Enhancing an Interactive Recommendation System with Review-based Information Filtering

Jan Feuerbach
University of Duisburg-Essen
Duisburg, Germany
jan.feuerbach@stud.uni-due.de

Benedikt Loepp
University of Duisburg-Essen
Duisburg, Germany
benedikt.loepp@uni-due.de

Catalin-Mihai Barbu
University of Duisburg-Essen
Duisburg, Germany
catalin.barbu@uni-due.de

Jürgen Ziegler
University of Duisburg-Essen
Duisburg, Germany
juergen.ziegler@uni-due.de

ABSTRACT

Integrating interactive faceted filtering with intelligent recommendation techniques has shown to be a promising means for increasing user control in Recommender Systems. In this paper, we extend the concept of *blended recommending* by automatically extracting meaningful facets from social media by means of Natural Language Processing. Concretely, we allow users to influence the recommendations by selecting facet values and weighting them based on information other users provided in their reviews. We conducted a user study with an interactive recommender implemented in the hotel domain. This evaluation shows that users are consequently able to find items fitting interests that are typically difficult to take into account when only structured content data is available. For instance, the extracted facets representing the opinions of hotel visitors make it possible to effectively search for hotels with comfortable beds or that are located in quiet surroundings without having to read the user reviews.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; *Data mining*; *Information retrieval*; *Search interfaces*; • **Computing methodologies** → *Natural language processing*;

KEYWORDS

Interactive Recommending, Faceted Filtering, User Reviews

1 INTRODUCTION

Conventional automated *Recommender Systems* (RS) that pro-actively suggest items of potential interest to users make it often difficult to influence and to understand their outcome [21]. Interactive RS have been proposed that particularly aim at giving users more control over the recommendation process and at improving transparency. For instance, *TasteWeights* [4], *SetFusion* [27], *MyMovieMixer* [23], or *uRank* [10] allow users to vary the degree to which datasources, algorithms, product facets, or mined keywords are taken into account when generating recommendations. However, they often require up front availability of information such as existing user preference profiles or rich item data. In contrast, the increasing amount of user-provided content that is available online today has not yet been extensively exploited for interactive recommending. Social media, such as product reviews written by users in online shops or opinions about hotels on booking platforms, has been used so far primarily to deal with data sparsity and to increase algorithmic precision [7].

The concept of *blended recommending* [23] combines advantages of conventional automated RS, e.g. low user effort and high accuracy, with those of interactive information filtering, e.g. high level of control and transparency. For this, *faceted filtering* [15], which has shown to be an intuitive and efficient means for browsing large items spaces [15, 32], is integrated with different recommender techniques in a hybrid fashion. Consequently, users are enabled to select and weight criteria from facets leading to items being recommended based on their weighted relevance as determined by *Collaborative Filtering* (CF) or content-based techniques. In *MyMovieMixer* [23], an interactive RS based on this concept, users can select a movie as a facet value that in the following serves to suggest other movies that are similar with respect to their latent factor representation as derived from ordinary rating data. Users can also express that they want, for example, movies from a particular director, starring certain actors, or related in terms of user-generated tags, while being able to specify the weight of each of these criteria. *MyMovieMixer* was found especially promising for cold-start situations, i.e. without an existing rating profile for the current user, and when users only have a vague search goal in mind. Moreover, the approach allowed to significantly increase the perceived level of user control [23, 24].

In this paper, we build upon our prior work and extend the concept by extracting meaningful facets and corresponding values from user reviews by means of *Natural Language Processing* (NLP). So far, *blended recommending* has only been implemented based on ratings and structured content information. Relying on social media has several advantages. User-written product reviews, in particular, play an important role in buying decisions [9, 31]. They form a useful source

IntRS '17, August 27, 2017, Como, Italy

https://doi.org/

of information about what users liked or disliked, especially in case of "experience products". In the hotel domain, for instance, reviews may contain references to amenities that are typically not available as filters on online booking websites. Manually looking through dozens of reviews to check whether the recommended hotels have "comfortable beds" or are in "quiet locations" would be time-consuming and require a lot of cognitive effort. By automatically exploiting review data, in contrast, we allow users to directly express their preferences with respect to such subjective dimensions, consequently being able to better take their actual interests into account. Besides, from an information provider's perspective, user-generated content can be considered a useful addition to conventional objective product data that might be difficult to prepare for each item at the same level of quality.

The remainder of this paper is organized as follows: First, we discuss relevant related work. Next, we describe our proposed method for extracting facets from reviews using NLP. Then, we elaborate on how these facets can be used in *blended recommending* and introduce a demonstrator system we implemented in the hotel domain based on a real-world dataset[1]. Afterwards, we present a user study we conducted to examine the value of facets extracted according to our method in comparison to facets based on typical features as defined by content providers in terms of user experience. Finally, we conclude the paper and discuss avenues for further research.

2 RELATED WORK

Today's RS pro-actively suggest items that match individual preferences based on long-term user models. Producing well-fitting results can thereby reduce interaction effort and cognitive load [29]. However, the process of generating recommendations is often not controllable by users. Generally improving user experience, giving users more control, and increasing system transparency, have therefore been identified as important goals [21, 29], which are still only partially addressed in many real-world systems. RS research has for a long time been focused on algorithmic issues as well [21, 29]. For instance, in order to improve accuracy, several attempts have been made to integrate CF algorithms with additional data, including user-generated content such as tags, and in few cases also topics or opinions automatically extracted from user reviews [e.g. 1, 7, 18, 26, 37]. Only more recently, model-based CF has been enhanced for other purposes. One example is *TagMF* [12], a method that allows users to select and weight tags in order to manipulate the set of recommendations generated as in conventional RS based on ratings.

In general, many interactive recommending approaches have been proposed to overcome the issues of automated RS [14, 17, 22]. The cold-start problem, which occurs when no historical data is available for new users, has, for instance, been addressed algorithmically [e.g. 38], by taking reviews into account [e.g. 7], and by eliciting user preferences in an interactive manner [e.g. 25]. One prominent type of interactive RS are critique-based variants that allow users to iteratively refine the results by critiquing features of currently recommended items [8]. While this usually requires availability of well-defined product data, more recently, several attempts have been made that instead rely on, for instance, latent factors automatically derived from user ratings [25] or user-generated tags [35]. *MovieTuner* [35], as an example, first determines the relevance of tags and presents users with the most important ones. Then, users can indicate their preferences by critiquing recommendations in terms of these well understandable dimensions—which could represent subjective aspects not adequately describable by the often more technical objective product attributes. While it seems promising to elicit preferences this way, interaction is still typically limited to one single kind of item features, i.e. predefined metadata, latent factors *or* tags. Besides, user-provided content has only been exploited to a limited extent. The richness of social media such as user reviews has, to our knowledge, not yet been used for integrating RS with more interactivity.

In hybrid RS, multiple algorithms, and often multiple datasources, are combined to generate results with higher precision. This has consequently led to the development of corresponding interactive approaches that give users control over the recommendation process in terms of several dimensions at once. *TasteWeights* [4], a hybrid music recommender, allows users to directly weight different information types and social datasources, thereby increasing perceived recommendation quality and comprehensibility. The datasources used include social media artifacts such as Wikipedia articles, Facebook profiles or Twitter tweets, but without processing the content data on a semantic level to e.g. infer inherent user preferences. *SetFusion* [27] also employs a standard hybridization strategy, but enables users to weight each algorithm individually. In addition, the system provides a number of interactive features. However, it requires a persistent user profile and does not offer interaction with respect to any content-related criteria. *MyMovieMixer* [23] lets users select and weight facet values based on different types of recommender algorithms and related background data. The system increases the perceived amount of control by successfully enabling users to manipulate the result set not only with respect to explicitly defined product features, but also latent factors as well as user-generated data such as tags. Although the underlying concept of *blended recommending* is easily extendable, it has yet only been implemented using structured data that is directly given through the applied datasets. *uRank* [10] is one exception where keywords are extracted from background data especially to promote interactivity. The system focuses on exploration of document collections and supports users when their interests shift while browsing. The extraction is performed after some preprocessing steps by creating a vector space model using TF-IDF. Then, the keywords are presented to users as an interactive means to influence the recommendations by selecting and weighting

[1] We crawled metadata and overall 838 780 user reviews for 11 544 hotels located in five major European cities from Booking.com (http://www.booking.com/).

them. To increase transparency, their occurrences in the documents are visualized by means of a stacked bar chart and they are also used to generate an overview of the collection. Other approaches that make extensive use of visualizations for the purpose of increasing system transparency comprise, among others, *MoodPlay* [2] or *Conference Navigator* [34].

Overall, all of these works attempt to give users a high degree of control over hybrid RS. While this line of RS research to some extent converges with information filtering, there exists a wide range of manual filtering approaches outside the field of RS that can also be considered highly supportive for users finding the right items. *Faceted filtering* is one of the most prominent methods that supports exploration and discovery in large product spaces [15]. By selecting values from facets, the product space is iteratively constrained until the desired product is found. This principle also allows to e.g. facilitate keyword search and navigation in digital libraries or online shops [15]. Early attempts as well as contemporary real-world examples that can be found on many websites (e.g. accommodation booking platforms) usually rely on predefined features, support only Boolean filtering and conjunctive queries, and consider all selected facet values with equal importance [30, 32, 33, 36]. Only few exceptions employ fuzzy methods for value matching [13].

More recently, facets and facet values have also been automatically extracted, and adaptive techniques have been applied to faceted search based on e.g. semantic or social datasources [6, 16, 33]. For instance, *RevMiner* [16] extracts attribute-value pairs from restaurant reviews (e.g. "delicious pizza"), associates each value with a positive-negative score representing the sentiment, groups the attributes, and eventually presents them to the user in form of facets and facet values. When applying filter criteria, the restaurants in the results are ranked according to sentiment, strength and frequency of the selected value. Moreover, users can receive recommendations for other places with similar attributes. In general, previous attempts have however been focused on supporting users to select appropriate filter criteria and to deal with lack of metadata. Yet, the user's influence on the current filter setting is still limited. *VizBoard* [36], in contrast, allows users to prioritize selected criteria. Other work has also investigated user experience of faceted search as well as integrating visualizations. For example, in [32], a matrix visualization is used to display documents and their relevance with respect to the selected facets. While research in faceted filtering has thus brought numerous advances, the respective methods neither have yet made extensive use of recommender functionalities nor social media.

Summarizing the state-of-the-art, there exist various attempts that give users more control over RS, also in complex hybrid scenarios. Only to a limited extent, social media has thereby been utilized as a means to increase interactivity. Building on *MyMovieMixer* and extending the concept of *blended recommending* seems promising to go beyond integrating the datasources used so far (i.e. rating data, structured content information and explicit user-generated data such as tags), and to exploit the rich knowledge found in unstructured user-provided information such as reviews.

3 EXTRACTING MEANINGFUL FACETS FROM USER REVIEWS

In order to apply the concept of *blended recommending* based on a social datasource, we propose the procedure illustrated in Figure 1 to extract facets from user reviews.

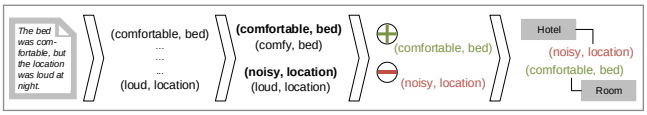

Figure 1: First, we identify attribute-value pairs such as "comfortable bed" in the reviews. Then, these pairs are merged with others that have the same meaning, e.g. "comfy bed". Next, using sentiment analysis, pairs are classified as positive or negative item properties. Finally, pairs that describe properties all related to e.g. hotel rooms are classified and grouped together to serve as values of a corresponding facet.

In the following, we elaborate on each of the steps involved, and also describe how we actually implemented them.

3.1 Identifying Attribute-Value Pairs

First, we split user reviews into sentences. Then, in each sentence, we need to identify nouns that describe properties of the respective item as well as adjectives which represent the opinion of the user who has written the review regarding these properties, e.g. "the *bed* was *comfortable*".

With the help of a *Part of Speech* tagger, we determine the word form of each term in a sentence. Based on the results, we establish grammatical relations using a dependency parser. Especially the relations *amod* (adjectival modifier) and *nsubj* (nominal subject) are of interest as they describe relations between adjectives and nouns. Besides, we take other relations into account to analyze more complex sentence structures, e.g. negations and relative clauses. Eventually, this gives us a set of attribute-value pairs which we then reduce to those pairs that appear a minimum number of times in all reviews.

For actually implementing this, we decided to use the *Stanford CoreNLP* toolkit[2], a lightweight framework with all the NLP functionality required in this step.

3.2 Merging Values

After attribute-value pairs have been identified, there might be multiple values sharing the same meaning, e.g. "large room" and "big room". Thus, we need to merge values that describe the same concept (i.e. synonyms) and replace them with a representative value to avoid confusion and redundancy.

For this purpose, we employ a lexical database providing links between synonyms. Since values may have different meanings depending on context (e.g. "big" and "heavy"), we do not directly use these links as criteria to merge them with others, but instead take the proportion of intersection of the

[2]https://stanfordnlp.github.io/CoreNLP/

sets of synonyms for each value into account. Then, a value is classified as being similar if a specified threshold is met. Otherwise, it is assumed to be a new representative itself. To improve the quality even further, we define groups of values and related representatives manually for terms that carry very different meanings across contexts, e.g. "good".

In the same step, we also identify pairs with opposite meaning, e.g. "comfortable bed" and "uncomfortable bed", and associate them by looking up the respective terms (i.e. antonyms) in a lexical database as well. Although we assume negative pairs to be less meaningful as filter criteria that can be selected by the user, we need them for later calculating item relevances (see Section 4).

As lexical database, we use *WordNet 2.1*[3].

3.3 Analyzing Sentiments

Next, in order to explicitly distinguish between positive and negative pairs for the recommendation process, we need to detect the sentiment of the pairs.

Adjectives often already represent a certain sentiment, e.g. "comfortable" can be considered positive. Thus, we decided to use an approach which determines sentiments for single words. Sentiment lexicons are databases where each term is assigned a sentiment value or class based on a certain algorithm or by human judgment. In contrast, in [16], a computational method specifically for reviews is proposed: By averaging helpfulness scores of user reviews in which a pair occurs, the respective values are classified as positive or negative based on a specified threshold.

To obtain adequate results on the dataset we used[1], we compared *SentiWordNet 3.0*[4], an algorithmically labeled sentiment lexicon, the *Stanford CoreNLP* toolkit which uses the *Sentiment Treebank*, a manually labeled dataset, and the method from [16]. By choosing the "right" threshold value, the latter achieved perfect accuracy, i.e. every value was labeled correctly in our test. Among the other two approaches, the *Stanford CoreNLP* toolkit yielded better results (accuracy of .933) than *SentiWordNet* (.667). Consequently, in cases where a threshold leading to adequate results could not be set for the computational method, we use the *Stanford CoreNLP* toolkit for performing the sentiment analysis as well.

3.4 Classifying Pairs

Finally, to support users in finding facet values that fit their goal and to reduce cognitive load, we aim at assigning the pairs to predefined categories based on the attributes.

For this purpose, we rely on the categories presented in [11] which resulted from collecting and grouping relevant hotel properties. Accordingly, we distinguish between pairs that either describe qualities of the "Hotel", the "Room", or related to the "Service". As classification method, we employ a semantic similarity metric utilizing the graph structure of a lexical database. There exist several interchangeable metrics that rely on different aspects of the underlying database. In any case, to assign an attribute to one of the categories, e.g. "bed" to "Room", a term to calculate the similarity with is required. Comparing with the class name itself would make it difficult to distinctly assign attributes. Thus, we employ typical terms from [11] and from taxonomies of popular booking websites for each category (e.g. the set for "Room" contains "bathroom", "bed" and "air conditioning"), and determine their average similarity with the respective attribute.

For implementing the classification, we use the *WS4J*[5]-API that offers a range of similarity metrics based on *WordNet*. When examining the results obtained on the dataset we used[1] with a set of manually labeled pairs, *HirstStOnge* yielded highest accuracy (.737).

4 BLENDED RECOMMENDING WITH EXTRACTED FACETS

In order to use the previously extracted attribute-value pairs for *blended recommending*, the corresponding facet values (e.g. "comfortable bed") have to be taken into account for calculating the relevance of the items when selected by the user. In the following, we describe how individual relevances are determined for each facet type, and how these relevance values eventually lead to an aggregated score for each hotel.

Standard Facets. As in [23], we use Boolean filtering for nominal facets ("Location") and fuzzy filtering for numerical facets ("Price", "Stars", "Score"). In case of Boolean filtering, items matching a selected value are additionally ranked using a criterion that establishes an ordering, e.g. score. In case of fuzzy filtering, the distance between a selected value and the respective item property determines the relevance (e.g. if the user selects the facet value "*score* = 8.0", items with a score of 7.5 are considered more relevant than items with 7.0).

Extracted Facets. For the extracted facets, i.e. related to "Hotel", "Room" and "Service", we in principle follow the way keywords are treated in [23], i.e. relevance calculation is based on the TF-IDF heuristic. Therefore, pairs are considered as terms, and sets of pairs associated with the hotels as documents. Table 1 shows an example where a user is looking for a hotel with a "comfortable bed".

Table 1: Results of different TF-IDF variants for an example where a user is looking for a hotel with a "comfortable bed".

	Hotel A	Hotel B	Hotel C
number of pairs	5	10	2
"comfortable bed"	3	2	1
"uncomfortable bed"	2	0	0
$tfidf_{baseline}$	.90	.60	.30
$tfidf_{norm}$	.13	.04	.11
$tfidf_{pair}$	.06	.06	.15

The baseline heuristic $tfidf_{baseline}$ results in hotel B being more relevant than C. Although hotel B is associated more often with the desired criterion than C, more pairs are associated with B in total, i.e. "comfortable bed" cannot be assumed to be a very distinctive characteristic of this hotel.

[3] https://wordnet.princeton.edu/
[4] http://sentiwordnet.isti.cnr.it/
[5] https://github.com/Sciss/ws4j/

Consequently, we additionally normalize the frequency using the overall number of associated pairs. According to the modified heuristic $tfidf_{norm}$, hotel A is still more relevant than C. However, even though "comfortable bed" is relatively associated more often with hotel A than with C, hotel A is also associated with the opposite pair. Hence, we not only consider the TF-IDF value for the positive, but also for the negative pair:

$$tfidf_{pair} = tfidf_{positive} - tfidf_{negative} \quad (1)$$

Item Relevance. Finally, individual relevance scores rel_i[6] for each facet value f_i are aggregated using the corresponding weights w_i by means of arithmetic mean as in [23]. The overall relevance rel of a hotel h is thus calculated as follows:

$$rel(h, f_1, ..., f_n, w_1, ..., w_n) = \frac{\sum_{i=1}^{n} w_i \cdot rel_i(h, f_i)}{\sum_{i=1}^{n} w_i} \quad (2)$$

5 DEMONSTRATOR SYSTEM

To finally demonstrate how *blended recommending* can be implemented based on facets extracted from user reviews, we developed a web application in the hotel domain using the dataset we crawled from Booking.com[1]. The demonstrator system which generally follows the design of *MyMovieMixer* [23] is shown in Figure 2.

On the left side, a list comprising all facets users can choose from is presented. Clicking on a facet expands it and shows corresponding facet values in form of tiles (A). Initially, the values of each facet are hidden (B) to reduce cognitive load. For the facets "Price", "Stars" and "Score", some tiles represent predefined values (e.g. "$30 - 45$ Euro"). In addition, users can create tiles themselves by manually specifying preferred ranges (e.g. "$25 - 45$ Euro"). The "Location" facet presents users with predefined tiles for each location in the dataset. For the extracted facets, i.e. related to "Hotel", "Room" and "Service", we initially show those attribute-value pairs that occur most often in the reviews assuming they are generally more important for users. Thereby, we consider only positive values because it is unlikely that users want to receive recommendations related to negative properties. Yet, users can request more tiles, i.e. the next most frequent pairs, by clicking the respective button (C). Moreover, to look for specific values which might be useful to pursue a particular search goal, users can also perform a text-based search.

As soon as users drag a tile into the preference area in the middle of the screen, the corresponding facet value is considered for generating recommendations, i.e. its individual relevance rel_i is now used in (2) when calculating overall item relevances. In this area, each tile is accompanied by a slider that allows users to weight the respective criterion, i.e. to modify w_i (D). In case users are no longer interested in applying a specific criterion, tiles can be removed from the preference area (E). Adding or removing tiles as well as changing their weight immediately updates the results.

[6]Scores are determined as described above. Note that this might not be possible when a criterion is not referred to in the reviews of a hotel.

The resulting recommendations are shown on the right side (F). We deliberately limit their number to reduce choice difficulty and to motivate users to manipulate the results by selecting criteria and weighting them, this way being able to explore the effects of their preference settings in an interactive manner. However, if users are not satisfied with particular recommendations, the respective hotels can be removed from the list so that the next most relevant item appears. Each recommendation is displayed with a photo and the top-3 attribute-value pairs[7] occurring in the related reviews (the number of occurrences is shown in brackets). By clicking on a recommendation, a list of all associated pairs as well as further metadata is shown in a dialog. Since negative opinions can also have an impact on the decision-making process [9], users are here presented with both pairs having positive (green) or negative (red) sentiments.

When users are satisfied with a recommendation, the respective hotel can be dragged into the basket in the top-right corner (G). This area serves to store items that users would like to take into consideration for their final decision. To enable further refinement, the system also suggests new tiles as soon as an item is put into the basket: The last values from each of the extracted facets on the left side are replaced by the top-3 pairs of the respective hotel. Moreover, the basket is used for evaluation purposes.

6 EVALUATION

To examine the benefits extracting facets from user reviews has in terms of user experience in comparison to using explicitly defined features as they are typically found on booking websites, we performed a user study. In this study, we compared the demonstrator system described in the previous section[8] with an almost identical variant where instead of facets extracted according to our method we used facets based on well-defined provided features.

6.1 Method

Participants and Materials. We recruited 30 participants (24 female) with an average age of $M = 21.63$ ($SD = 3.32$). Most of the participants were students; only two of them were employed. Participants were asked to use the system under controlled conditions in a lab-based setting. During the course of the study, they used a desktop PC with $24''$ LCD (1920×1200 px resolution) and a standard web browser to fill in a questionnaire and to perform several tasks.

Procedure. Participants were assigned (in counter-balanced order) in a between-subject design[9] to one of the two following

[7]Note that these are pairs *after* performing all steps described in Section 3, i.e. they do not necessarily appear in exactly the same way in all underlying reviews (some reviews may refer to e.g. an "excellent location", which however would increase the count for "great location").

[8]For the user study, we used an earlier version of the demonstrator system which was slightly different, in particular, with respect to the level of detail for presented recommendations (among others, frequent attribute-value pairs were not visible right away).

[9]We decided against a within-subject design to avoid carry-over effects and to reduce participants' workload.

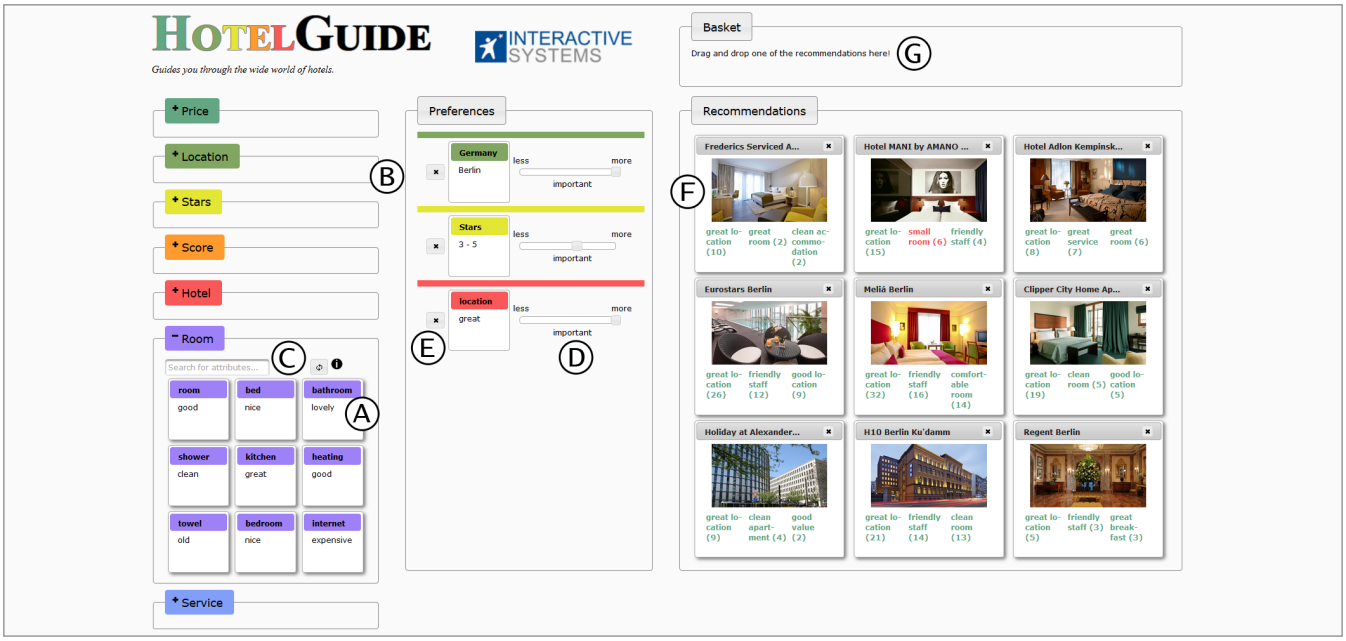

Figure 2: Screenshot of our demonstrator system: Facet values are shown as tiles on the left side (A). Users can expand and collapse each facet (B). For some facets, users can search and ask for more tiles (C). As soon as users drag tiles into the preference area in the middle, an accompanying slider allows to weight the corresponding value (D). If users do not want to consider a criterion anymore, they can remove it (E). Recommended hotels are shown on the right side with images and attribute-value pairs most frequently occurring in the related reviews (F). Users can put items they like from the results into a basket (G).

conditions (15 per condition), which varied regarding the source used for "Hotel", "Room", and "Service" facet:

Feature-Based Facets *(FF)*: Demonstrator system with values for the "Hotel" (89 values), "Room" (124), and "Service" (109) facets based on predefined features[10].

Extracted Facets *(EF)*: Demonstrator system with values for the "Hotel" (562 values), "Room" (266), and "Service" (1038) facets based on attribute-value pairs extracted from a dataset of reviews[1] as described in Section 3.

The user study comprised four tasks which were presented (in random order) as scenarios described as follows:

Task 1 *"You want to do a weekend trip to London or Berlin with a friend of yours. You want to spent a maximum of 40 Euros per night. The accommodation should have reliable Wifi and a bar."*

Task 2 *"You are going on vacation to Brussels with your parents. The accommodation should have 3 stars and should cost about 80 Euros. Furthermore, your parents want a nice view and a good breakfast."*

Task 3 *"You want to surprise your partner with a short getaway to Rome. You have some money, so you can spend 60 Euros per night. Since you would like to have some private time, the place should be quiet and you want to have your own (large and clean) bathroom."*

Task 4 *"You get a one-week holiday as a gift. Money plays no role and you are able to freely choose the location."*

Questionnaires and Log Data. At the beginning of each session, we elicited demographics and domain knowledge. To assess participants' subjective perception of the respective system variant, we used a questionnaire primarily composed of existing constructs. Concretely, after each task, we used the evaluation framework proposed in [20] to assess perceived recommendation quality, perceived set variety, choice satisfaction as well as choice difficulty, usage effort, and perceived effectiveness. Since interaction influences user experience, we also assessed the intention to provide feedback [20] and tracked user behavior. Relying on [28], we additionally assessed perceived usefulness and overall satisfaction. Furthermore, we formulated questionnaire items ourselves to particularly address helpfulness and understandability of the facets, their suitability for expressing preferences, and participants' satisfaction with them. Finally, at the end of each session, we asked the same questions again, now regarding participants' general impression independent of specific tasks. In addition, we applied the *System Usability Scale* (SUS) [5]. All items were assessed on a positive 5-point Likert scale.

6.2 Results

Domain Knowledge and Usability. Overall, participants reported average domain knowledge with no significant difference ($t(28) = .54$, $p = .595$) between conditions (FF: $M = 2.87$, $SD = 1.13$; EF: $M = 2.67$, $SD = 0.90$). Regarding usability, both

[10] Hotel features were crawled from Booking.com and manually assigned to the three facets. For example, we associated the feature "non-smoking room", that relying on their taxonomy is explicitly given to hotels at the Booking.com website, with the "Room" facet.

variants of the demonstrator system received "good" scores on the SUS, with 76 in the FF, and 83 in the EF condition.

User Experience. Table 2 shows the results regarding participants' general impression[11]. We conducted t-tests ($\alpha = .05$) to assess differences between conditions. EF was rated superior to FF with respect to all constructs. As highlighted, there were significant differences (with medium to large effect size) in terms of perceived variety of the recommendation set, choice difficulty, and intention to provide feedback.

Table 2: t-test ($df = 28$) results with means and SDs for the overall comparison of the two conditions (* indicates significance at 5 % level; d represents Cohen's effect size value).

	Feat.-Based Facets		Extracted Facets				
	M	*SD*	*M*	*SD*	*T*	*p*	*d*
Perc. Rec. Quality	3.97	0.83	**4.03**	0.79	-0.23	.824	.07
Perc. Set Variety	3.60	0.51	**4.13**	0.83	-2.12	.043*	.77
Choice Satisfaction	4.27	0.70	**4.40**	0.63	-0.55	.590	.20
Choice Difficulty	2.73	1.34	**3.67**	0.98	-2.19	.037*	.80
Perc. Effectiveness	3.60	1.18	**4.07**	1.10	-1.12	.273	.41
Usage Effort	3.70	0.94	**4.07**	0.86	-1.11	.276	.41
Feedback Intention	3.20	0.78	**3.87**	0.83	-2.27	.031*	.83
Usefulness	3.69	0.73	**4.18**	0.85	-1.69	.103	.62
Overall Satisfaction	3.73	1.10	**4.07**	0.96	-0.88	.384	.33

Moreover, we found significant gender differences ($t(28) = -2.48$, $p = .019$) regarding recommendation quality, with men giving higher ratings ($M = 4.67$, $SD = 0.41$) than women ($M = 3.83$, $SD = 0.79$), with large effect size ($d = 1.34$). Women (M=367.00 sec, $SD = 138.91$) also took on average significantly longer ($t(28) = 2.08$, $p = .047$) to accomplish tasks than men (M=245.17 sec, $SD = 63.77$), with large effect size ($d = 1.13$).

Facets. After each task, we assessed participants' opinions specifically on the facets. We conducted two-way RM ANOVA to examine effects of condition and task. Interaction terms were not significant, and we found only few small differences between tasks. Table 3 shows that participants perceived the suitability of the facets for expressing their preferences significantly higher in the EF condition than in the other, independent of the task. With respect to all other variables, EF was assessed superior to FF as well, but without significances.

Table 3: ANOVA ($df1 = 1$, $df2 = 28$) results with means and SEs for the comparison of the conditions across tasks (* indicates significance at 5 % level; we aggregated scores assessed individually for the "Hotel", "Room", and "Service" facet).

	Feat.-Based Facets		Extracted Facets			
	M	*SE*	*M*	*SE*	*F*	*p*
Suit. for Exp. Preferences	3.38	0.179	**3.93**	0.179	4.604	.041*
Helpfulness	3.82	0.185	**4.12**	0.185	1.322	.260
Understandability	4.26	0.176	**4.27**	0.176	0.002	.965
Satisfaction	3.96	0.161	**4.16**	0.161	0.776	.386

[11] We examined effects of condition and task using two-way RM ANOVA. Interaction terms were only significant for two variables (variety, effort), each showing differences in only one pairwise comparison. Since the results obtained after each task were overall tendentially similar, we thus omit reporting them separately. Instead, we present the scores from the final assessment where participants were asked regarding their general impression after completing all tasks.

Interaction Behavior. Concerning actual user behavior, no significant differences were found for the number of facet values being selected (FF: $M = 3.77$, $SD = 1.53$; EF: $M = 3.57$, $SD = 1.55$), ($t(28) = .36$, $p = .72$), the number of times more facets values were requested (FF: $M = 24.93$, $SD = 39.53$; EF: $M = 4.67$, $SD = 13.06$), ($t(17) = 1.87$, $p = .08$), and the number of recommendations removed from the results (FF: $M = 5.67$, $SD = 12.84$; EF: $M = 5.87$, $SD = 12.71$), ($t(28) = -.04$, $p = .97$).

6.3 Discussion

In conclusion, the user study shows that the concept of *blended recommending* can be successfully applied relying on social datasources. When compared to the system variant with facets based on features from a well-established taxonomy (FF), the variant with facets extracted from user reviews (EF) obtained overall superior results after the individual tasks as well as in the end after participants finished all tasks. Regarding participants' general impression, significant differences were identified for several relevant variables. Concretely, set diversity was perceived to be higher and it was easier for participants to settle on one of the recommended hotels (which is in line with earlier research [3]). Furthermore, participants' feedback intention was higher, i.e. they preferred to provide feedback in the EF condition. This is corroborated by their answers to the questionnaire items that specifically addressed the perception of facets. Apparently, in all tasks, participants valued the possibility to express their preferences with respect to the more subjective dimensions represented through facets extracted from reviews. At the same time, it can be considered promising that we did not find a negative effect in terms of facet understandability. In contrast, our proposed method seems able to extract facets from a real-world review dataset in a meaningful way.

By showing an effect of gender on perceived quality, our study partly validates earlier findings that such factors influence how important reviews are considered as an information source by individual users [19]. Thus, it is subject of future studies to explore more deeply the impact of personality variables on the use of different information sources. In this regard, it is important to note that there might have been confounding factors. In particular, with the present experimental design, participants did not know the source of the facet values, i.e. that they were extracted from user reviews. However, as reviews particularly in the hotel domain influence the perception of trust [31], knowing where the information comes from might positively affect perceived usefulness and possibly also trustworthiness—in particular, if it would be possible to trace back from mined attribute-value pairs to underlying reviews. Besides, we were not able to establish all relations between pairs and hotels due to time restrictions, potentially contributing negatively to the assessment in the EF condition. The lack of differences in actual user behavior may in contrast be attributed to the almost identical interfaces of the two demonstrator variants. In summary, the assessment however yielded promising results with respect to all variables. Since interaction terms of condition and task

were significant only in two cases, we deduce that this applies independent of the task and its complexity. Nevertheless, further investigating task-related differences is subject of future work. Overall, participants seemed more satisfied in the EF condition, which is reflected accordingly in effect sizes.

7 CONCLUSIONS AND OUTLOOK

In this paper, we have presented an extension to the concept of *blended recommending*. Relying on social media, in particular, reviews for hotels, we allow users to specify their preferences with respect to meaningful criteria that usually are not available as filter options, but especially useful for adequately choosing from sets of recommended "experience products". In line with that, the user study we conducted has shown significant improvements with respect to diversity and choice difficulty, but also promising results in terms of other relevant variables. Thus, it seems that user reviews can be successfully exploited for interactive recommending, and that the contained information is of value for users even without actually reading them. Besides, the study demonstrates that the concept can be applied—also in absence of structured content data—in other domains than movies.

In future work, we aim at improving and possibly using different NLP methods to extract the facets. Moreover, reviews could be further exploited to improve the presentation of recommendations. For instance, by identifying users with shared interests, results could be accompanied with summaries of their opinions in order to make suggestions easier to understand and to increase the system's trustworthiness. Finally, all of this would go hand in hand with conducting more user studies, e.g. to evaluate the specific improvements.

REFERENCES

[1] A. Almahairi, K. Kastner, K. Cho, and A. Courville. 2015. Learning distributed representations from reviews for collaborative filtering. In *RecSys '15*. ACM, 147–154.

[2] I. Andjelkovic, D. Parra, and J. O'Donovan. 2016. Moodplay: Interactive mood-based music discovery and recommendation. In *UMAP '16*. ACM, 275–279.

[3] D. Bollen, B. P. Knijnenburg, M. C. Willemsen, and M. P. Graus. 2010. Understanding choice overload in recommender systems. In *RecSys '10*. ACM, 63–70.

[4] S. Bostandjiev, J. O'Donovan, and T. Höllerer. 2012. TasteWeights: A visual interactive hybrid recommender system. In *RecSys '12*. ACM, 35–42.

[5] J. Brooke. 1996. SUS – A quick and dirty usability scale. In *Usability Evaluation in Industry*. Taylor & Francis, 189–194.

[6] I. Celik, F. Abel, and P. Siehndel. 2011. Towards a framework for adaptive faceted search on twitter. In *DAH '11*. 11–22.

[7] L. Chen, G. Chen, and F. Wang. 2015. Recommender systems based on user reviews: The state of the art. *UMUAI* 25, 2 (2015), 99–154.

[8] L. Chen and P. Pu. 2012. Critiquing-based recommenders: Survey and emerging trends. *UMUAI* 22, 1-2 (2012), 125–150.

[9] J. A. Chevalier and D. Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *J. Marketing Res.* 43, 3 (2006), 345–354.

[10] C. di Sciascio, V. Sabol, and E. E. Veas. 2016. Rank as you go: User-driven exploration of search results. In *IUI '16*. ACM, 118–129.

[11] S. Dolnicar and T. Otter. 2003. *Which hotel attributes matter? A review of previous and a framework for future research.* Technical Report. University of Wollongong.

[12] T. Donkers, B. Loepp, and J. Ziegler. 2016. Tag-enhanced collaborative filtering for increasing transparency and interactive control. In *UMAP '16*. ACM, 169–173.

[13] A. Girgensohn, F. Shipman, F. Chen, and L. Wilcox. 2010. DocuBrowse: Faceted searching, browsing, and recommendations in an enterprise context. In *IUI '10*. ACM, 189–198.

[14] C. He, D. Parra, and K. Verbert. 2016. Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities. *Expert Syst. Appl.* 56, 1 (2016), 9–27.

[15] M. A. Hearst. 2009. *Search user interfaces*. Cambridge University Press.

[16] J. Huang, O. Etzioni, L. Zettlemoyer, K. Clark, and C. Lee. 2012. RevMiner: An extractive interface for navigating reviews on a smartphone. In *UIST '12*. ACM, 3–12.

[17] M. Jugovac and D. Jannach. 2017. Interacting with recommenders - Overview and research directions. *ACM TiiS* (2017).

[18] A. Karatzoglou, X. Amatriain, L. Baltrunas, and N. Oliver. 2010. Multiverse recommendation: N-dimensional tensor factorization for context-aware collaborative filtering. In *RecSys '10*. ACM, 79–86.

[19] E. E. K. Kim, A. S. Mattila, and S. Baloglu. 2011. Effects of gender and expertise on consumers' motivation to read online hotel reviews. *Cornell Hosp. Q.* 52, 4 (2011), 399–406.

[20] B. P. Knijnenburg, M. C. Willemsen, and A. Kobsa. 2011. A pragmatic procedure to support the user-centric evaluation of recommender systems. In *RecSys '11*. ACM, 321–324.

[21] J. A. Konstan and J. Riedl. 2012. Recommender systems: From algorithms to user experience. *UMUAI* 22, 1-2 (2012), 101–123.

[22] B. Loepp, C.-M. Barbu, and J. Ziegler. 2016. Interactive recommending: Framework, state of research and future challenges. In *EnCHIReS '16*. 3–13.

[23] B. Loepp, K. Herrmanny, and J. Ziegler. 2015. Blended recommending: Integrating interactive information filtering and algorithmic recommender techniques. In *CHI '15*. ACM, 975–984.

[24] B. Loepp, K. Herrmanny, and J. Ziegler. 2015. Merging interactive information filtering and recommender algorithms – Model and concept demonstrator. *i-com* 14, 1 (2015), 5–17.

[25] B. Loepp, T. Hussein, and J. Ziegler. 2014. Choice-based preference elicitation for collaborative filtering recommender systems. In *CHI '14*. ACM, 3085–3094.

[26] J. McAuley and J. Leskovec. 2013. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *RecSys '13*. ACM, 165–172.

[27] D. Parra, P. Brusilovsky, and C. Trattner. 2014. See what you want to see: Visual user-driven approach for hybrid recommendation. In *IUI '14*. ACM, 235–240.

[28] P. Pu, L. Chen, and R. Hu. 2011. A user-centric evaluation framework for recommender systems. In *RecSys '11*. ACM, 157–164.

[29] P. Pu, L. Chen, and R. Hu. 2012. Evaluating recommender systems from the user's perspective: Survey of the state of the art. *UMUAI* 22, 4-5 (2012), 317–355.

[30] G. M. Sacco. 2006. Dynamic taxonomies and guided searches. *J. Am. Soc. Inf. Sci. Tec.* 57, 6 (2006), 792–796.

[31] B. A. Sparks and V. Browning. 2011. The impact of online reviews on hotel booking intentions and perception of trust. *Tourism Manage.* 32, 6 (2011), 1310–1323.

[32] V. T. Thai, P.-Y. Rouille, and S. Handschuh. 2012. Visual abstraction and ordering in faceted browsing of text collections. *ACM TIST* 3, 2 (2012), 21:1–21:24.

[33] M. Tvarožek, M. Barla, G. Frivolt, M. Tomša, and M. Bieliková. 2008. Improving semantic search via integrated personalized faceted and visual graph navigation. In *SOFSEM '08*. Springer, 778–789.

[34] K. Verbert, D. Parra, and P. Brusilovsky. 2016. Agents vs. users: Visual recommendation of research talks with multiple dimension of relevance. *ACM TiiS* 6, 2 (2016), 11:1–11:42.

[35] J. Vig, S. Sen, and J. Riedl. 2011. Navigating the tag genome. In *IUI '11*. ACM, 93–102.

[36] M. Voigt, A. Werstler, J. Polowinski, and K. Meißner. 2012. Weighted faceted browsing for characteristics-based visualization selection through end users. In *EICS '12*. ACM, 151–156.

[37] Y. Zhang, G. Lai, M. Zhang, Y. Zhang, Y. Liu, and S. Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *SIGIR '14*. ACM, 83–92.

[38] K. Zhou, S.-H. Yang, and H. Zha. 2011. Functional matrix factorizations for cold-start recommendation. In *SIGIR '11*. ACM, 315–324.