Journal of Information Technology and Tourism manuscript No.

Leveraging multi-criteria customer feedback for satisfaction analysis and improved recommendations

Dietmar Jannach · Markus Zanker ·
Matthias Fuchs

Abstract Travel websites and online booking platforms represent today's major sources for customers when gathering information before a trip. In particular, community-provided customer reviews and ratings of various tourism services represent a valuable source of information for trip planning. With respect to customer ratings, many modern travel and tourism platforms – in contrast to several other e-commerce domains – allow customers to rate objects along multiple dimensions and thus to provide more fine-granular post-trip feedback on the booked accommodation or travel package.

In this paper, we first show how this multi-criteria rating information can help to obtain a better understanding of factors driving customer satisfaction for different segments. For this purpose, we performed a Penalty-Reward Contrast analysis on a data set from a major tourism platform, which reveals that customer segments significantly differ in the way the formation of overall satisfaction can be explained. Beyond the pure identification of segment-specific satisfaction factors, we furthermore show how this fine-granular rating information can be exploited to improve the accuracy of rating-based recommender systems. In particular, we propose to utilize user- and object-specific factor relevance weights which can be learned through linear regression. An empirical evaluation on datasets from different domains finally shows that our method helps us to predict the customer preferences more accurately and thus to develop better online recommendation services.

Keywords Online booking platforms · multi-criteria rating feedback · customer satisfaction · recommender systems

Dietmar Jannach
TU Dortmund, Germany, E-mail: dietmar.jannach@tu-dortmund.de

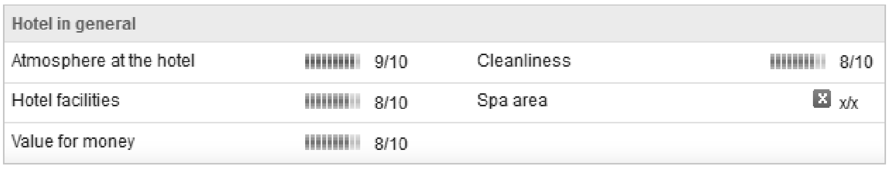
Markus Zanker
Alpen-Adria-Universität Klagenfurt, Austria, E-mail: markus.zanker@aau.at

Matthias Fuchs
Mid Sweden University, Östersund, Sweden, E-mail: mathias.fuchs@miun.se

1 Introduction

The World Wide Web has become the major source of information for customers in the travel and tourism domain and the importance of travel websites and online booking platforms has continually increased over the last years. Today's assumedly largest travel web site, TripAdvisor, reports in 2013 to have more than 200 million unique visitors per month and provides information about more than one million accommodations[1]. The probably most valuable information on this platform, however, that is also prominently reported, are the more than 100 million customer reviews and opinions shared by their user community that have a measurable impact on the customers' decision processes [12].

However, beside plain-text reviews, all of today's major travel and tourism sites including TripAdvisor, Booking.com, Expedia.com or HRS.com allow customers to formulate structured feedback on the accommodation, the booked travel packages or the destination itself in terms of multi-criteria ratings. Figure 1 depicts an example of the rating values for the different criteria for an arbitrary hotel at HRS.com



Hotel in general			
Atmosphere at the hotel	9/10	Cleanliness	8/10
Hotel facilities	8/10	Spa area	☒ x/x
Value for money	8/10		

Fig. 1 Part of the multi-criteria ratings at HRS.com.

Beside the usual overall rating, this more fine-grained feedback should allow customers to identify the strengths and weaknesses, e.g., of a certain hotel, more quickly without reading the reviews in detail. Furthermore, since customers might have different preferences and perceptions of what is important for them, multi-criteria ratings let customers assess in a more efficient way, whether or not a tourism offer matches their expectations.

Multi-dimensional feedback of one customer is however not only an important piece of information for other customers. Also the providers of the travel or tourism web site can exploit this information in different ways. First, the available rating data can be analyzed with respect to the relative importance of different quality factors[2] for different customer groups. This information can then be forwarded to the actual tourism service providers, who can react according to this feedback. Based on this information they can, for example, improve different aspects of their service or change their service to better match their target customers' expectations.

[1] `http://www.tripadvisor.com/PressCenter-c4-Fact_Sheet.html`, retrieved April 2013.

[2] Also termed "quality domains" in the literature.

On the other hand, rating information can be used to automatically filter and rank the often large number of available options[3] in a personalized way with the help of recommender systems (RS). Recommender systems are nowadays part of many e-commerce sites and are designed to help online customers finding relevant and interesting items within large product assortments [20]. A popular class of such systems is based on “collaborative filtering” (CF), which uses the explicit ratings of a larger user community to make predictions about the relevance of items the user has not seen before. Given that these detailed ratings of modern travel and tourism platforms carry more information about user preferences than single overall ratings alone, special algorithms have been proposed, e.g., in [1], that take these additional pieces of information into account in the recommendation process.

In this paper, we start from an empirical exploration of data harvested from a popular travel platform which gives us new insights on the relative importance of different rating criteria. In addition, we present an algorithmic approach to better exploit multi-criteria ratings in recommender systems. The contributions of the paper as thus as follows.

- First, we analyze the multi-criteria ratings and dig deeper in order to see if traditional customer segments differ in their expectations and requirements with respect to the available quality factors. Our analysis is based on a data set from the TripAdvisor platform and reveals significant differences between these analyzed segments. The insights of this study can thus serve as a basis for further investigations with respect to segment-oriented adaptation of the travel or tourism offerings but also of the corresponding travel and tourism information systems.
- Second, we propose a new automated recommendation technique that takes the detailed multi-criteria rating information into account when calculating suitable booking proposals for online customers. The particularity of our approach is that we estimate user- and item-specific importance weights for the different quality factors from the rating data through linear regression and combine the resulting models. A comparative evaluation using datasets from different domains gives evidence that these models help us estimate the customer preferences more accurately and thus to make better online recommendations when compared with previous multi-criteria based techniques as well as state-of-the-art recommendation techniques based on matrix factorization.

2 Analysis of satisfaction factors

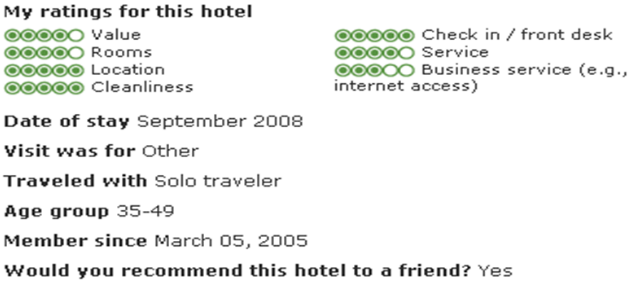
In this section, we aim to analyze if individual customer segments use different “weights” for particular quality dimensions when they determine their overall assessment (rating) [25].

[3] For presentation purposes, we will limit our discussion to hotels and not general tourism offerings. The analysis and algorithms presented later on are, however, not limited to accommodation services.

2.1 Data set and customer segments

We base our analysis on data from the TripAdvisor platform, which contains both detailed rating information as well as demographic information about users and the context of their trips, e.g., whether they are traveling solo or with their family. The dataset, which we obtained in January 2010 through a web crawling process, comprises ratings of 62,290 different users for hotels located in 14 global metropolitan destinations, such as London, New York or Singapore.

On TripAdvisor.com, customers can rate items in 7 different dimensions: value for money, quality of rooms, location of the hotel, cleanliness of the hotel, quality of check-in, overall quality of services and particular business services. In addition, users can provide an uni-dimensional overall satisfaction rating for the hotel (not depicted in Figure 2). These standardized evaluation items are consistently measured on the base of a 5-point scale, from excellent to terrible. Finally, users are explicitly asked if they would recommend the hotel to a friend. This customer-based recommendation to visit a hotel is measured by a separate binary rating (recommend: yes/1; no/0). Exemplarily for one customer, Figure 2 depicts such a detailed item rating as well as basic demographics and context parameters.



My ratings for this hotel

●●●●○ Value
●●●●○ Rooms
●●●●● Location
●●●●● Cleanliness

●●●●● Check in / front desk
●●●●○ Service
●●●○○ Business service (e.g., internet access)

Date of stay September 2008

Visit was for Other

Traveled with Solo traveler

Age group 35-49

Member since March 05, 2005

Would you recommend this hotel to a friend? Yes

Fig. 2 Detailed view on a user rating on TripAdvisor.com

Segment	Description	N	Share
Senior couples	Aged above 50, on a leisure trip, staying with spouse in 4-star or 5-star hotel.	1,284	8.8%
Business tourist solo	Aged between 35 and 50, business trip, staying alone in 4-star or 5-star hotel.	1,366	9.3%
Budget family tourist	Aged between 35 and 50, leisure trip, staying with partner & children in 0-3 star hotel.	2,302	15.7%
Youth tourists & friends	Aged below 25, leisure trip, staying with friends in 0-3 star hotel.	875	6%

Table 1 Tourist segments at TripAdvisor.com

Based on the demographics and the travel context parameters each review on the platform can be assigned to a certain travel segment. We clustered the reviews into four major tourist segments and subsequently focused our analysis on the most traditional tourist segments shown in Table 1.

2.2 Relative importance of quality domains

The goal of the subsequent data analysis is to identify empirical relationships between the overall ratings, the ratings of the perceived value, and the users' assessments of the detailed rating criteria. As suggested in the literature, such an analysis of potential relationships can be done using linear structural equation models (SEM), a statistic technique that allows us to incorporate falsifiable causal assumptions into the model [48]. These hypothesized relationships between model variables (i.e. independent = exogenous, intervening = mediator variables, dependent = endogenous variables) can be tested against empirical data in order to determine how well the SEM model fits the data [26]. Path-coefficients are measuring how strongly exogenous and mediator variables influence the endogenous variable(s) [41].

Figure 3 shows a linear structural equation model (SEM) for the given problem setting, which contains six exogenous rating variables (room quality, cleanliness, service, business services, location and check-in), two mediator variables (value and overall rating), as well as one endogenous variable (willingness to recommend).

The explanatory power of such a model is measured using the Coefficient of Determination R^2, which corresponds to the proportion of variability in the data accounted for by the model. Overall, our model shows high explanation power. Regarding the mediator variables, we observe $R^2 = 60\%$ for the *value* variable and $R^2 = 78\%$ for the variable *overall rating*. For the endogenous variable *willingness to recommend*, we observe $R^2 = 54\%$ for the coefficient.

The path-coefficients $\beta_{Std.}$, which express the strength of the influence of individual variables on others in SEM models, are shown as arrow labels in Figure 3. Looking into the details of the analysis, we can in particular observe that the three hotel quality domains room quality, cleanliness, and service are the strongest single drivers behind the overall rating yielded by hotel guests, which, in turn, significantly affects the willingness to recommend. The path-coefficients of these most influential factors are as follows.

- Room quality (rating-overall: $\beta_{Std.} = .28$; rating-value: $\beta_{Std.}. = .33$)
- Cleanliness (rating-overall: $\beta_{Std.} = .13$, rating-value: $\beta_{Std.} = .10$)
- Service (rating-overall: $\beta_{Std.} = .21$, rating-value: $\beta_{Std.} = .19$)

2.3 Identifying the relative importance of quality factors for tourist segments

Furthermore, we identified the relative level of importance (determination) of the above mentioned seven hotel quality domains on the overall assessment

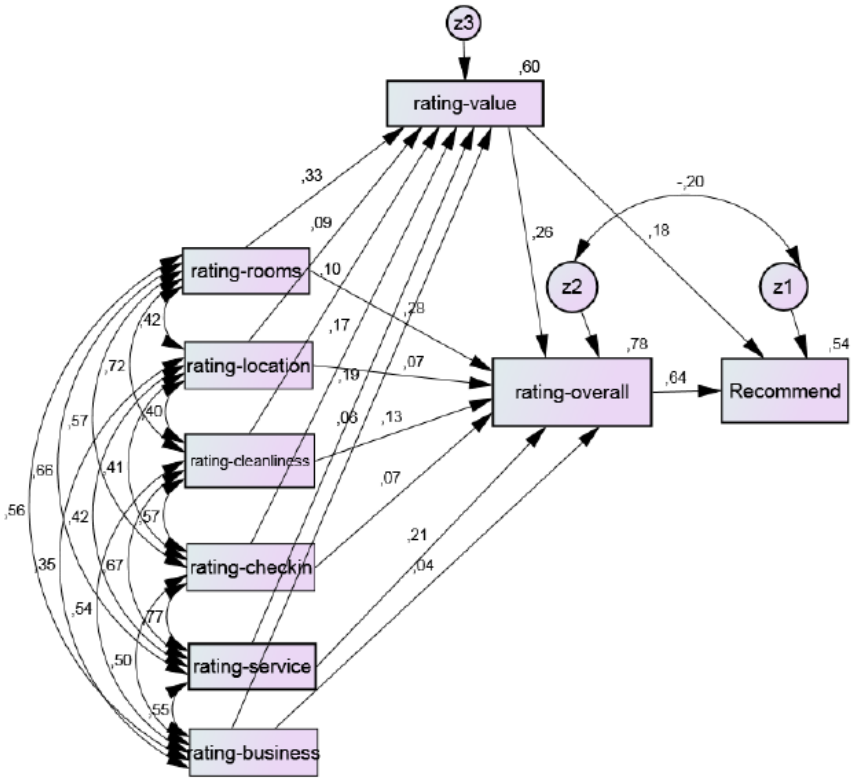


Fig. 3 Structural Equation Model of TripAdvisor Data

through multiple regression [14, 49]. In order to obtain insights on potential differences between the customer segments, we applied the measurements for all four segments listed in Table 1 individually.

Table 2 shows how the *overall rating* is statistically determined through multiple regression models exploiting the 7 quality domains as independent variables and indicating the different perceptions by the four major tourist segments found in our TripAdvisor dataset (i.e. senior tourist couples, business tourist solo, budget family tourist, and youth tourist & friends). Technically, a high and significant Beta (i.e. T-Value ≈ 2) indicates strong impact on the *overall rating* for the respective customer segment.

All models show a strong explanation power (Adj. R^2), are statistically significant (F-Value), and are free of auto-correlated residuals (Durbin Watson) and multi-correlated variables. The magnitude of multicollinearity is typically analyzed by the size of the Variance Inflation Factor (VIF) for each estimated regression coefficient. A common rule of thumb is that if the VIF is greater than 5, then multicollinearity is high [28]. The empirically obtained VIF values shown in Table 2 range between 1.175 and 3.272 and the magnitude of

A-priori	Segment 1:			Segment 2:		
Segments	Senior Tourist Couples			Business Tourist Solo		
	Adj. $R^2 = .787$ $F = 114.21$			Adj. $R^2 = .807$ $F = 260.24$		
	$DW = 1.91$			$DW = 1.85$		
Quality domains	Beta	T-Value	VIF	Beta	T-Value	VIF
Value	**0.384**	**8.274**	2.169	**0.332**	**9.211**	2.927
Rooms	0.247	4.861	2.610	0.271	7.386	3.031
Locations	0.039	1.157	1.175	0.091	3.788	1.273
Cleanliness	0.128	2.692	2.276	0.122	3.264	3.138
Check-in	0.081	1.755	2.089	0.048	1.475	2.386
Service	0.096	1.742	3.057	0.161	4.242	3.221
Business	0.178	4.427	1.625	0.252	2.885	1.718

A-priori	Segment 3:			Segment 4:		
Segments	Budget Family Tourist			Youth tourist & friends		
	Adj. $R^2 = .769$ $F = 183.36$			Adj. $R^2 = .698$ $F = 45.99$		
	$DW = 2.23$			$DW = 2.19$		
Quality domains	Beta	T-Value	VIF	Beta	T-Value	VIF
Value	**0.444**	**11.026**	2.687	0.266	3.442	2.692
Rooms	0.203	4.684	3.103	**0.459**	**5.411**	3.245
Locations	0.055	2.010	1.236	0.185	3.443	1.308
Cleanliness	0.179	4.342	2.820	0.081	0.947	3.272
Check-in	0.107	1.991	1.990	0.021	0.287	2.458
Service	0.044	1.077	2.760	0.131	1.662	2.801
Business	0.058	1.952	1.455	0.093	1.514	1.684

* Remarks:

– The (adjusted) Coefficient of Determination R^2 is the proportion of variability in data accounted for by the statistical model;
– Beta is a measure of how strongly a predictor influences the dependent variable;
– An F- or a T-test are statistical tests in which the test statistic has an F or a T-distribution under the null hypothesis.
– The null hypothesis is rejected if the F or T-value calculated from the data is greater than the critical value of the F- or T-distribution for some desired false-rejection probability (e.g. 0.05).
– The Durbin Watson (DW) Test detects autocorrelation (i.e. residuals from a multiple regression model are independent).
– The Variance Inflation Factor (VIF) quantifies the degree of multicollinearity (i.e. correlated predictor variables) in regression analyses.

Table 2 Multiple Regression Results - Determination of Overall Assessment by Partial Hotel Quality Domains

multicollinearity can thus be considered to be low. Therefore, the TripAdvisor data seems to be appropriate for being used to identify how the various hotel quality domains determine the overall assessment among different customer segments ([14]).

The results, for instance, show that a relatively strong and most general determination across all segments stems from the quality factors "value for money" and "room quality", see the Beta and T-values in Table 2. For young tourists, the factor "room quality" appears to be the most important one. For the budget family tourist, on the other hand, the "value-for money" aspect emerges as the most critical quality domain.

The remaining quality domains are playing quite different roles when determining the overall quality assessments for customers in different segments as shown in Table 2. A convenient “business environment”, for example, seems to be a crucial quality domain both for business tourists as well as for senior tourist couples. At the same time, “location” only plays a minor and insignificant role. “Service quality” is, not so surprisingly, of particular importance for business tourists. For other customer segments, this aspect seems to have less relevance. The “cleanliness of the hotel” is quite important for budget family tourists (ranked third); the “location factor”, in contrast, determines the overall quality assessment of young tourists significantly stronger than in other customer segments.

2.4 Applying the Penalty-Reward-Model

Beside the purely quantitative role of quality domains in determining the overall assessment, the literature also discusses their relevance from a qualitative point of view [6]. Since the 1990s, researchers have begun to tackle problems related to the empirical analysis of service quality perception with a multi-factor structure model of customer satisfaction [23]. This model has been adopted and empirically validated both in a service marketing and tourism context, see [8,9,35,36] and [38]. The three-factor structure of customer satisfaction was originally defined by Kano in [24]. Based on his model, quality attributes can be grouped into three categories, each of which exerts a different impact on customer satisfaction:

- *Basic factors* are minimum requirements that cause dissatisfaction if not fulfilled, but do not lead to customer satisfaction if fulfilled or exceeded. Negative performance in these quality domains has a greater impact on overall satisfaction than a positive one. Hence, basic factors are expected by the customer (i.e. regarded as prerequisites).
- *Excitement factors* are factors that increase customer satisfaction if delivered, but do not cause dissatisfaction if they are not delivered. Thus, only positive performance on these quality dimensions has an impact on the overall satisfaction.
- *Performance factors* lead to satisfaction if performance is high and lead to dissatisfaction if performance is low. In this case, the relationship between the attribute performance and overall satisfaction is linear and symmetric [9].

2.4.1 Problem encoding

In order to decipher the factor structure of customer satisfaction in the hotel booking domain we applied Brandt’s [5] Penalty-Reward-Contrast analysis method on the TripAdvisor dataset. The method employs a dichotomized regression analysis using dummy variables [14]. One set of dummy variables represents the excitement factors in quantitative form, while a second set expresses the basic factors. In order to carry out the analysis using our TripAdvisor data, we recoded the 5 point scales of the given ratings (i.e. the independent variables) which range from 5=excellent to 1=terrible in a way where scores of 5 correspond to a value of 1 for the dummy variable representing the *excitement factor*. Comparably low rating scores of 1, 2, and 3 were translated into the value 1 for a second dummy variable representing the *basic factor*[4]. Finally, empty cells of both dummy variables were recoded with a value of zero.

With the help of this recoding multiple regression analyses were carried out to quantify basic requirements and excitement factors using the *overall rating* assessment as the dependent variable and the two dummy variables for each of the seven quality domains as independent variables. “Penalties” can now be expressed as the incremental decline associated with low levels of satisfaction, while “rewards” become expressed as the incremental increase associated with high satisfaction for a certain hotel quality domain.

The results obtained from the Penalty-Reward-Contrast-Analysis can be interpreted as follows. If penalty levels surpass reward levels, the respective quality domain represents a basic factor. If, on the other hand, the reward index surpasses the penalty value, the quality dimension can be interpreted as an excitement factor. Finally, if the reward and penalty values are rather similar, the quality domain will contribute to tourist satisfaction only when its level of performance is high. At the same time, the quality factor will lead to dissatisfaction in case the performance level is low.

2.4.2 Observed results and discussion

When applying the Penalty-Reward approach to the seven quality factors in the TripAdvisor dataset the results for the 4 customer segments are as shown in Figure 4.

As in the previous tests, all multiple regressions show a strong explanation power, are statistically significant, free of auto-correlated residuals and multi-correlated variables[5]. Therefore, again, the TripAdvisor data seems to be suitable for identifying the factor structure of customer satisfaction with hotel quality domains among different customer segments [14].

Figure 4 has to be interpreted as follows. If the reward index for satisfaction ($+\beta$) is significant and surpasses the penalty value, the quality dimension can be interpreted as an excitement factor. By contrast, if the penalty levels for satisfaction ($-\beta$) are significant and surpass reward levels, the respective quality domain represents a basic factor. Finally, if reward and penalty values are rather similar, the quality domain contributes to tourist satisfaction only when its level of performance is high (i.e. performance factor).

[4] This simple encoding approach shows some degree of arbitrariness. However, as the empirical distribution of the raw data is also taken into consideration, the approach is recommended in the literature, e.g., in [6,9,35] and [36].

[5] Adjusted R^2 values are between 0.681 and 0.723, F-values range from 74.28 to 429.24, Durbin Watson is between 1.87 and 2.02, Variance Inflation Factor between 1.224 and 2.087

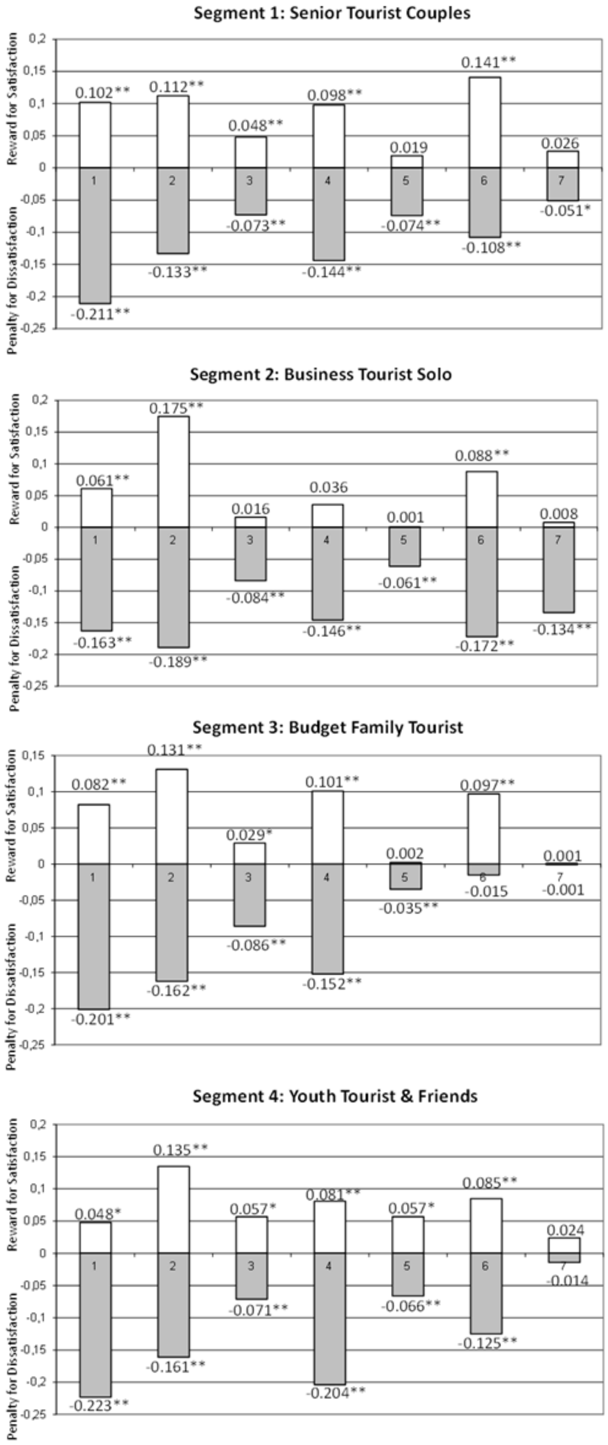


Fig. 4 Penalty Reward Contrast Analysis for TripAdvisor Data. Indices are regression coefficients. Significance levels: 5%*, 1%**. 1: Value, 2: Rooms, 3: Location, 4: Cleanliness, 5: Checkin, 6: Service, 7: Business

Somewhat surprisingly, for none of the customer segments, a clear *excitement factor* could be identified, except for the “general service quality” dimension: in the segment “budget family tourists” the positive reward index clearly surpasses the negative penalty value. However, all remaining TripAdvisor rating dimensions can be classified either as *basic* or *performance* factors.

The following more detailed observations can be made. For all customer segments, “room quality” has a comparably strong potential to increase overall satisfaction if its performance level is high. For senior tourist couples, also the “general service quality” shows similar potentials. On the other hand, the quality factors “value for money”, “room quality”, and “cleanliness” can lead to significantly lower overall assessments if their performance is low. The business tourist segment has a different focus. Here, poor performance on “business convenience” and “general service quality” can easily lead to strong penalties in the overall assessment. Finally, for budget family tourists and also the young tourist segments, low performance levels with respect to “room quality”, “cleanliness” but also the “location factor” make them assign lower overall scores for a hotel.

Overall, our exploratory analyses revealed both significantly differing factor importance weights (Tables 2) and penalty-reward profiles (Figure 4) for the examined customer segments. Being aware of these empirical phenomena can be particularly valuable, as customers implicitly might use weighting schemes when they assess the overall quality of the travel or tourism offering [8, 25].

With respect to the results related to the factor importance weights, we could observe that the 7 rating dimensions capture most of the signal that determines the overall rating value, as the adjusted R^2 value is clearly above 0.75 for most regression models in Table 2. However, the relationship between multi-criteria ratings and the overall rating is moderated by the tourist segments, which we determined with the help of user profile data and the travel context. Thus, the relative influence of the specific rating dimensions varies for different customer segments.

The penalty-reward-contrast analysis also unveiled differences between the studied individual customer groups. The results showed that high or low performance values for certain quality dimension can have a significantly different impact on the customers’ overall assessment.

2.5 Summary

Table 3 summarizes the insights gained so far from our empirical analysis. These findings can not only be leveraged to build more accurate recommendation services as will be shown in next section but they also help to give service providers a better understanding of what is important for their customers and where there is room for service improvements.

As a side note we would like to mention that TripAdvisor has modified its rating criteria since data extraction for this paper took place. The two dimensions *Check-in* and *Business Services* have been deleted, while the dimension

Sleep Quality has been introduced. This partly corresponds to our findings (Figure 4) as Business Services only insignificantly influence the overall rating assessments for the customer segments 3 and 4. However, in our analysis the *Check-in* dimension has been shown to be a basic factor for all four customer segments.

Analysis	**Main findings**
Linear SEM	*Room quality*, *cleanliness* and *service* are the strongest single drivers behind the perception of *value for money* and the *overall rating.*
Segment Analysis	Hotel quality domains play quite different roles when determining the *overall rating* for different segments, e.g. *room quality* is most important for youth tourists while *value for money* dominates for budget family tourists.
Penalty-Reward Model	*Room quality* is a performance factor for all segments; *value for money* and *cleanliness* are mostly basic factors and *general service quality* is the sole excitement factor (but only for segment budget family tourists).

Table 3 Summary of satisfaction analysis

3 Multi-dimensional rating feedback for recommendation

In the first part of this paper, we showed how existing rating data can be used to identify customer group-specific weighting schemes. In our approach, the weights were empirically determined by regression models explaining the overall rating through the quality domains behind it.

In the second part of the paper we will demonstrate how we can make use of these insights to further increase the accuracy of recommender applications in tourism. In particular, we will demonstrate that we can apply regression type models to identify item and customer-specific weighting schemes in order to improve the information filtering and recommendation services on travel websites that target individual customers.

Thus, we propose to use the multi-criteria ratings to learn factor importance weights from the data for each user and hotel individually[6]. Then we aim at combining the resulting models in a weighted approach to achieve higher accuracy when recommending hotels to customers. Thus, instead of determining the aggregate weights for a specific customer segment as described in the previous section, we follow an even more fine-granular automated approach and estimate the relative weights for each customer individually.

[6] In the following, we will use the term "rating" when we refer to a customer's known or estimated quality assessment for a hotel or its individual quality factors. The assessments for the quality factors are termed "multi-criteria ratings", as this term is more common in the recommender systems literature.

3.1 Multi-criteria collaborative filtering recommender systems

In traditional collaborative filtering (CF) recommender systems, the only input to the system consists of item ratings of a larger user community [20]. The rating scales typically range from one to five, as in the case of TripAdvisor and Amazon.com. The task of a corresponding algorithm usually consists of predicting ratings for the items that the customer has not seen yet and which represent the potential recommendations. These unseen items can then be ranked according to the predicted rating value, i.e., the expected overall quality assessment of the customer.

Generally, the goal of a CF recommender is to estimate a rating function R: $Users \times Items \rightarrow R_0$, where R_0 is a totally ordered set, typically consisting of real-valued numbers ranging between the lowest to the highest possible rating value. While the input for learning the rating function R in the single-rating case is a usually sparse user-item rating matrix, we additionally assume that we know the detailed ratings in the multi-criteria case. Table 4 shows an example for such a multi-criteria rating database.

Row	User	Item	**Overall**	Value	Rooms	Location	Cleanliness
1	u1	i1	**4**	3	3	5	5
2	u1	i2	**3**	2	3	4	2
3	u2	i1	**4**	5	4	2	3
4	u2	i2	**5**	5	5	3	4

Table 4 Multi-criteria rating database fragment for hotels.

When examining the fictitious example in Table 4 in more detail, we can make the following observations. For user $u1$ (row 1 and row 2), it seems that the overall rating roughly corresponds to the average of the detailed ratings for the hotel. In contrast, user $u2$'s overall assessment seems to be biased towards *Value* and *Rooms* and he gives high ratings even when his assessments for *Location* and *Cleanliness* are comparably low. Thus, the first dimensions might be more important for this user and one could try to recommend hotels, which also obtained high ratings by other users in these dimensions.

In [1], two basic schemes for making rating predictions based on multi-criteria ratings are proposed. The general idea of these schemes can be summarized as follows.

1. *Similarity-based approaches*: In traditional neighborhood-based collaborative filtering algorithms, the first task is to find a set of like-minded users (also called neighbors or peers) for a *target user u*, for whom a rating prediction is sought for. This is usually done by comparing the ratings of the users with the help of some similarity function or a correlation measure. The rating prediction can then be made by combining the ratings of u's peers in a weighted approach. The idea of similarity-based multi-criteria approaches is to retain the usual prediction function but use a more fine-granular and multi-dimensional similarity function, which also considers

the detailed rating information. Looking at the example in Table 4, we observe that both users gave a four-star rating for hotel $i2$. A closer look, however, reveals that they might have assigned this rating for different reasons and the users appreciated different aspects of the hotel. Thus, the interest and preference similarity of these two users might be not as high as the overall ratings suggest.

2. *Aggregation-function based approaches*: This class of techniques consists of two steps. In a first step, the detailed item ratings for the different criteria are estimated for an unseen item. This can, for example, be done by considering the ratings for each dimension as an individual recommendation problem. Therefore, any existing single-rating recommendation algorithm can be applied, e.g., to predict the rating for the *Location* factor for a given user-item pair. In the second phase, the estimated criteria ratings are combined with the help of an aggregation function f to generate the prediction for the overall rating, i.e., $R_0 = f(R_1, ..., R_k)$. In the simplest form, f could simply return the average of the input values. However, another, more promising approach on which we also base our work, is to learn the combination weights for f from the available data.

In [1], Adomavicius and Kwon propose a method of the latter type of techniques and approximate the function f for each item i using multiple linear regression. The overall rating R_0 can thus be viewed to be dependent on a linear combination of the criteria ratings, where each criterion is assigned a weight w_i, that is $R_0 = w_1R_1 + ... + w_kR_k + c$ where the weights w_i and the constant c are estimated from the data.

For our the TripAdvisor dataset, the prediction function f for a certain hotel could for example look like the one shown in Equation 1. In the example, the weight factor learned through regression from the data for the *Rooms* aspect of the hotel is lower than the other aspects.

$$R_0 = 0.12 * Value + 0.08 * Rooms + 0.19 * Location + \cdots + c \qquad (1)$$

Once these weight factors are learned, we furthermore need an estimate of the user's ratings for the different quality dimensions like *Value* or *Location*. Since these ratings (R_1 to R_k) are also unknown, the idea is to estimate them from the data as well. This can be accomplished by viewing each quality dimension as a recommendation problem and applying any standard collaborative filtering algorithm.

In [1], an experimental evaluation of the different techniques was conducted using a comparably small and dense dataset from the movie domain. The experiments showed that multi-criteria recommendation approaches and in particular aggregation-function based approaches using regression can outperform traditional baseline techniques such as the above-mentioned nearest-neighbor approaches in terms of their predictive accuracy.

3.2 Proposed enhancements

In this section, we propose different enhancements to the regression-based approach from [1], which target not only on accuracy improvements, but should also help us to deal with the often very sparse data situation in the tourism domain.

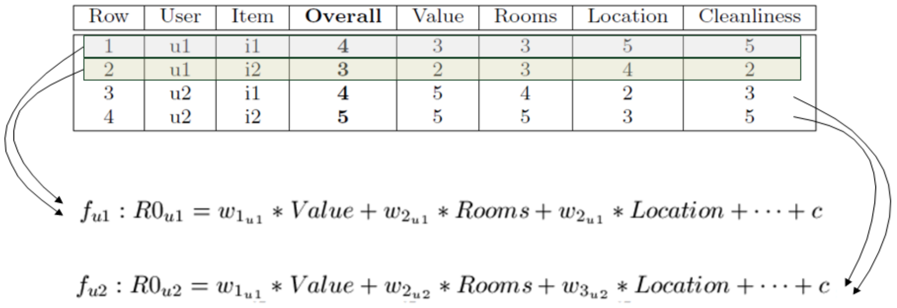
1. *Combining user- and item specific models:* A particular aspect of the work of [1] is that they do not rely on one single set of "global" weights, but learn such weights for each individual item as shown in Figure 5. However, given the empirical evidence from the first part of this article, we additionally propose to learn regression functions per user and combine the predictions in a weighted approach where factor weights are automatically determined for each user and item through an optimization procedure.
2. *Applying feature selection:* In the tourism domain, the number of quality factors for which the user can provide ratings, can be comparably large. However, not all of these factors might be relevant as shown in the explorative analysis of the Tripadvisor data previously presented in this paper. Therefore, we propose to apply a feature selection procedure to factor out rating dimensions which carry only little information and might introduce noise.
3. *Using Support Vector Regression:* Instead of using least squares regression as done in [1], we propose to use Support Vector Regression (SVR) [7], because this technique is also applicable when there are very few data points and many coefficients to be determined, which is a typical situation, e.g., for hotel booking platforms. Furthermore, SVR has a limited tendency of overfitting and has been successfully applied before to solve recommendation problems, e.g., [11] or [46].

3.3 Combining user- and item-specific models

The empirical analysis in Section 2 in this paper has clearly shown that different customer groups can be considered as quite heterogeneous with respect to which quality factors are most important for them. Therefore, it would be in fact more intuitive to learn preference weights rather per user or user groups than per item as has been done and suggested in previous works [1].

Learning such user-based regression models can be done in the same way as for the item-based approach. The only difference is how we split the available data. In our example in Table 4, we would use row 1 and row 2 to learn the preference weights for user $u1$ across all hotels he or she has visited and rated so far. Row 3 and row 4 would correspondingly be the input to the regression problem for user $u2$.

The regression models can be learned in an offline training phase. At the end, we obtain one model for each user. Learning these possibly many models is not particular time consuming, since the number of input data points per model are comparably small, e.g., 3 to 20 ratings of a particular user. The



Row	User	Item	**Overall**	Value	Rooms	Location	Cleanliness
1	u1	i1	**4**	3	3	5	5
2	u1	i2	**3**	2	3	4	2
3	u2	i1	**4**	5	4	2	3
4	u2	i2	**5**	5	5	3	5

$$f_{u1} : R0_{u1} = w_{1_{u1}} * Value + w_{2_{u1}} * Rooms + w_{2_{u1}} * Location + \cdots + c$$

$$f_{u2} : R0_{u2} = w_{1_{u1}} * Value + w_{2_{u2}} * Rooms + w_{3_{u2}} * Location + \cdots + c$$

Fig. 5 Learning regression functions per user.

small number of data points per model however led us to the choice of Support Vector Regression as a learning technique, since techniques like Ordinary Least Squares for example require that there are at least as many data points as there are dimensions. SVR is based on the same principles as Support Vector Machines (SVM), a popular supervised machine learning technique used for classification tasks [45]. Like SVMs it is based on projecting the original data into higher-dimensional space and, in the case of regression, do a linear regression in this space, which corresponds to a non-linear regression in the ordinary space [39].

First experiments indicated that relying only on user-based models does not lead to satisfactory results for those users who have only rated very few items. The number of available ratings per hotel is usually higher than for users. Therefore, since both aspects might be relevant and contain relevant information, we propose to learn both user- and item-based regression models and combine their predictions as shown in Equation 2.

$$\hat{r}_{u,i} = w_u * \hat{r}^{user}_{u,i} + w_i * \hat{r}^{item}_{u,i} \tag{2}$$

The estimated overall rating $\hat{r}_{u,i}$ is thus computed as a weighted sum of the user-model prediction $\hat{r}^{user}_{u,i}$ and the item-model $\hat{r}^{item}_{u,i}$. One difference to existing approaches is that we propose not to use global weights or a static weighting scheme as in [18] but rather try to learn a weight parameter for each user and item that minimize the difference between the predicted and the true rating for the given data.

Technically, we apply an optimization procedure that is similar to the gradient descent procedure that is also used in modern matrix factorization based recommendation algorithms [10,27]. After assigning initial default values to the weight parameters, we iterate over each user/item rating tuple in the training data and compute the system's current prediction by calculating a weighted combination of the user-specific and item-specific regression functions. The difference between the prediction and the true rating is then used to adapt the weights to better fit the data. A detailed listing of the weight learning algorithm and the optimization goal is given in the Appendix; further details about the algorithmic approach are given in [19].

Generally, even though assessing the relative importance of, e.g., the cleanliness of a hotel, based on only very few data points of an individual user might not be very reliable, our empirical evaluation shows that these models do not hurt the accuracy when combined with the item-based model. However, to obtain even more accurate results for such cold-start users, we currently investigate the user of clustering techniques. In such an approach, we would group users, items, or ratings into clusters and learn regression functions for each cluster.

3.4 Feature selection

The number of available criteria ratings in the tourism domain can be relatively high. In the data set which we obtained from HRS.com, for example, criteria ratings for up to twenty different dimensions are available[7]. While all these detailed rating can carry valuable information, our hypothesis is, however, that it might be advantageous to use only a subset of the data in the prediction process. In particular, given the comparably high effort for the end user to fill out the relatively lengthy feedback form, users might be tempted to fill out the forms not very carefully, which might introduce noise into the data. At the same time, there might be dimensions which are misinterpreted by the user. Thus, using a larger number of features could finally lead to overfitting effects, such that the learned regression models are optimized for the historical data, but are not general enough to work well for new visitors.

Feature selection is a common practice in various applications of machine learning. In our context it means that we simply do not take certain quality dimensions into account in the learning process and only retain those ratings in the dataset that are related to quality dimensions which we assume to be particularly relevant. In principle, the selection of relevant features can be done by a domain expert. We are, however, interested in an at least partially automated process through which the "optimal" (or at least "sufficiently good") subset of features is identified. A basic strategy could be to determine the correlation between the overall rating of a hotel with the ratings of each of the quality dimensions. If we observe that, e.g., the rating for *Rooms* strongly correlates with the overall rating for the hotel, we can assume that this quality dimension was important for the customer and should be retained in the learning process. More details and a classification of different systematic or heuristic approaches to feature selection can be found in [13][8].

In order to find a good subset of features (i.e. rating dimensions) for the hotel booking domain, we evaluated the following three strategies.

[7] The web site is regularly updated such that the number of rating dimensions varies over time.

[8] An alternative idea to find the most important factors in the data and to avoid noise could be to apply Principal Component Analysis.

- *ST1*: In this strategy, we first order the individual rating dimensions based on their estimated relevance or influence on the overall rating. In particular, we use the chi-square statistic [31] as a measure of relevance; other relevance or correlation metrics are possible, but did not lead to largely different feature rankings in our experiments. Then, we incrementally add one feature after the other to the training data, make predictions for the test data and measure the prediction error. At the end, we determine the subset of features that leads to the smallest prediction error.
- *ST2*: This strategy is similar to ST1, but this time we remove individual features from the current set when we observe that they lead to a deterioration of the prediction error.
- *ST3*: Beside the straightforward incremental strategies ST1 and ST2, we also made experiments with an optimizing technique based on an Evolutionary Algorithm (EA). We used the methods available in the RapidMiner[9] toolkit to determine "good" feature sets through heuristic optimization [37].

3.5 Experimental Analysis

We conducted different experimental analyses using historical rating data from two different domains. The goal of the evaluation was to assess to which extent individual importance weights for the different quality factors can help us to generate more accurate recommendations.

We rely on a typical experimental evaluation design used, e.g., in the fields of Information Retrieval and Recommender Systems where the datasets are split into a training and a (hidden) test dataset. Then, given the data in the training dataset, the task for a recommender system is to predict the hidden ratings in the test dataset or to rank them according to their predicted relevance for an individual user. To assess the quality of the rating predictions, the aggregate prediction error can be calculated. Similarly, measures exist to assess the generated personalized item rankings.

3.5.1 Datasets

The two datasets from the tourism domain contain multi-dimensional rating feedback from hotel booking platforms (TripAdvisor, HRS.com). A third data set from a different domain – ratings from the movie platform Yahoo!Movies – was used for the purpose of demonstrating the external validity of the proposed techniques that can also be applied in other application domains.

- TripAdvisor: The dataset has been already described in Section 2, based on which we made the explorative analysis of the weighing schemes related to the satisfaction factors (quality domains). The sparsity of this dataset is comparable to the HRS dataset discussed next.

[9] http://www.rapidminer.com

- HRS.com: The dataset contains multi-dimensional ratings for up to 20 criteria, which are provided on a 1-10 scale. In addition, an overall rating on a 1-3 scale is given. Compared to previous works on multi-criteria RS in which the data is pre-processed and filtered [1], our real-world data set is extremely sparse and the number of ratings per user and item is very low.
- Yahoo!Movies: A dataset which we obtained from the Yahoo!Movies website through a crawling process. The dataset contains an overall rating for movies as well as sub-ratings for four dimensions (Story, Acting, Direction, Visuals). We transformed the ratings which were given on a 13-point rating scale (A+ to F) to the usual 1-5 rating scale to make our work comparable with previous works who used a similar dataset.

In order to evaluate how the density of the dataset influences the prediction accuracy we created subsamples for the HRS, TripAdvisor (TA) and Yahoo!Movies (YM) datasets, in which we varied the minimum number of ratings per user and item. The dataset characteristics are summarized in Table 5. The dataset names indicate the minimum number of ratings per user and item. HRS-5-5 for example means that each user in the dataset has rated at least 5 items and that for each hotel at least 5 ratings are available.

Name	#Users	#Items	#Overall ratings
HRS-5-5	1,162	1,203	4,564
HRS-3-3	1,768	1,762	9,712
HRS-RAW	1,582	2,277	10,347
TA-5-5	2,321	2,119	16,907
TA-3-3	13,048	12,342	83,397
TA-RAW	40,970	44,098	137,566
YM-20-20	429	491	18,504
YM-10-10	1,827	1,471	48,026
YM-5-5	5,978	3,079	82,599

Table 5 Dataset characteristics.

3.5.2 Algorithms & evaluation method

Algorithms. We compared the performance of the following recommendation algorithms.

- Single-rating prediction (ignoring multi-criteria ratings):
  1. SlopeOne: A single-rating algorithm proposed in [29], whose performance is comparable to classical nearest neighbor approaches but is less computationally intensive.
  2. FunkSVD: A more recent technique based on matrix factorization[10]. Approaches based on matrix factorization (MF) have shown to lead to

[10] http://sifter.org/˜simon/journal/20061211.html

accurate results in the Netflix prize competition. We also made experiments with Koren's MF approach [27], which led to similar results.

- Multi-criteria rating algorithms:
  1. MC-Similarity: A similarity-based approach as described in Section 3.1. Experiments showed that the worst-case similarity as proposed in [1] worked best for our setting.
  2. LS-Regress-*: This technique corresponds to the aggregation-function based approach from [1], who use Ordinary Least Squares regression. We made experiments both with per-item regression models (LS-Regress-I) and per-user models (LS-Regress-U).
  3. SV-Regress-*: Basically the same as LS-Regress-* with the difference that we use Support Vector Regression as an underlying technique.
  4. WeightedSVM: Our newly proposed method described in Section 3.2 which combines the predictions of SV-Regress-U and SV-Regress-I in a weighted approach.

For the aggregation-function based approaches (i.e. all regression-based ones), we need an additional technique to predict the criteria ratings of the target item as described in Section 3.1. We used a traditional neighborhood-based method with Pearson correlation as a similarity function for all algorithms. Experiments in which we used MF-based approaches for that purpose interestingly led to worse results.

Quality measures. As typical in the literature, we use the accuracy measures root-mean-square-error (RMSE), Precision and Recall to assess the quality of the recommendations of the different algorithms [20]. To measure the RMSE, we randomly split the data into 95% training and 5% test data and repeated the experiments to factor out effects of randomness. The reported values correspond to the average RMSE of 30 evaluation runs.

To determine precision and recall, we used the protocol variant of [40]. In particular, we transform the rating predictions into "like" and "dislike" statements, where ratings above the user's mean rating are interpreted as "like" statements. We then compare the Existing Like Statements (ELS) with the Predicted Like Statements (PLS) returned by the recommender, where $|PLS| \leq |ELS|$. *Precision* is defined as $\frac{|PLS \cap ELS|}{|PLS|}$ and *Recall* is measured as $\frac{|PLS \cap ELS|}{|ELS|}$. Since precision and recall represent a trade-off, we report the harmonic mean of *precision* and *recall*, i.e. the F1 measure $2 \times \frac{precision \times recall}{precision + recall}$, obtained from a five-fold cross-validation procedure. Finally, we also report the *coverage* numbers, where we use a coverage metric that counts the fraction of ratings in the test set for which an algorithm could make a prediction[11].

[11] For the Yahoo!Movies dataset, we also made experiments in which we measured the Mean Absolute Error (MAE) as well as Precision@5 and Precision@7 to compare our work with previous results from the literature. The results are reported in detail in [19].

3.5.3 Accuracy results

HRS dataset. In Table 6, we report the obtained accuracy results for the different HRS datasets in terms of the RMSE. Technically, the RMSE measure aggregates how much the rating value predicted by a recommender system deviates from the true (hidden) rating. Thus, the lower the RMSE value, the better an algorithm is capable of predicting the user's assessment for an unseen item.

The results show that `WeightedSVM`, the proposed weighted combination of user- and item-based regression models, consistently outperforms the other methods for all datasets. The results confirm that the availability of more data leads to smaller errors. Even though the absolute numbers cannot be directly compared because of the different dataset sizes, the best results are achieved with the smallest but most dense dataset `HRS-5-5`. For this dataset, we can also observe that the similarity-based approach `MC-Similarity` outperforms `SlopeOne`, which has a correspondence to the work of [1] who could show that taking into account multi-criteria ratings can be better than using traditional approaches. Our work, however, shows that better accuracy values can be achieved even when compared with more recent matrix factorization techniques.

Coverage numbers are unfortunately not reported in [1]. In all our experiments we observed a coverage problem for the similarity-based approaches, which make them less appropriate for real-world scenarios; the scalability of this neighborhood-based approach is similarly limited.

Algorithm	HRS-5-5	HRS-3-3	HRS-RAW
SlopeOne	0.68 (1.0)	0.71 (0.99)	0.77 (0.72)
Funk-SVD	0.60 (1.0)	0.64 (0.99)	0.66 (0.73)
MC-Similarity	0.65 (0.32)	0.71 (0.12)	0.77 (0.31)
SV-Regress-I	0.59 (1.0)	0.62 (0.99)	0.72 (0.73)
SV-Regress-U	0.61 (1.0)	0.66 (0.99)	0.66 (0.72)
WeightedSVM	**0.52** (1.0)	**0.56** (0.99)	**0.61** (0.73)

Table 6 RMSE results for the HRS datasets; coverage is shown in parentheses.

The approaches based on Ordinary Least Squares regression cannot be computed for this dataset, because they require that there are at least as many data points per user or item as there are coefficients to be estimated in the regression model. In our particular setting, however, we have up to twenty dimensions but for most users and hotels only very few ratings.

TripAdvisor dataset. In case of the TA dataset with the highest rating density (`TA-5-5`), the SVM-based scheme again outperforms the other techniques as shown in Table 7. The similarity-based approach performed worst on all datasets and has a very low coverage. When the sparsity of the data set is increased, the predictions of the matrix factorization technique are, however, either equally accurate as those of the SVM-based method (`TA-3-3`) or even slightly better (`TA-RAW`). This indicates that a hybridization strategy which switches between recommenders depending on the available amount of ratings or combines the different predictions in a weighted approach might be appropriate in this scenario.

Algorithm	TA-5-5	TA-3-3	TA-RAW
SlopeOne	0.99 (1.0)	1.04 (1.0)	1.07 (0.82)
Funk-SVD	0.94 (1.0)	**1.00** (1.0)	**1.01** (0.82)
MC-Similarity	1.20 (0.3)	1.12 (0.07)	1.18 (0.23)
SV-Regress-I	1.00 (1.0)	1.08 (1.0)	1.14 (0.82)
SV-Regress-U	1.03 (1.0)	1.12 (1.0)	1.17 (0.82)
WeightedSVM	**0.91** (1.0)	**1.00** (1.0)	1.05 (0.82)

Table 7 RMSE results for the Trip Advisor datasets; coverage is shown in parentheses.

Movie dataset. When finally measuring the predictive accuracy on the dataset from the movie domain, we can again make the observation that the weighted approach works best in terms of the RMSE, see Table 8. What can however be observed is that all regression- and multi-criteria-based models work much better here than the single-rating approaches, which we attribute to the much denser rating information that is available in this setting.

Algorithm	YM-20-20	YM-10-10	YM-5-5
SlopeOne	0.81 (1.0)	0.89 (1.0)	0.97 (1.0)
Funk-SVD	0.83 (1.0)	0.87 (1.0)	0.91 (1.0)
MC-Similarity	0.87 (0.99)	0.93 (0.56)	0.99 (0.24)
LS-Regress-U	0.65 (1.0)	0.72 (1.0)	0.83 (0.97)
LS-Regress-I	0.70 (1.0)	0.79 (1.0)	0.82 (0.97)
SV-Regress-I	0.66 (1.0)	0.69 (1.0)	0.72 (1.0)
SV-Regress-U	0.60 (1.0)	0.65 (1.0)	0.73 (1.0)
WeightedSVM	**0.57** (1.0)	**0.60** (1.0)	**0.63** (1.0)

Table 8 RMSE values for the Yahoo!Movies datasets.

Table 9 reports the F1 values (i.e. the harmonic mean of precision and recall measures [20]) for the HRS and Yahoo!Movies dataset. Generally, the ranking of algorithms follows the trend of the ranking based on the RMSE. The results for the F1 measure for the TripAdvisor dataset are finally shown in Table 10. Similar to the other datasets, the regression-based approaches work particularly well on this measure across all dataset variations, despite the fact that the matrix factorization method was better on the RMSE measure for the sparse TripAdvisor datasets. Another interesting aspect here is that the performance of the similarity-based approach as well as SlopeOne strongly degrade in these experiments.

Discussion. Our analysis shows that the proposed multi-criteria approaches can predict the user's quality assessment better than previous techniques and can thus serve as a basis for building high-quality recommendation services

Algorithm	HRS-5-5	HRS-3-3	HRS-RAW	YM-20-20	YM-10-10	YM-5-5
SlopeOne	68.40	46.40	8.31	78.47	82.64	87.39
Funk-SVD	85.33	88.36	69.13	78.62	83.30	89.07
MC-Similarity	26.99	13.55	5.91	75.97	52.47	32.87
LS-Regress-I	-	-	-	86.35	88.14	90.07
LS-Regress-U	-	-	-	87.04	88.43	73.66
SV-Regress-I	88.82	90.60	69.72	87.64	89.93	93.37
SV-Regress-U	87.69	89.50	**71.05**	86.35	88.18	91.42
WeightedSVM	**90.39**	**91.67**	71.00	**88.70**	**91.53**	**94.32**

Table 9 F1 values for the HRS and Yahoo!Movies datasets.

Algorithm	TA-5-5	TA-3-3	TA-RAW
SlopeOne	70.6	48.4	40.8
Funk-SVD	87.8	88.8	75.2
MC-Similarity	13.4	2.7	2.6
SV-Regress-I	88.9	89.5	75.7
SV-Regress-U	89.4	89.5	76.0
WeightedSVM	**90.6**	**91.0**	**76.8**

Table 10 F1 results for the TripAdvisor datasets.

on tourism platforms. Depending on the dataset characteristics, the accuracy improvements can be substantial. A deeper and systematic analysis of the factors that influence the accuracy of the different techniques as done, e.g., in [3] has however not been done so far.

3.5.4 Features selection results

In a further set of experiments we finally tried different feature selection strategies as described in Section 3.4. In Table 11, we show the results when applying the incremental strategy ST1 on the HRS dataset. To apply the strategy, we first ranked the rating dimensions based on the strength of their relationship with the overall rating using the chi-square statistic. The most important features for the HRS data were "value for money" and "hotel ambiance", which was also confirmed to be plausible by the domain experts of HRS[12].

An interesting and somewhat surprising aspect that can be seen in Table 11 is that using only the most important quality factor leads to results that are close to the result that we obtain when using all the dimensions. Adding a small number of additional quality factors, however helps us to further decrease the RMSE. (Re-)Adding the remaining dimension leads to a slight deterioration of the results.

These observations, therefore, confirm our assumption that not all criteria ratings are equally valuable and that the corresponding feedback dimensions should be taken with care.

Applying strategy ST2, which includes the removal of features that cause the RMSE to increase, led to a similar slight improvement of the RMSE after

[12] We limited our tests to the 14 most relevant dimensions according the chi-square statistic.

Features	HRS-RAW	HRS-5-5
1	0.61	0.52
2	0.61	0.49
3	**0.59**	0.49
4	**0.59**	0.49
5	0.61	**0.48**
6	0.60	0.49
7	0.60	0.50
8	0.61	0.50
9	0.61	0.49
10	0.60	0.49
11	0.60	0.50
12	0.61	0.50
13	0.62	0.50
14	0.62	0.51

Table 11 RMSE values for feature selection ST1 for the HRS-RAW and HRS-5-5 datasets.

the first few dimensions. Removing dimensions which did not lead to an RMSE improvement did however not help us to get significantly better results.

Finally, the evolutionary optimization procedure used in strategy ST3 led to RMSE results that were comparable to the results that were obtained when all dimensions are included as shown in Table 6.

When applying strategy ST1 on the Yahoo!Movies dataset, we could observe that leaving out detailed rating feedback in every case led to a decrease of the accuracy. In contrast to the HRS dataset, the number of rating dimensions for the movie dataset is very low and comprises only 4 dimensions and all of them can be helpful to better estimate the user's true preferences.

Discussion. Some booking platforms allow customers to evaluate the hotels along quite a number of different dimensions. The following observations can be made. First, the correlation analysis corroborates our findings from Section 2 that not all features are equally relevant for the user. Second, there seem to be a number of key quality factors like "value-for-money" which strongly determine the overall evaluation while others are not particularly relevant or even introduce noise. From a practical perspective, the selection of quality dimensions on which customers can give feedback should be done with care. The selected rating dimensions should first of all be both understandable and relevant for the decision process for the customers. At the same time, the number of rating dimensions should probably be kept small in order to obtain high-quality reliable feedback.

4 Relation to other works

Recommender systems are nowadays used in a variety of domains as a tool to support the online customer is the information filtering, decision making and buying process. In the travel and tourism domain, such systems are for example developed to help the customer in the pre-trip information search and decision making process. Examples of recent research include knowledge-based and conversational approaches to filter destinations and select travel packages [21], [22], [51], context-aware recommendation of places of interest [4], mobile recommenders [42], or the development of more intelligent user interaction strategies [33].

Recommendation in the tourism domain has some specific particularities and challenges, which are not present in more classical RS application domains, in which especially collaborative filtering (CF) techniques have been successfully applied in the past. Customers in the tourism domain, for example, do not purchase items as frequently as customers of an online book store or movie rental system. Thus, the amount of user feedback and the buying history available for building systems based on collaborative filtering techniques may be limited, which is why conversational approaches are often chosen. Furthermore, the context of the traveler or tourist is particularly relevant. Think, e.g., of making recommendations for a group of people traveling together. Also, the type of the trip (business or private) or seasonal aspects can be important when recommending tourism products.

At the same time, multi-dimensional ratings are quite common in the tourism domain but not so popular yet in other domains. While there exists quite a body of research in multi-criteria decision making and optimization, see [34], exploiting multi-criteria rating information for collaborative filtering is comparably new. In [2], an overview of recent research in this area is given, in which the following approaches are identified which can be considered to be multi-criteria based:

1. Classical information retrieval systems (content-based recommenders), which try to learn content-based preferences of a user based, for example, on the given overall ratings for the items.
2. Retrieval systems which allow users to state their *general* or specific preferences using a set of predefined categories. Typical examples are given in knowledge-based or critique-based recommendation systems [16].
3. Multi-criteria rating recommenders, in which users are allowed to specify their preferences (ratings) *for individual items* along different dimensions.

The work presented in this paper clearly falls into the third category and we will, thus, limit the discussion to related works of this category.

An extension of the Flexible Mixture Model (FMM) for collaborative filtering of [47] to incorporate multi-criteria ratings was proposed in [44] and [43]. In their work, the authors try to automatically detect existing dependency structures within the criteria ratings (which they call multi-component ratings). These dependencies are then incorporated in the probability calculations. Based on a dataset obtained from Yahoo!Movies they could empirically show that their extended FMM model can lead to a higher prediction accuracy than when only the single-rating model is used, at least for low-density data situations.

Zhang et al. later on in [53] proposed a related probabilistic approach which extends the the Probabilistic Latent Semantic Analysis (PLSA) model

[15] to the multi-criteria rating case. In their analysis they could show that their model is capable to outperform a traditional item-based nearest neighbor approach which uses only the overall rating on a Yahoo!Movies dataset.

The application of multi-linear singular value decomposition (SVD) to exploit information about the user's current context as well as multi-criteria ratings in the recommendation process was proposed by Li et al. in [30]. Their experimental analysis in a restaurant recommendation scenario showed that their approach was better in terms of precision and recall than a comparably weak baseline algorithm that only used the overall ratings as an input.

Since the datasets used in the above-mentioned evaluations are not publicly available, a direct comparison with our approach is not possible. In our view, however, the single-rating matrix-factorization model used in our experiments represents a much stronger baseline than traditional nearest-neighbor approaches used in previous works on multi-criteria recommender systems.

Another recent multi-criteria recommendation approach based on clustering of users was presented by Liu et al. in [32]. In their work, they assume that for each user there are quality factors (and rating criterions) which are more important than others and thus dominate the decision process. Correspondingly, they try to cluster users according to their criteria preferences and base their rating predictions on users in the same cluster. Similar to our work, they evaluate their approach based on a dataset from TripAdvisor and can show that the prediction accuracy can be significantly improved with their approach when compared to a traditional single-rating approach.

Their work is similar to ours in that we try to determine a user-specific weight for the different quality factors. The clustering technique is in our view complementary to our approach. One possible limitation of their approach could be that in the experiments of Liu et al. the dataset is preprocessed in a way that only users are considered which have rated at least 20 hotels, which we believe is a relatively strong assumption. On the other hand, basing personalized recommendations only on a few ratings can be a risky strategy. In our view, more work is therefore required to determine the right point for switching, e.g., from the "safe" recommendation of top-ranked hotels to personalized recommendations.

5 Limitations

The empirical study in the first part of the paper was based on data systematically harvested from a tourism platform, where all available ratings for 14 metropolitan destinations located on 4 different continents (America, Europe, Asia and Australia) have been collected. Therefore, the findings mainly apply to the assessment of city hotels and ratings of hotels situated in other traditional travel destinations such as beaches or mountains are not included.

However, the algorithmic evaluation in the second part showed that exploiting the individual weights of the users' multi-criteria ratings leads to accuracy improvements on data from two different tourism platforms and also on non-tourism related data. Still, while our experiments showed that relying on multi-criteria rating information for recommendation can be helpful not only in the tourism domain, all datasets used in our experiments are comparably small. Additional experiments with larger datasets are therefore required to analyze if and to which extent the effectiveness of the algorithms varies depending on the amount of available data. Unfortunately, no such dataset containing multi-criteria ratings is publicly available yet.

In general, the evaluation of the predictive accuracy is based on a standard experimental methodology for benchmarking recommendation algorithms. In particular in the tourism domain, the suitability of a certain offering can however be highly dependent on the situational context of the user. For a customer being on a business trip quality factors like Internet availability or other business services might be important decision criteria. When the same customer looks for a hotel for a private stay over the weekend with the family, other aspects might be more relevant. While context-aware recommendation techniques have obtained increased interest in recent years, in our view more research is required to understand how to consider these contextual factors into the recommendation processes.

Finally, prediction accuracy is only one possible evaluation criterion when comparing different recommendation techniques. Aspects like the diversity, homogeneity, familiarity, novelty or serendipity of the recommendations can be important factors that determine the success and user acceptance of a recommender system [17]. Some of these aspects can however not be evaluated based on offline experimentation with data but only by involving real users.

6 Summary and Conclusions

The paper presented an empirical analysis of multi-criteria customer feedback on TripAdvisor that provided clear evidence that different customer segments weight the relative importance of rating dimensions differently when making their overall assessments of an accommodation. A more detailed analysis based on the Kano-model led also to a qualitative assessment of the rating dimensions by classifying them into *basic*, *excitement* and *performance factors* that supports service providers to specifically design services for different segments.

Based on the empirical evidence of different weight assessments for the rating dimensions by different tourist segments and assuming a long rating history of users we developed a recommendation and prediction mechanism following the collaborative filtering paradigm that takes the different weights for rating dimensions into account and learns a prediction model that not only employs optimal weights for each segment but for each user and each item individually. In an extensive evaluation on real-world datasets from two tourism platforms (TripAdvisor and HRS) as well as from the movie domain we can show that our proposed approach consistently outperforms other state-of-the-art techniques in terms of the traditional accuracy measures used in the recommender systems research community.

In case of cold-start users a tourism platform could also incorporate segment-specific weights for ranking search results. However, in the latter case an algorithm must be able to automatically assign users to customer segments based on their self-reports, demographics or other criteria. Furthermore, knowing about the qualitative differences in the appreciation of different criteria can be used to generate segment-specific item descriptions and explanations [52] in order to not only more accurately predict items of interest but also to create more persuasive [50] interaction experiences.

References

1. Gediminas Adomavicius and YoungOk Kwon. New recommendation techniques for multicriteria rating systems. *IEEE Intelligent Systems*, 22:48–55, May 2007.
2. Gediminas Adomavicius, Nikos Manouselis, and YoungOk Kwon. Multi-criteria recommender systems. In Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, editors, *Recommender Systems Handbook*, pages 769–803. Springer US, 2011.
3. Gediminas Adomavicius and Jingjing Zhang. Impact of data characteristics on recommender systems performance. *ACM Transactions on Management Information Systems*, 3(1):3:1–3:17, April 2012.
4. Linas Baltrunas, Bernd Ludwig, Stefan Peer, and Francesco Ricci. Context-aware places of interest recommendations for mobile users. In *Proc. International Conference on Human-Computer Interaction (HCII 2011)*, pages 531–540, Orlando, FL, 2011.
5. Randall D. Brandt. How service marketers can identify value enhancing service elements. *Journal of Services Marketing*, 2(3):35–41, 1988.
6. Bruno Busacca and Giovanna Padula. Understanding the relationship between attribute performance and overall satisfaction: Theory, measurement and implications. *Marketing Intelligence & Planning*, 23(6):543–561, 2005.
7. Harris Drucker, Chris, Burges L. Kaufman, Alex Smola, and Vladimir Vapnik. Support vector regression machines. In *Advances in Neural Information Processing Systems 9*, volume 9, pages 155–161, 1997.
8. Matthias Fuchs and Klaus Weiermair. New perspectives on satisfaction research in tourism destinations. *Tourism Review*, 58(3):6–14, 2003.
9. Matthias Fuchs and Klaus Weiermair. Destination benchmarking: An indicator-system's potential for exploring guest satisfaction. *Journal of Travel Research*, 42:212–225, 2004.
10. Simon Funk. Try this at home. http://sifter.org/˜simon/journal/20061211.html (last accessed: 03/2013), 2006.
11. Fatih Gedikli and Dietmar Jannach. Improving recommendation accuracy based on item-specific tag preferences. *ACM Transactions on Intelligent Systems and Technology*, 4, 2013.
12. Ulrike Gretzel and Kyung Hyan Yoo. Use and impact of online travel reviews. In *Proc. ENTER 2008*, pages 35–46, Innsbruck, Austria, 2008.
13. Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and Lotfi A. Zadeh. *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*. Springer-Verlag New York, Inc., 2006.
14. Joseph F. Hair, Rolph E. Anderson, Barry J. Bubin, Ronald L. Tatham, and William C. Black. *Multivariate Data Analysis. 6th edn.* Prentice-Hall, New York, 2006.
15. Thomas Hofmann. Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems*, 22:89–115, 2004.
16. D. Jannach and G. Kreutler. Personalized user preference elicitation for e-services. In *Proceedings of the IEEE International Conference one-Technology, e-Commerce and e-Service, EEE '05*, pages 604–611, 2005.
17. D. Jannach, L. Lerche, F. Gedikli, and G. Bonnin. What recommenders recommend - an analysis of accuracy, popularity, and sales diversity effects. In *Proceedings of the 21st International Conference on User Modeling, Adaptation and Personalization (UMAP 2013)*, Rome, Italy, 2013.

18. Dietmar Jannach, Fatih Gedikli, Zeynep Karakaya, and Oliver Juwig. Recommending hotels based on multi-dimensional customer ratings. In *Proceedings ENTER 2012 eTourism Conference*, pages 320–331, Helsingborg, Sweden, 2012.
19. Dietmar Jannach, Zeynep Karakaya, and Fatih Gedikli. Accuracy improvements for multi-criteria recommender systems. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC 2012)*, pages 674–689, 2012.
20. Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. *Recommender Systems - An Introduction*. Cambridge University Press, 2010.
21. Dietmar Jannach, Markus Zanker, and Matthias Fuchs. Constraint-based recommendation in tourism: A multiperspective case study. *International Journal of Information Technology and Tourism*, 11(2):139–155, 2009.
22. Dietmar Jannach, Markus Zanker, Markus Jessenitschnig, and Oskar Seidler. Developing a conversational travel advisor with ADVISOR SUITE. In *Proceedings ENTER 2007 eTourism Conference*, pages 43–52, Ljubljana, Slovenia, 2007.
23. Robert Johnston. The determinants of service quality: Satisfiers and dis-satisfiers. *International Journal of Service Industry Management*, 6(1):53–71, 1995.
24. Noriaki Kano. Attractive quality and must-be quality. *Hinshitsu: The Journal of the Japanese Society for Quality Control*, 14(2):39–48, 1984.
25. Peter Klaus. Quality epiphenomenon: The conceptual understanding of quality in face-to-face service encounters. In *The Service Encounter: Managing Employee Customer Interaction in Service Business*, pages 17–33. Lexington, 1985.
26. R. B. Kline. *Principles and practice of structural equation modeling*. Guilford Press, London, 2005.
27. Yehuda Koren. Factor in the neighbors: Scalable and accurate collaborative filtering. *ACM Transactions on Knowledge Discovery from Data*, 4:1–24, 2010.
28. MH Kutner, CJ Nachtsheim, and J Neter. *Applied Linear Regression Models, 4th edition*. McGraw-Hill Irwin, 2004.
29. Daniel Lemire and Anna Maclachlan. Slope one predictors for online rating-based collaborative filtering. In *Proceedings of the 5th SIAM International Conference on Data Mining (SDM'05)*, pages 471–480, Newport Beach, CA, 2005.
30. Qiudan Li, Chunheng Wang, and Guanggang Geng. Improving personalized services in mobile commerce by a novel multicriteria rating approach. In *Proceedings of the 17th International Conference on World Wide Web (WWW 2008)*, pages 1235–1236, Beijing, China, 2008.
31. Huan Liu and Rudy Setiono. Chi2: Feature selection and discretization of numeric attributes. In *Proceedings of the Seventh International Conference on Tools with Artificial Intelligence (ICTAI 1995)*, pages 388–391, Vancouver, Canada, 1995.
32. Liwei Liu, Nikolay Mehandjiev, and Dong-Ling Xu. Multi-criteria service recommendation based on user criteria preferences. In *Proceedings of the fifth ACM Conference on Recommender Systems (RecSys 2011)*, pages 77–84, Chicago, IL, USA, 2011.
33. Tariq Mahmood, Francesco Ricci, and Adriano Venturini. Improving recommendation effectiveness: Adapting a dialogue strategy in online travel planning. *International Journal of Information Technology and Tourism*, 11(4):285–302, 2009.
34. Nikos Manouselis and Constantina Costopoulou. Analysis and classification of multi-criteria recommender systems. *World Wide Web*, 10:415–441, December 2007.
35. Kurt Matzler, Franz Bailom, Hans Hinterhuber, Birgit Renzl, and Johann Pichler. The asymmetric relationship between attribute-level performance and overall customer satisfaction: A reconsideration of the importance-performance analysis. *Industrial Marketing Management*, 33:271–277, 2004.
36. Kurt Matzler and Elmar Sauerwein. The factor structure of customer satisfaction: An empirical test of the importance grid and the penalty-reward-contrast analysis. *International Journal of Service Industry Management*, 13(4):314–332, 2002.
37. Ingo Mierswa. *Non-convex and multi-objective optimization in data mining*. PhD thesis, Department of Computer Science, TU Dortmund, Germany, 2009.
38. Josip Mikulic and Darko Prebeac. Prioritizing improvement of service attributes using impact range-performance analysis and impact-asymmetry analysis. *Managing Service Quality*, 18(6):559–576, 2008.

39. K.-R. Müller, A.J. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik. Predicting time series with support vector machines. In *Artificial Neural Networks - ICANN'97*, volume 1327 of *Lecture Notes in Computer Science*, pages 999–1004. Springer Berlin Heidelberg, 1997.
40. Miki Nakagawa and Bamshad Mobasher. A hybrid web personalization model based on site connectivity. In *Proceedings of the Workshop on Web Mining and Web Usage Analysis (WebKDD'03)*, pages 59–70, Washington, DC, USA, 2003.
41. Yvette Reisinger and Lindsay Turner. Structural equation modeling with lisrel: application in tourism. *Tourism Management*, 20(1):71 – 88, 1999.
42. Francesco Ricci. Mobile recommender systems. *International Journal of Information Technology and Tourism*, 12(3):205–231, 2011.
43. Nachiketa Sahoo, Ramayya Krishnan, George Duncan, and James P. Callan. Collaborative filtering with multi-component rating for recommender systems. In *Proceedings of the Sixteenth Annual Workshop on Information Technologies and Systems (WITS'06)*, Milwaukee, USA, 2006.
44. Nachiketa Sahoo, Ramayya Krishnan, George Duncan, and James P. Callan. The Halo Effect in multi-component ratings and its implications for recommender systems: The case of Yahoo! Movies. *Information Systems Research*, 23(1):231–246, 2012.
45. Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, Cambridge, MA, USA, 2001.
46. Shilad Sen, Jesse Vig, and John Riedl. Tagommenders: Connecting users to items through tags. In *Proceedings of the 18th International World Wide Web Conference (WWW'09)*, pages 671–680, Madrid, Spain, 2009.
47. Luo Si and Rong Jin. Flexible mixture model for collaborative filtering. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML 2003)*, pages 704–711, Washington, DC, 2003.
48. Jan-Benedict E.M Steenkamp and Hans Baumgartner. On the use of structural equation models for marketing modeling. *International Journal of Research in Marketing*, 17(2/3):195 – 202, 2000.
49. Klaus Weiermair and Matthias Fuchs. Measuring tourist judgments on service quality. *Annals of Tourism Research*, 26(4):1004–1021, 1999.
50. Kyung-Hyan Yoo, Ulrike Gretzel, and Markus Zanker. *Persuasive Recommender Systems - Conceptual Background and Implications.* Springer, 2013.
51. Markus Zanker, Matthias Fuchs, Wolfram Höpken, Mario Tuta, and Nina Müller. Evaluating recommender systems in tourism - a case study from Austria. In *Proceedings ENTER 2008 eTourism Conference*, pages 24–34, Amsterdam, Netherlands, 2008.
52. Markus Zanker and Daniel Ninaus. Knowledgable explanations for recommender systems. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI/IAT)*, pages 657–660. IEEE, 2010.
53. Yin Zhang, Yueting Zhuang, Jiangqin Wu, and Liang Zhang. Applying probabilistic latent semantic analysis to multi-criteria recommender system. *AI Communications*, 22(2):97–107, 2009.

Appendix - Parameter optimization for weighted prediction model

The goal of the weight optimization process described in Section 3.3 is to find weight parameters w_u* and w_i* that minimize the prediction error on the training data and at the same time do not overfit the model to the data. The optimization goal is given in Equation 3, where K corresponds to the user-item rating tuples in the training set and lambda is the penalty factor.

$$\min_{w_i*,w_u*} \underset{(u,i)\in K}{\Sigma} (r_{u,i} - (w_u * \hat{r}_{u,i}^{user} + w_i * \hat{r}_{u,i}^{item}))^2 + \lambda(\underset{u}{\Sigma} w_u^2 + \underset{i}{\Sigma} w_i^2) \qquad (3)$$

Algorithm 1 shows our procedure to iteratively optimize the weights similar to the gradient descent approach from [27] and other recent works. The algorithm starts with randomly chosen initial weights and iterates over all ratings in the training set. It generates predictions with the current weights and compares them with the true ratings. Based on the observed error, the weights are then slightly adjusted. This procedure is repeated for a pre-defined number of iterations (e.g., 50). The parameters γ and λ determine the step size for the weight adaptation and a penalty factor for overfitting [19].

Algorithm 1 Optimization of weights.

```
Require: $\#iterations$, $\gamma$, $\lambda$
  // Do defined number of iterations
  for 1 to $\#iterations$ do
    for each user $u$ do
      for each rated item $i$ of user $u$ do
        // compute prediction with current weights
        $\hat{r}_{u,i} \leftarrow w_u \cdot \hat{r}^{user}_{u,i} + w_i \cdot \hat{r}^{item}_{u,i}$
        // compare with real rating $r_{u,i}$ and determine the error $e_{u,i}$
        $e_{u,i} \leftarrow r_{u,i} - \hat{r}_{u,i}$
        // Adjust $w_u$
        $w_u \leftarrow w_u + \gamma \cdot (e_{u,i} - \lambda \cdot w_u)$
        // Adjust $w_i$
        $w_i \leftarrow w_i + \gamma \cdot (e_{u,i} - \lambda \cdot w_i)$
      end for
    end for
  end for
  return $w_u$ for each user $u$ and $w_i$ for each item $i$
```