings given by other users are enough to book homes (statement 2: $M = 3.19$, SD = 0.86).
- *Trust in Technology (PC2).* Participants positively evaluated technology. In fact, statements 1–3, which deny the trust in technology, have low mean values. Differently, participants tended to be suspicious towards people and things that they do not know (statement 4: $M = 2.83$, SD = 0.89).

8.2 Analysis of participants' experience with the justification models

Table 6 shows the results of the post-task questionnaire of Table 5. According to a Kruskall–Wallis test, the differences between user experience constructs across the five models are statistically significant:

- *Perceived User Awareness Support* $[H = 13.40, df = 4, p < 0.008]$;
- *Interface Adequacy* $[H = 10.21, df = 4, p < 0.035]$;
- *Satisfaction* $[H = 8.07, df = 4, p < 0.084]$.

Moreover, a *post-hoc* comparison based on a Mann-Whitney test showed that:

- *Perceived User Awareness Support.* M- THUMBS ($M = 3.66$, SD=1.05) is the best justification model. The difference compared to M- SUMMARY, M- OPINIONS and M- REVIEWS is statistically significant. Specifically, in M- THUMBS, participants perceived the information about the homes as the easiest to interpret and understand (statement U3). Moreover, M- THUMBS best supported them in quickly finding the data about homes (U4). On the other hand, M- REVIEWS received the lowest evaluation regarding both the easy of interpretation/understanding of information about items, and the speed in finding it. As all the justification models present the amenities offered by the homes in the same way, we think that M- REVIEWS challenged the participants because it does not summarize the reviews. Conversely,

Table 6 Post-task questionnaire results describing participants' experience with the justification models

Construct	Factor	Justification model				
		M- THUMBS M(SD)	M- ASPECTS M(SD)	M- SUMMARY M(SD)	M- OPINIONS M(SD)	M- REVIEWS M(SD)
Perceived User Awareness Support	U1	3.61 (0.95)	3.44 (1.10)	2.59 (1.12)	3.08 (1.16)	**3.73 (1.08)**
	U3	**3.58 (1.04)**	3.42 (1.04)	3.56 (1.10)	3.07 (1.19)	3.07 (1.22)
	U4	**3.80(1.16)**	3.53 (1.09)	3.63 (1.07)	3.20 (1.17)	2.80 (1.28)
	Average	**3.66 (1.05)**	3.46 (1.07)	3.26 (1.19)**	3.12 (1.17)***	3.20 (1.25)**
Interface Adequacy	I1	**3.46 (1.06)**	3.34 (1.14)	2.98 (1.18)	3.12 (1.13)	2.93 (1.26)
	I2	3.46 (1.25)	3.36 (1.06)	**3.81 (1.01)**	3.44 (1.10)	3.53 (1.02)
	I3	**3.64 (0.98)**	3.47 (1.01)	2.41 (1.02)	3.14 (1.14)	3.39 (0.89)
	Average	**3.52 (1.10)**	3.39 (1.07)	3.07 (1.21)***	3.23 (1.13)*	3.28 (1.09)
Satisfaction	S1	**3.29 (1.22)**	3.10 (1.18)	2.58 (1.10)	2.86 (1.17)	3.00 (1.08)
	S3	3.61 (0.98)	3.41 (0.89)	**3.85 (0.96)**	3.31 (1.00)	3.51 (1.02)
	S4	**3.51 (1.14)**	3.37 (0.91)	2.90 (1.21)	3.05 (1.09)	3.27 (1.05)
	Average	**3.47 (1.12)**	3.29 (1.01)	3.11 (1.22)**	3.07 (1.10)**	3.26 (1.07)

Results are grouped by user experience construct. For each construct, three rows show the values obtained for the questions of Table 5 (factors). The "Average" row reports the mean value of the factors. The highest values are in boldface. Stars denote the statistical significance of the difference between the best-performing model and the other ones. Significance levels: (***)$p < 0.01$, (**)$p < 0.05$, (*)$p < 0.1$

M- REVIEWS best supported the understanding of previous guests's opinions about the homes (U1) because it presents the full set of reviews received by the visualized items. However, M- THUMBS is the second best model concerning this evaluation aspect.

- *Interface Adequacy*. Participants perceived M- THUMBS as the best justification model (M=3.52, SD=1.10); the differences with respect to M- SUMMARY and M- OPINIONS are statistically relevant. People felt that M- THUMBS helps the user understand why some homes are good or bad and is sufficiently informative. However, they perceived M- SUMMARY as the most intuitive user interface, probably because it summarizes information in a simple text.

- *Satisfaction*. Participants perceived M- THUMBS ($M = 3.47$, SD=1.12) as the best model and the differences with respect to M- SUMMARY and M- OPINIONS are statistically relevant. Users declared that they would like to frequently use M- THUMBS to evaluate homes. Moreover, they felt more confident using M- THUMBS than the other justification models. However, they perceived M- SUMMARY as the easiest model to use. Also in this case, the reason might be the bare user interface of this model, which presents previous guests' experience with items using few sentences. Anyway, M- THUMBS is the second best model concerning the easy of use.

8.3 Opting outs

In the user study, the 59 participants evaluated 25 homes, thus providing 295 item ratings in total. Below, we report the number and rate of opting outs, which correspond to "I don't know" selections and thus denote a lack of confidence in the evaluations:

- M- THUMBS: 10 opting outs (3.39 %);
- M- ASPECTS: 15 (5.08 %);
- M- SUMMARY: 30 (10.17 %);
- M- OPINIONS: 33 (11.19 %);
- M- REVIEWS: 32 (10.85 %).

These results are in line with the fact that M- THUMBS is the best justification model from the viewpoint of the *Perceived User Awareness Support*. When using M- SUMMARY and M- REVIEWS, a definitely higher number of participants did not feel like evaluating the homes. This can be explained with the fact that (i) in spite of its simplicity, M- SUMMARY poorly describes previous guests' opinions about the homes, and (ii) M- REVIEWS forces the user to read the reviews to evaluate them. Differently, M- THUMBS shows a summary of consumer feedback but it also supports the on demand retrieval of detailed information about the homes.

8.4 Structured Equation Model analysis

We performed Knijnenburg and Willemsen (2015)'s Structured Equation Model analysis to deeply understand the user experience with the justification models. This analysis is useful to find the relationship between unobserved constructs (latent variables) by leveraging observable variables.

Based on the post-task questionnaire of Table 5, we associated two constructs (*Perceived User Awareness Support* and *Interface Adequacy*) to the Subjective Systems Aspects (SSA) and one construct (*Satisfaction*) to User Experience Aspects (EXP). We tested five justification models (Objective System Aspects) that we represented as dummy variables: M- THUMBS, M- ASPECTS, M- SUMMARY, M- REVIEWS and M-OPINIONS. Moreover, we selected the following constructs to represent the Personal Characteristics (PC): *Trust in Booking Systems*, *Trust in Technology*, *Curiosity and Exploration Inventory* and *Need for Cognition*. As these constructs include at least three statements each, they are good candidates for a Structured Equation Model.

We performed the Confirmatory Factor Analysis to examine the validity of the constructs. First, we computed the convergent validity to check that their statements are related. For this purpose, we examined the Average Variance Extracted (AVE) of each construct that must be over 0.50 to keep the validity. Then, we computed the discriminant validity to check if the statements of different constructs are too highly correlated. To obtain the discriminant validity, the largest correlation value must be less than the squared root of the AVE value of both factors. All the constructs respected the constraints:

- *Perceived User Awareness Support*: $AVE = 0.64$, $\sqrt{AVE(0.64)} = 0.80$, largest correlation = 0.59;
- *Interface Adequacy*: $AVE = 0.47$, $\sqrt{AVE(0.47)} = 0.69$, largest correlation = 0.68;
- *Satisfaction*: $AVE = 0.62$, $\sqrt{AVE(0.62)} = 0.79$, largest correlation = 0.70.

Figure 9 shows the Structural Equation Model we obtained. The graph reports the dependencies, β-coefficients and standard error values that indicate the correlations between the constructs. We removed *Trust in Technology* and *Need for Cognition* from the Personal Characteristics because during the iterations they had high p values.

All the models have a positive effect on the *Perceived User Awareness Support* except M- OPINIONS that has a neutral correlation. Specifically, M- THUMBS has the strongest positive correlation ($+ 0.285$; $p > 0.001$) followed by M- ASPECTS ($+ 0.151$; $p > 0.05$). This is in line with the results of the post-task questionnaire, in which participants perceived M- THUMBS and M- ASPECTS as the best models from this viewpoint. Moreover, since there is a positive correlation between the *Perceived User Awareness Support* and the *Satisfaction* construct ($+ 1.151$; $p < 0.001$), we infer that M- THUMBS also satisfies users in the analysis of previous guests' feedback when they explore the homes.

All the models negatively influence the *Interface Adequacy*, probably because their user interfaces require some effort to understand the representation of consumer feedback. However, M- THUMBS has the less negative value, i.e., it is better than the other models. Notice that the *Interface Adequacy* and *Satisfaction* constructs are positively correlated ($+ 1.204$; $p < 0.001$). From this information we infer that the models that best correlate with the *Interface Adequacy* also lead to higher satisfaction. See Table 6.

We notice a positive correlation between the *Curiosity and Exploration Inventory* construct and the *Interface Adequacy* one with lower statistical significance ($+ 0.332$; $p < 0.05$). Finally, there is a positive correlation between the *Trust in Booking Systems* construct and the *Interface Adequacy* one ($+ 0.251$; $p < 0.001$). This finding suggests
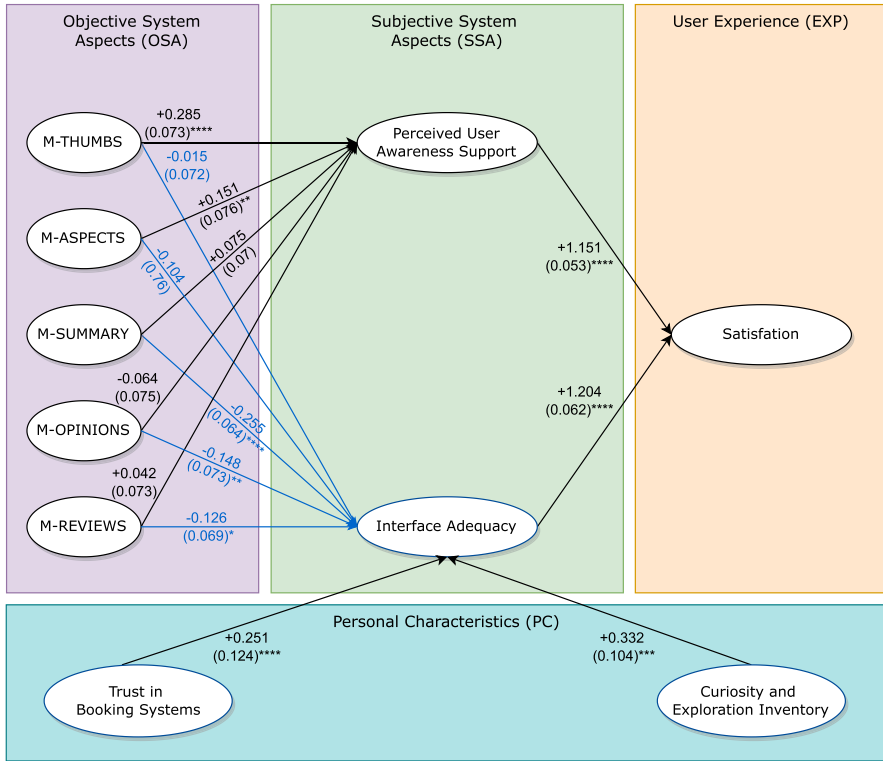
Fig. 9 Structural Equation Model. Significance levels: (****)$p < 0.001$, (***)$p < 0.01$, (**)$p < 0.05$, (*)$p < 0.1$. The numbers on the arrows represent the β-coefficients and standard error of the effect

that the people who have high trust in booking systems perceive the user interfaces of the justification models positively and they are more satisfied than the other users.

8.5 User experience analysis based on personality traits and cognitive styles

To further understand participants' perceptions of the justification models we analyzed the *Perceived User Awareness Support*, *Interface Adequacy* and *Satisfaction* by taking personality traits and cognitive style into account.

8.5.1 Curiosity trait

We grouped participants depending on the value they obtained when they answered the Curiosity and Exploration Inventory-II (CEI-II) questionnaire, which measures the motivation to seek out knowledge and new experiences (Stretching) and the willingness to embrace the novel, uncertain, and unpredictable nature of everyday life (Embracing). Participants with a high (or low) CEI-II value have high (low) Stretching and Embracing levels. The first group includes 24 people having a CEI-II value <3.5; the second one includes 35 people with a value ≥ 3.5. We could not obtain more

Table 7 Post-task questionnaire results grouped by CEI-II value

		CEI-II<3.5	CEI-II>=3.5
M- THUMBS	*Perceived User Awareness Support*	**3.67**	**3.66**
	Interface Adequacy	**3.64**	3.44
	Satisfaction	**3.51**	**3.44**
M- ASPECTS	*Perceived User Awareness Support*	3.29	3.58
	Interface Adequacy	3.29	**3.46**
	Satisfaction	3.21	3.35
M- SUMMARY	*Perceived User Awareness Support*	3.11**	3.36
	Interface Adequacy	3.08**	3.06**
	Satisfaction	2.94**	3.22
M- OPINIONS	*Perceived User Awareness Support*	2.92***	3.26*
	Interface Adequacy	3.06**	3.35
	Satisfaction	2.92**	3.18
M- REVIEWS	*Perceived User Awareness Support*	3.24	3.17
	Interface Adequacy	3.26	3.30
	Satisfaction	3.13	3.35**

The highest values for each group of participants are in boldface. Stars denote the statistical significance of the difference between the best-performing model and the other ones. Significance levels: (***)$p < 0.01$, (**)$p < 0.05$, (*)$p < 0.1$

balanced groups according to the results of the questionnaire. Table 7 shows the user experience results:

- The participants having low CEI-II value evaluated M- THUMBS as the best performing model in all the user experience constructs; see the values in bold in column "CEI-II<3.5". The difference between this model and M- SUMMARY and M- OPINIONS is statistically significant. The second best model is M- ASPECTS, regarding all three constructs.
- The situation for the participants having a high CEI-II value is mixed but supports the service-based justification models we propose. In this case, M- THUMBS obtained higher values than the baselines in all the constructs. However, M- ASPECTS excelled in *Interface Adequacy*.

These findings confirm that M- THUMBS is the best justification model, regardless of the user's curiosity. However, the *Interface Adequacy* results deserve further investigation. The main difference between M- THUMBS and M- ASPECTS is the fact that M- THUMBS summarizes consumer feedback through bar graphs and thumbs up/down. Differently, M- ASPECTS combines bar graphs with interactive filters representing the individual adjectives that are associated with the aspects. We can thus explain the diverse participants' perceptions with the fact that, as users having a low CEI-II value don't like to seek out knowledge, they prefer a model that offers a simple and quick (schematic) summary of consumer feedback, rather than a user interface offering advanced data exploration functions. Conversely, highly curious people are inclined towards interacting with fine-grained widgets and a larger amount of information.

Table 8 Post-task questionnaire results grouped by Need for Cognition (NfC) value

		NfC<3.5	NfC>=3.5
M- THUMBS	*Perceived User Awareness Support*	**3.73**	**3.61**
	Interface Adequacy	**3.68**	3.39
	Satisfaction	**3.60**	3.36
M- ASPECTS	*Perceived User Awareness Support*	3.63	3.33
	Interface Adequacy	3.51	3.29
	Satisfaction	3.44	3.18
M- SUMMARY	*Perceived User Awareness Support*	3.19**	3.31
	Interface Adequacy	2.99***	3.13*
	Satisfaction	3.00**	3.19
M- OPINIONS	*Perceived User Awareness Support*	3.21**	3.05***
	Interface Adequacy	3.36	3.13
	Satisfaction	3.04**	3.10
M- REVIEWS	*Perceived User Awareness Support*	3.03**	3.33
	Interface Adequacy	3.08**	**3.44**
	Satisfaction	3.05**	**3.42**

We use the same notation as in Table 7
The highest values for each group of participants are in boldface

8.5.2 Need for cognition

We also grouped participants depending on the value they obtained in the Need for Cognition (NfC) questionnaire, which measures the tendency to engage in and enjoy thinking. The first group includes the participants with a NfC value <3.5 (26 people); the second includes those having a NfC value ≥3.5 (33 people). Table 8 shows the results of the user experience analysis:

- On the group of participants having NfC<3.5, M- THUMBS is the best performing model and M- ASPECTS is the second best. Both models obtained higher user experience values than the baselines in all the constructs and most differences between M- THUMBS and the baselines are statistically significant.
- On the people having a high NfC value, M- THUMBS obtained the best evaluation regarding the *Perceived User Awareness Support* construct, followed by M- ASPECTS and M- REVIEWS. However, M- REVIEWS obtained the best results concerning *Interface Adequacy* and *Satisfaction*.

In summary, both participant groups perceived that our service-based justification models support awareness in a more efficacious way than the baselines. Moreover, M- THUMBS emerged as the best model for the people with low NfC while M- REVIEWS was the best one for the other users. We explain these observations with the fact that the people with low Need for Cognition prefer the models offering a schematic and organized summary of consumer feedback, such as bar graphs and thumbs. The other models are more suitable to the users having a high Need for Cognition because these people prefer to autonomously analyze and interpret data, as it happens in M- REVIEWS.

8.5.3 Discussion

M- THUMBS offers a concise view of information that requires little effort to understand consumer feedback. This makes it particularly good for the people with low curiosity or need for cognition. Moreover, M- THUMBS enriches the summary of information by enabling people to inspect the details of items they care about. For this reason, all users, regardless of their CEI-II and NfC values, appreciate its awareness support.

Differently, the people with high need for cognition prefer models such as M- REVIEWS and M- OPINIONS (which allow direct access to information without summarization) as far as the interface adequacy and satisfaction are concerned. This is probably due to the fact that these models are less assistive and offer larger freedom in data exploration. For instance, they show the plain text of the item reviews or a detailed list of features to analyze. In all such cases, users actively interact with the system to extract the data they are interested in.

8.6 Log analysis

Tables 9 and 10 summarize the analysis of the logged actions considering both the complete group of participants and the splits by CEI-II or NfC value. For each justification model, the tables report the mean time spent to explore the five homes presented during the user study, and the mean number of clicks or visualized data during the interaction with the test application. Specifically:

Table 9 Log analysis

		Total	CEI-II<3.5	CEI-II>=3.5
M- THUMBS	Time spent to explore 5 homes	175.58	205.79	154.86
	#clicks on the bar graphs	38.17	36.83	39.09
	#clicks on fine-grained dimensions	15.41	18.83	13.06
	#clicks to view more aspects	14.36	17.13	12.46
	#clicks on thumbs up/down	24.20	27.08	22.23
M- ASPECTS	Time spent to explore 5 homes	169.03	193.92	151.97
	#clicks on the bar graphs	29.00	36.25	24.03
	#clicks on fine-grained dimensions	13.24	17.29	10.46
	#clicks to view more aspects	12.36	15.92	9.91
	#clicks on the aspects	24.56	26.88	22.97
M- SUMMARY	Time spent to explore 5 homes	76.14	89.08	67.26
M- OPINIONS	Time spent to explore 5 homes	152.54	185.88	129.69
	#clicks on aspects	59.61	64.08	56.54
	#visualized aspects	80.66	88.25	75.46
M- REVIEWS	Time spent to explore 5 homes	270.07	310.71	242.20
	#visualized reviews	30.24	32.29	28.83

The Total column reports mean values for all the participants of the user study. The last two columns refer to the CEI-II groups

Table 10 Log analysis

		Total	NfC<3.5	NfC>=3.5
M- THUMBS	Time spent to explore 5 homes	175.58	124.96	215.45
	#clicks on the bar graphs	38.17	25.73	47.97
	#clicks on fine-grained dimensions	15.41	11.46	18.52
	#clicks to view more aspects	14.36	10.96	17.03
	#clicks on thumbs up/down	24.20	15.81	30.82
M- ASPECTS	Time spent to explore 5 homes	169.03	139.73	192.12
	#clicks on the bar graphs	29.00	22.23	34.33
	#clicks on fine-grained dimensions	13.24	12.23	14.03
	#clicks to view more aspects	12.36	11.54	13.00
	#clicks on aspects	24.56	19.15	28.82
M- SUMMARY	Time spent to explore 5 homes	76.14	69.92	81.03
M- OPINIONS	Time spent to explore 5 homes	152.54	104.58	190.33
	#clicks on the aspects	59.61	46.85	69.67
	#visualized aspects	80.66	72.46	87.12
M- REVIEWS	Time spent to explore 5 homes	270.07	225.04	305.55
	#visualized reviews	30.24	28.81	31.36

We use the same indicators and notation as in Table 9 and we repeat column Total for clarity. Participants are grouped by NfC

- "#clicks on the bar graphs" (M- THUMBS, M- ASPECTS) is the mean number of clicks to open the widget that shows the fine-grained dimensions associated with a specific coarse-grained one.
- "#clicks on fine-grained dimensions" (M- THUMBS, M- ASPECTS) is the mean number of clicks to view the aspects of a fine-grained dimension.
- "#clicks to view more aspects" (M- THUMBS, M- ASPECTS) is the mean number of clicks to visualize more aspects of a fine-grained dimension.
- "#clicks on thumbs up/down" (M- THUMBS) is the mean number of clicks to view the positive/negative quotes of reviews concerning an aspect.
- "#clicks on aspects" (M- ASPECTS) is the mean number of clicks to view the quotes of the reviews mentioning a specific aspect.
- "#visualized aspects" (M- OPINIONS) is the mean number of displayed aspects, given that the user can scroll down the list.
- "#visualized reviews" (M- REVIEWS) is the mean number of visualized reviews, given that the user can scroll down the list.

8.6.1 Complete group of participants

M- REVIEWS engaged participants in the interaction for the longest time because, to make an opinion about the homes, they had to read and analyze the reviews. In average, people visualized about 30 reviews, i.e., 6 for each home. As the user interface typically shows a maximum of 3 reviews (unless they are very short), we infer that participants scrolled down the review lists to inspect more opinions about the homes.

The opposite situation holds for M- SUMMARY that summarizes the opinions emerging from the reviews in a short text.

M- THUMBS and M- ASPECTS are in the middle of the ranking of all the justification models and engaged participants a bit longer than M- OPINIONS. Moreover, people spent slightly more time in interacting with M- THUMBS than M- ASPECTS. We explain these findings by analyzing the clicks on the widgets:

- When participants used M- THUMBS, on average they expanded the bar graphs 38.17 times, which roughly corresponds to 8 times per home. This means that they tended to go back and forth from one to the other one. They opened the widgets of specific fine-grained dimensions (mean total number: 15.41 times) and expanded the evaluation dimensions (14.36) to see the complete list of aspects about 3 times per home. The mean number of clicks on the thumbs up/down to view the quotes from the reviews is 24.20, i.e., about 5 clicks per home. This finding suggests that people considered the thumbs (which also show the number of reviews supporting the evaluations) a good synthesis of previous guests' perceptions of the homes. Thus, they did not need to inspect many quotes from the reviews.
- On M- ASPECTS, participants explored the coarse-grained evaluation dimensions about 6 times per home (mean total number: 29). Moreover, they performed about the same number of clicks on the fine-grained dimensions and they clicked about 5 aspects per home (24.56), similar to the number of clicks on thumbs up/down of M- THUMBS. As each aspect might have more than one adjective, this finding suggests that participants selectively inspected the quotes associated with the adjectives.
- On M- OPINIONS, participants visualized several aspects of the presented homes (about 16 per home, 80.66 in total). The reason is that M- OPINIONS directly shows the list of aspects, without grouping them by fine-grained dimension. Therefore, to make an opinion about a home, users had to check several aspects, looking for the relevant ones.

8.6.2 Curiosity trait

Table 9 shows the log analysis on the groups of participants having low and high CEI-II value, respectively. The time spent to evaluate the 5 homes of the experiment, and the clicking behavior on the user interfaces, have similar trends to those of the complete user group, confirming the interpretation we gave in the previous section. However, when comparing behavior of the participants with low and high CEI-II values on the same models, we can see that the people with low CEI-II spent more time in the evaluation than the other group. Moreover, they performed a larger number of clicks to inspect the aspects of the homes or their reviews. This is probably due to the fact that the less curious users are less efficient in finding the information they need and thus browse the user interface longer to make an opinion about the homes.

8.6.3 Need for cognition

Table 10 shows the log analysis on the groups of participants organized by NfC value. Also in this case, M- REVIEWS engages participants in the longest interaction to evaluate

homes and M-SUMMARY has the shortest evaluation times. However, there are differences concerning M-THUMBS and M-ASPECTS. The people having high NfC exhibited a similar pattern in the evaluation of homes as the complete group of participants. Differently, when using M-THUMBS, low NfC users spent less time in the evaluation but performed a higher number of clicks than in M-ASPECTS. While this finding seems to be contradictory, we explain it with the fact that, by summarizing the evaluation of aspects through thumbs up/down, M-THUMBS supports a more efficient evaluation of homes than M-ASPECTS that requires to investigate individual aspects.

8.6.4 Discussion

The log analysis, either performed on the whole group of participants, or on the subgroups split by curiosity trait/cognitive style, shows that M-REVIEWS engages users for the longest time to evaluate homes. This is because, other than checking the mean rating received by the items, it requires reading possibly long lists of reviews. In contrast, M-SUMMARY engages users for the shortest time because it proposes a summary of consumer feedback.

The incremental access to data supported by M-THUMBS and M-ASPECTS through bar graphs and fine-grained information exploration widgets (thumbs and clickable adjectives) engages users in the interaction a bit longer than M-SUMMARY and M-OPINIONS. However, the analysis of the clicks shows that participants evaluated the homes by inspecting a relatively low amount of data. Moreover, M-THUMBS and M-ASPECTS received the best evaluation of their information awareness support from all the participants. Thus, the longer time spent in interacting with the widgets can be interpreted as a sign of interest.

9 Lessons learned

The main finding that emerged from our study is that the *Perceived User Awareness Support* of our service-based justification models (especially by M-THUMBS) is higher than that of the state-of-the art models we considered. As this support is key to decision making and it is agnostic to the algorithms underlying item suggestion, our experiments encourage the introduction of service-based justification in recommender systems. However, participants' perception of the interface adequacy and satisfaction with these models depends on their curiosity traits and cognitive styles:

- The analysis of the user feedback shows that the overall group of participants evaluated M-THUMBS and M-ASPECTS as the best justification models on all the user experience constructs. Moreover, the Structured Equation Model analysis, and the fact that the service-based models received the smallest numbers of opting outs in the evaluation of the homes, confirms their superiority to the baselines. From the log analysis we also found that the service-based justification models enabled participants to make an opinion about the homes by inspecting a relatively low amount of information.
- However, the highly curious participants perceived the *Interface Adequacy* of M-ASPECTS (that supports the inspection of the individual adjectives of aspects)

higher than that of M-THUMBS (which summarizes opinions by means of thumbs up/down). Moreover, participants having low Need for Cognition preferred M-REVIEWS, which directly presents the item reviews, as far as the *Interface Adequacy* and *Satisfaction* constructs are concerned.

We interpret these findings by saying that the less curious users, and those with a low Need for Cognition, are comfortable with a justification model that organizes and summarizes data around the stages of interaction with items, offering a quick way to retrieve the data they care about. Differently, curious people and users with a high Need for Cognition benefit from the service-based organization of information but prefer to be autonomous in the analysis of consumer feedback. In line with the findings of other works, such as (Millecamp et al. 2022; Kouki et al. 2019, 2020), this suggests that, to support all users in an informed item selection, we should extend service-based justification of results with the personalization of the user interface to the user's characteristics.

10 Limitations and future work

As our justification models are broadly applicable but they fail to fully satisfy the users having a high Need for Cognition, we plan to personalize their user interfaces, e.g., by enabling the user to dynamically select the justification model to be used on an individual basis. Another possibility is to combine the information exploration functions offered by different models in a way that enables the user to choose the preferred ones. For instance, the service-based justification model might also offer a widget that enables the user to make a direct access to item reviews. In our future work, we plan to improve the user interface of our models by integrating, possibly in a personalized way, some of the functions offered by the other models. For this purpose, it will be crucial to understand how to recognize the user's Need for Cognition value, or how to make the user interface adaptable so that the user can select the preferred interaction features. The adaptability of the user interface might also be strategic to comply with user interface requirements implied by the level of risk of the task that users have to perform. In the user study, participants engaged in a realistic but artificial task because they did not spend their money to book the homes. Therefore, we could not measure their preferences for the user interfaces of the justification models under this perspective.

Another limitation of our service-based models is that they sort item aspects on a relevance basis by considering the number of reviews that mention them. In other words, we currently adhere to a conformity principle to decide which aspects are worth promoting. While the opinion of the mass is important, the user's interests might be considered as well to reflect individual priorities, as suggested in works such as (Pu and Chen 2007) and since the early faceted information exploration approaches (Tvarožek et al. 2008; Abel et al. 2011). In our future work, we plan to tailor the presentation of aspects by analyzing the user's preferences for the fine-grained evaluation dimensions corresponding to the stages of interaction with items, and by steering the content presented in the user interface accordingly. This type of personalization might be

applied to both aspect-based and service-based recommender systems, which take item aspects into account.

In our future work, we also plan to extend our Service Blueprint to model amendment episodes (e.g., due to exceptions at the consumers' side, or failures in the appliances of the rented home) that might involve interacting with the host, and thus impact consumers' experience as well.

Finally, we plan to test our service-based justification models on other domains such as the e-commerce one to assess their applicability to heterogeneous types of items. In this respect, it is worth mentioning that the specification of a new application domain is frequently supported by existing service models that can be adapted to the peculiarities of the selected domain. For instance, there are Service Blueprints defining the sales of products in online retailer platforms (Gibbons 2017) and food and beverage service systems (Nam et al. 2018). Specifically, in Gibbons (2017)'s blueprint, customer actions include visiting the website, visiting the store and browsing for products, discussing options and features with a sales assistant, purchasing the product, getting a delivery-date notification, and finally receiving the product. Moreover, the onstage/visible contact employee actions include the event of meeting customers in the store, the interaction with a chat assistant on the website to get more information about products, and so forth. Finally, the physical evidence layer includes, among the others, the products, the physical stores, the website, and the tutorial videos available on the website. All these elements of the blueprint can be exploited to define coarse-grained and fine-grained experience evaluation dimensions for our models.

11 Ethical issues

In planning the user study, we complied with literature guidelines on controlled experiments[3] (Kirk 2013). Through the user interface of our test application, and the informed consent, we notified participants about their rights: (i) the right to stop participating in the experiment at any time, without giving a reason; (ii) the right to obtain further information about the purpose, and the outcomes of the experiment; (iii) the right to have their data anonymized.

We did not collect participants' names, nor any data that could be used to identify them. During the user study, and the analysis of its results, we worked with anonymous codes, generated on the fly when users started the experiment. As described in Sect. 7, before starting the experiment, participants had to (i) confirm that they were 18 years old or over, (ii) read a consent form describing the nature of the experiment and their rights (https://bit.ly/3LypcZp), and (iii) confirm that they had read and understood their rights by clicking on the user interface of the test application. Every participant was given the same instructions before the experimental tasks. Our experiment has been approved by the ethics committee of the University of Torino (Protocol Number: 0421424).

---

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

12 Conclusions

We described a novel approach to justify recommender systems results based on an explicit representation of the service underlying the suggested items. Our goal was that of enhancing users' awareness about the suggested items with a holistic view of previous consumers' experience that considers the services and actors that the user might encounter during the overall interaction with an item, from its selection to its usage.

The existing work on the justification of recommender systems' results generates short descriptions (Musto et al. 2019, 2021) or flat lists of aspects (Chen et al. 2014) to present the suggested items. Differently, we leverage the Service Blueprints to summarize consumer feedback at different granularity levels, taking the stages of the service underlying items into account. Our approach supports an incremental access to the information about previous consumers' experience with items. By explicitly representing service-based data, the system provides users with a holistic description of items aimed at enhancing their awareness of the available options and their confidence in item evaluation.

We applied our approach to the home booking domain that is is particularly challenging because it exposes the user to the interaction with multiple entities, such as homes, hosts and surroundings, which might dramatically impact the renting experience. A user test involving 59 participants has shown that our service-based justification models obtain higher results than state-of-the-art baselines as far as the *Perceived User Awareness Support*, *Interface Adequacy* and *Satisfaction* user experience constructs are concerned. However, we noticed that the people having a high Need for Cognition prefer a direct analysis of consumer feedback, without the mediation of a summarization model. These results encourage the adoption of service-based justification models in recommender systems but they suggest to investigate the adaptation of these models to the individual user.

Appendix A Rotated figures
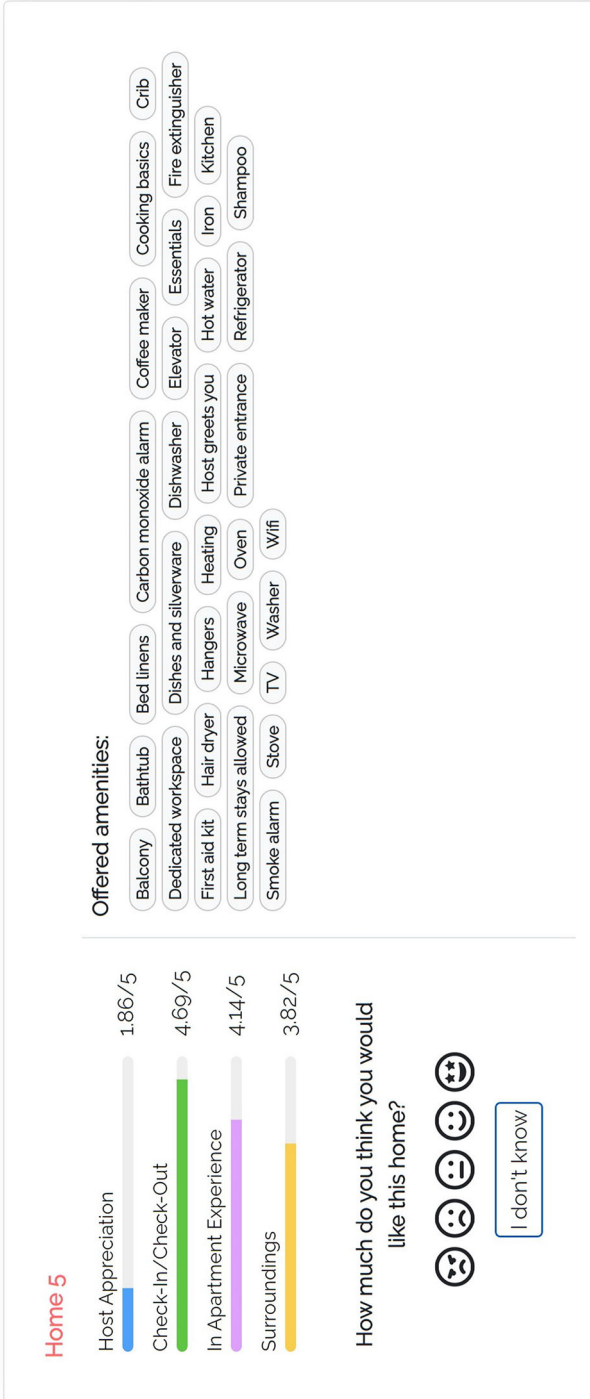
See Figs. 10, 11, 12, 13, 14.

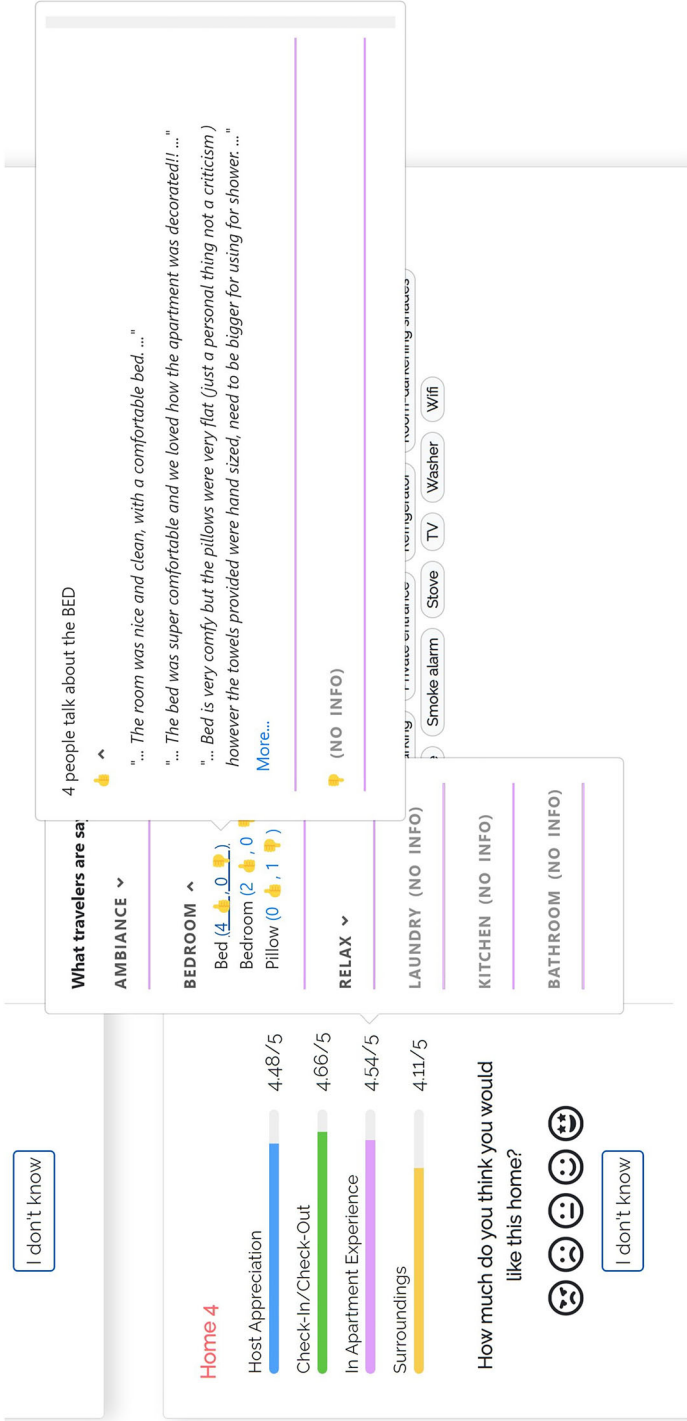**Fig. 10** Portion of the user interface shared by the proposed justification models
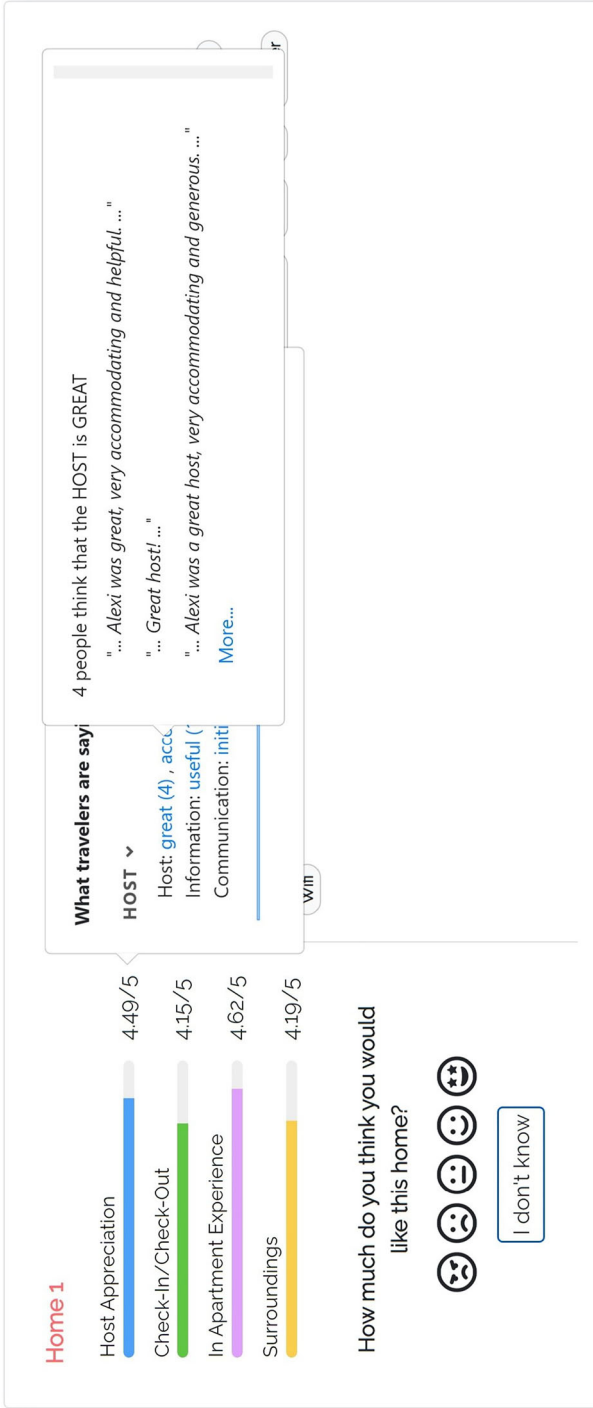
Fig. 11 User interface of justification model M- THUMBS

Fig. 12 User interface of justification model M- ASPECTS

Home 5

How much do you think you would like this home?

😣 😦 😐 🙂 😊

I don't know

Offered amenities:

Bed linens · Carbon monoxide alarm · Cooking basics · Dedicated workspace · Dishes and silverware · Essentials · Fire extinguisher · Hair dryer · Hangers · Heating · Hot water · Iron · Kitchen · Lock on bedroom door · Long term stays allowed · Microwave · Oven · Private entrance · Refrigerator · Shampoo · Smoke alarm · Stove · TV · Wifi

People who rented this home think that the location is great and perfect. Besides, previous guests think that the bed is uncomfortable.
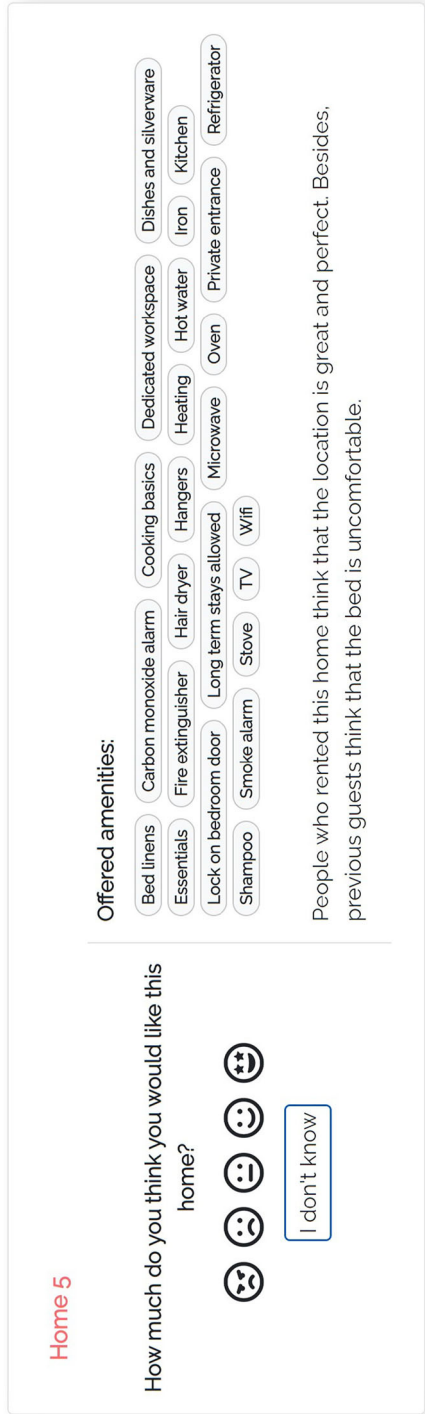
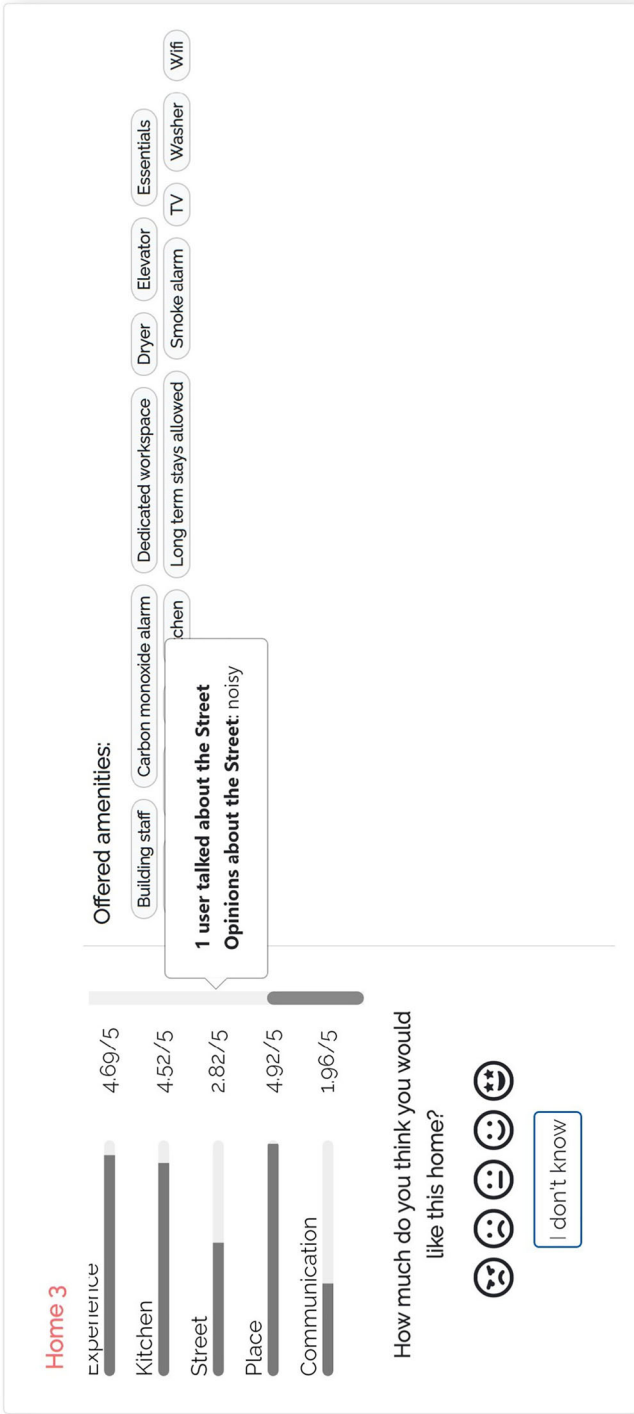**Fig. 13** User interface of justification model M- SUMMARY

Fig. 14 User interface of justification model M- OPINIONS

References

Abel, F., Celik, I., Houben, G.J., et al.: Leveraging the semantics of tweets for adaptive faceted search on twitter. In: Aroyo, L., Welty, C., Alani, H., et al. (eds.) The Semantic Web—ISWC 2011, pp. 1–17. Springer, Berlin (2011). https://doi.org/10.1007/978-3-642-25073-6_1

Airbnb (2022) Airbnb. https://airbnb.com

Amal, S., Tsai, C.H., Brusilovsky, P., et al.: Relational social recommendation: application to the academic domain. Expert Syst. Appl. **124**, 182–195 (2019). https://doi.org/10.1016/j.eswa.2019.01.061

Bitner, M.J., Ostrom, A.L., Morgan, F.N.: Service blueprinting: a practical technique for service innovation. Calif. Manag. Rev. **50**(3), 66–94 (2008). https://doi.org/10.2307/41166446

Cardoso, B., Sedrakyan, G., Gutiérrez, F., et al.: IntersectionExplorer, a multi-perspective approach for exploring recommendations. Int. J. Hum.-Comput. Stud. **121**, 73–92 (2019). https://doi.org/10.1016/j.ijhcs.2018.04.008

Chang, J. C., Hahn, N., Perer, A., et al.: SearchLens: composing and capturing complex user interests for exploratory search. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. ACM, New York, NY, USA, IUI '19, pp. 498–509 (2019). https://doi.org/10.1145/3301275.3302321

Chen, L., Wang, F.: Explaining recommendations based on feature sentiments in product reviews. In: Proceedings of the 22nd International Conference on Intelligent User Interfaces. Association for Computing Machinery, New York, NY, USA, IUI '17, pp. 17–28 (2017). https://doi.org/10.1145/3025171.3025173

Chen, L., Wang, F., Qi, L., et al.: Experiment on sentiment embedded comparison interface. Knowl.-Based Syst. **64**, 44–58 (2014). https://doi.org/10.1016/j.knosys.2014.03.020

Chen, L., Chen, G., Wang, F.: Recommender systems based on user reviews: the state of the art. User Model. User-Adapt. Interact. **25**(2), 99–154 (2015). https://doi.org/10.1007/s11257-015-9155-5

Cheng, M., Jin, X.: What do Airbnb users care about? An analysis of online review comments. Int. J. Hosp. Manag. **76**, 58–70 (2019). https://doi.org/10.1016/j.ijhm.2018.04.004

Coelho, G., Hanel, P.H.P., Wolf, L.J.: The very efficient assessment of need for cognition: developing a six-item version. Assessment **27**(8), 1870–1885 (2020). https://doi.org/10.1177/1073191118793208

Conati, C., Barral, O., Putnam, V., et al.: Toward personalized XAI: a case study in intelligent tutoring systems. Artif. Intell. **298**(103), 503 (2021). https://doi.org/10.1016/j.artint.2021.103503

Confalonieri, R., Coba, L., Wagner, B., et al.: A historical perspective of explainable artificial intelligence. WIREs Data Min. Knowl. Discov. **11**(1), e1391 (2021). https://doi.org/10.1002/widm.1391

Cramer, H.S.M., Evers, V., Ramlal, S., et al.: The effects of transparency on trust in and acceptance of a content-based art recommender. User Model. User-Adapt. Interact. **18**(5), 455–496 (2008). https://doi.org/10.1007/s11257-008-9051-3

Di Noia, T., Tintarev, N., Fatourou, P., et al.: Recommender systems under European AI regulations. Commun. ACM **65**(4), 69–73 (2022). https://doi.org/10.1145/3512728

Di Sciascio, C., Sabol, V., Veas, E. E.: Rank as you go: user-driven exploration of search results. In: Proceedings of the 21st International Conference on Intelligent User Interfaces. Association for Computing Machinery, New York, NY, USA, IUI '16, pp. 118–129 (2016). https://doi.org/10.1145/2856767.2856797

Di Sciascio, C., Brusilovsky, P., Trattner, C., et al.: A roadmap to user-controllable social exploratory search. ACM Trans. Interact. Intell. Syst. (2019). https://doi.org/10.1145/3241382

Dong, R., Smyth, B.: User-based opinion-based recommendation. In: Proceedings 26th IJCAI, pp. 4821–4825. Melbourne, Australia (2017)

European Commission General data protection regulation (GDPR) (2018). https://ec.europa.eu/info/law/law-topic/data-protection_en

Gibbons, S.: Service blueprints: definition (2017). https://www.nngroup.com/articles/service-blueprints-definition/

Herlocker, J. L., Konstan, J. A., Riedl, J.: Explaining collaborative filtering recommendations. In: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work. Association for Computing Machinery, New York, NY, USA, CSCW '00, pp. 241–250 (2000). https://doi.org/10.1145/358916.358995

Hernández-Rubio, M., Cantador, I., Bellogín, A.: A comparative analysis of recommender systems based on item aspect opinions extracted from user reviews. User Model. User-Adapt. Interact. **29**(2), 381–441 (2019). https://doi.org/10.1007/s11257-018-9214-9

Hutto, C., Eric, G.: 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. AAAI, New York, NY, USA, pp. 216–225 (2014).https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/viewPaper/8109

Jannach, D., Jugovac, M., Nunes, I.: Explanations and user control in recommender systems. In: Proceedings of the 23rd International Workshop on Personalization and Recommendation on the Web and Beyond. Association for Computing Machinery, New York, NY, USA, ABIS '19, pp. 31 (2019). https://doi.org/10.1145/3345002.3349293

Kashdan, T., Gallagher, M., Silvia, P., et al.: The curiosity and exploration inventory-II: development, factor structure, and psychometrics. J. Res. Personal. **43**, 987–998 (2009). https://doi.org/10.1016/j.jrp.2009.04.011

Kirk, R.E.: Experimental Design: Procedures for the Behavioral Sciences. SAGE Publications Inc., Thousand Oaks (2013). https://doi.org/10.4135/9781483384733

Knijnenburg, B.P., Willemsen, M.C.: Evaluating Recommender Systems with User Experiments, pp. 309–352. Springer, Boston (2015). https://doi.org/10.1007/978-1-4899-7637-6_9

Kouki, P., Schaffer, J., Pujara, J., et al.: User preferences for hybrid explanations. In: Proceedings of the Eleventh ACM Conference on Recommender Systems. Association for Computing Machinery, New York, NY, USA, RecSys '17, pp. 84–88 (2017). https://doi.org/10.1145/3109859.3109915

Kouki, P., Schaffer, J., Pujara, J., et al.: Personalized explanations for hybrid recommender systems. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. Association for Computing Machinery, New York, NY, USA, IUI '19, pp. 379–390 (2019). https://doi.org/10.1145/3301275.3302306

Kouki, P., Schaffer, J., Pujara, J., et al.: Generating and understanding personalized explanations in hybrid recommender systems. ACM Trans. Interact. Intell. Syst. (2020). https://doi.org/10.1145/3365843

Lee, C.K.H.: How guest-host interactions affect consumer experiences in the sharing economy: new evidence from a configurational analysis based on consumer reviews. Decis. Support Syst. **152**(113), 634 (2022). https://doi.org/10.1016/j.dss.2021.113634

Lewis, J.R., Sauro, J.: The factor structure of the system usability scale. In: Kurosu, M. (ed.) Human Centered Design, pp. 94–103. Springer, Berlin (2009)

Loepp, B., Herrmanny, K., Ziegler, J.: Blended recommending: integrating interactive information filtering and algorithmic recommender techniques. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, CHI '15, pp. 975–984 (2015). https://doi.org/10.1145/2702123.2702496

Loria, S.: TextBlob: simplified text processing (2020). https://textblob.readthedocs.io/en/dev/index.html

Lu, Y., Dong, R., Smyth, B.: Why i like it: Multi-task learning for recommendation and explanation. In: Proceedings of the 12th ACM Conference on Recommender Systems. Association for Computing Machinery, New York, NY, USA, RecSys '18, pp. 4–12 (2018). https://doi.org/10.1145/3240323.3240365

Mauro, N., Ardissono, L., Capecchi, S., et al.: Service-aware interactive presentation of items for decision-making. Appl. Sci. Spec. Issue Implic. Explicit Hum. Comput. Interact. **10**(16), 5599 (2020). https://doi.org/10.3390/app10165599

Mauro, N., Hu, Z. F., Ardissono, L.: Service-oriented justification of recommender system suggestions. In: Ardito, C., Lanzilotti, R., Malizia, A., et al (eds) Human-Computer-Interaction—INTERACT 2021, Lecture Notes in Computer Science, vol. 12936. Springer, Cham, Switzerland, pp. 321–330 (2021a). https://doi.org/10.1007/978-3-030-85613-7_23

Mauro, N., Hu, Z.F.F., Ardissono, L., et al.: A service-oriented perspective on the summarization of recommendations: preliminary experiment. In: Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization. Association for Computing Machinery, New York, NY, USA, pp. 213–219 (2021b)

Mauro, N., Hu, Z.F., Ardissono, L.: Service-aware personalized item recommendation. IEEE Access **10**, 26,715-26,729 (2022). https://doi.org/10.1109/ACCESS.2022.3157442

Millecamp, M., Htun, N. N., Conati, C., et al.: To explain or not to explain: the effects of personal characteristics when explaining music recommendations. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. Association for Computing Machinery, New York, NY, USA, IUI '19, pp 397–407 (2019). https://doi.org/10.1145/3301275.3302313

Millecamp, M., Htun, N. N., Conati, C., et al.: 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. Association for Computing Machinery, New York, NY, USA, UMAP '20, pp. 173–182 (2020). https://doi.org/10.1145/3340631.3394844

Millecamp, M., Conati, C., Verbert, K.: "Knowing me, knowing you": personalized explanations for a music recommender system. User Model. User-Adapt. Interact. **32**, 215–252 (2022). https://doi.org/10.1007/s11257-021-09304-9

Mohseni, S., Zarei, N., Ragan, E.D.: A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Trans. Interact. Intell. Syst. (2021). https://doi.org/10.1145/3387166

Muhammad, K. I., Lawlor, A., Smyth, B.: A live-user study of opinionated explanations for recommender systems. In: Proceedings of the 21st International Conference on Intelligent User Interfaces. Association for Computing Machinery, New York, NY, USA, IUI '16, pp. 256–260 (2016). https://doi.org/10.1145/2856767.2856813

Musto, C., Narducci, F., Lops, P., et al.: Linked open data-based explanations for transparent recommender systems. Int. J. Hum.-Comput. Stud. **121**, 93–107 (2019). https://doi.org/10.1016/j.ijhcs.2018.03.003

Musto, C., de Gemmis, M., Lops, P., et al.: Generating post hoc review-based natural language justifications for recommender systems. User-Model. User-Adapt. Interact. **31**, 629–673 (2021). https://doi.org/10.1007/s11257-020-09270-8

Nam, K.W., Kim, B.Y., Carnie, B.W.: Service open innovation; design elements for the food and beverage service business. J. Open Innov. Technol. Market Complex. (2018). https://doi.org/10.3390/joitmc4040053

Ni, J., Li, J., McAuley, J.: Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, pp. 188–197 (2019). https://doi.org/10.18653/v1/D19-1018

Nunes, I., Jannach, D.: A systematic review and taxonomy of explanations in decision support and recommender systems. User Model. User-Adapt. Interact. **27**(3–5), 393–444 (2017). https://doi.org/10.1007/s11257-017-9195-0

Parra, D., Brusilovsky, P.: User-controllable personalization: a case study with SetFusion. Int. J. Hum.-Comput. Stud. **78**, 43–67 (2015). https://doi.org/10.1016/j.ijhcs.2015.01.007

Pu, P., Chen, L.: Trust-inspiring explanation interfaces for recommender systems. Knowl.-Based Syst. **20**(6), 542–556 (2007). https://doi.org/10.1016/j.knosys.2007.04.004

Pu, P., Chen, L., Hu, R.: A user-centric evaluation framework for recommender systems. In: Proceedings of the Fifth ACM Conference on Recommender Systems. Association for Computing Machinery, New York, NY, USA, RecSys '11, pp. 157–164 (2011). https://doi.org/10.1145/2043932.2043962

Qiu, G., Liu, B., Bu, J., et al.: Opinion word expansion and target extraction through double propagation. Comput. Linguist. **37**, 9–27 (2011). https://doi.org/10.1162/coli_a_00034

Rana, A., D'Addio, R.M., Manzato, M.G., et al.: Extended recommendation-by-explanation. User-Model. User-Adapt. Interact. **32**, 91–131 (2022). https://doi.org/10.1007/s11257-021-09317-4

Ren, L., Qiu, H., Wang, P., et al.: Exploring customer experience with budget hotels: Dimensionality and satisfaction. Int. J. Hosp. Manag. **52**, 13–23 (2016). https://doi.org/10.1016/j.ijhm.2015.09.009

Ricci, F., Rokach, L., Shapira, B.: Recommender Systems: Techniques, Applications, and Challenges, pp. 1–35. Springer, New York (2022). https://doi.org/10.1007/978-1-0716-2197-4_1

Springer, A., Whittaker, S.: Progressive disclosure: empirically motivated approaches to designing effective transparency. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. Association for Computing Machinery, New York, NY, USA, IUI '19, pp. 107–120 (2019). https://doi.org/10.1145/3301275.3302322

Stickdorn, M., Schneider, J., Andrews, K.: This is Service Design Thinking: Basics, Tools, Cases. Wiley, Hoboken (2011)

Tintarev, N., Masthoff, J.: Evaluating the effectiveness of explanations for recommender systems. User Model. User-Adapt. Interact. **22**(4–5), 399–439 (2012)

Tintarev, N., Masthoff, J.: Beyond Explaining Single Item Recommendations, pp. 711–756. Springer, New York (2022). https://doi.org/10.1007/978-1-0716-2197-4_19

TripAdvisor.: Tripadvisor (2017). https://www.tripadvisor.it/

Tsai, C.H., Brusilovsky, P.: Exploring social recommendations with visual diversity-promoting interfaces. ACM Trans. Interact. Intell. Syst. **10**(1), 5:1-5:34 (2019). https://doi.org/10.1145/3231465

Tsai, C.H., Brusilovsky, P.: The effects of controllability and explainability in a social recommender system. User Model. User-Adapt. Interact. **31**(3), 591–627 (2021). https://doi.org/10.1007/s11257-020-09281-5

Tvarožek, M., Barla, M., Frivolt, G., et al.: Improving semantic search via integrated personalized faceted and visual graph navigation. In: Proceedings of the 34th Conference on Current Trends in Theory and Practice of Computer Science. Springer, Berlin, Heidelberg, SOFSEM'08, pp. 778–789 (2008). https://doi.org/10.1007/978-3-540-77566-9_67

Verbert, K., Parra, D., Brusilovsky, P.: Agents vs. users: visual recommendation of research talks with multiple dimension of relevance. ACM Trans. Interact. Intell. Syst. (2016). https://doi.org/10.1145/2946794

Wang, H., Zhang, F., Wang, J., et al.: Ripplenet: propagating user preferences on the knowledge graph for recommender systems. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. Association for Computing Machinery, New York, NY, USA, CIKM '18, pp. 417–426 (2018). https://doi.org/10.1145/3269206.3271739

Yi, J., Yuan, G., Yoo, C.: The effect of the perceived risk on the adoption of the sharing economy in the tourism industry: the case of Airbnb. Inf. Process. Manag. **57**(1), 102–108 (2020). https://doi.org/10.1016/j.ipm.2019.102108

Noemi Mauro, Ph.D., is an Assistant Professor at the Computer Science Department of the University of Torino, Italy where she obtained a Ph.D. in Computer Science with Honors. Her research interests concern user modeling, recommender systems, cultural heritage, information filtering, and information visualization. She is the author of several research articles related to intelligent systems and information filtering and she is a program committee member of the top conferences in her research areas and a reviewer for several related journals. She has been co-chair of several editions of the Workshop on Personalized Access to Cultural Heritage (PATCH) and she is a member of the Editorial Board of the international journal User Modeling and User-Adapted Interaction.

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 a co-author of some research papers related to information filtering and recommender systems. During his M.S. degree, he worked on the thesis in a research group at Alpen-Adria-Universität Klagenfurt.

Liliana Ardissono Ph.D., is a full Professor at the Computer Science Department of the University of Torino, Italy. Her research interests are User Modeling, Recommender Systems, and Information exploration support, with specific attention to geographic information search. She is a member of the Editorial Board of the international journal User Modeling and User-Adapted Interaction and a member of the Advisory Board of User Modeling Inc.