Recommending Judgment Targets for Rating Provision

Friederike Klan
Institute of Computer Science
Friedrich-Schiller-University of Jena
Email: friederike.klan@uni-jena.de

Birgitta König-Ries
Institute of Computer Science
Friedrich-Schiller-University of Jena
Email: birgitta.koenig-ries@uni-jena.de

***Abstract*—Existing rating and reviewing schemes typically come in the in the flavor of a single rating and/or a textual review. While a single judgment evaluating the overall quality of a product is of limited significance, textual customer reviews typically deliver more informative feedback at the attribute level. However, reading and comparing them to extract relevant information is time-consuming and mentally-demanding. Moreover, they are often biased and selective in the product features they consider and thus are less helpful for making informed buying decisions. We therefore propose a rating elicitation scheme that supports consumers in providing meaningful and machine-comprehensible, i.e automatically processable, responses in terms of multi-criteria ratings judging several features of a purchased product or used service. This is achieved by recommending suitable judgment targets and thereby accounting for a customer's willingness to provide ratings. Our evaluation results show that the proposed procedure effectively adjusts to a consumer's personal judgment preferences and thus provides helpful support for the elicitation of meaningful multi-criteria feedback.**

I. Introduction

Collaborative feedback schemes have been widely and successfully used to establish trust in online markets. Today, nearly every online store offers reviewing or rating facilities, since customer reviews have proven to be highly influential on buying decisions. According to Nielsen, 57% of the customers read online reviews before making a purchase [1]. Nearly three-quarters of the consumers consider online reviews as much trustful as personal recommendations [2]. However, the confidence that is put in electronic Word-of-Mouth is only partially justified and the usability of common rating and reviewing schemes that come in the flavor of a single 5-star rating and/or textual review capabilities is limited. This is due to several reasons.

Requirements and preferences regarding product or service properties and their importance differ among consumers. For instance, while one user might be more concerned about the price of a product, another one might be more interested in its technical properties. When it comes to make a purchase decision only feedback of those people sharing a similar taste, i.e. judging the same products in a similar way, is relevant. A single judgment, such as the typical 5-star rating evaluating the overall quality of a product, is of limited significance, since it does not reveal why a certain aggregated judgment was given and hence does not allow for the selective and personalized use of feedback information.

In opposition to that, textual customer reviews typically deliver attribute-specific feedback. They provide the information required to assess feedback relevance and thus are more valuable for customers. According to [2], 65% of the consumers study between 2 and 10 reviews before buying. However, reading and comparing reviews to extract relevant information is time-consuming and mentally-demanding. It is therefore desirable, that the meaning of judged product or service features is made explicit, consistent among the judgment providers and machine-comprehensible. This property would ensure the comparability of judgments provided by different users and allow for their automated processing to select and present consumer feedback in a personalized and easily comprehensible way.

Finally, customer judgments are often driven by emotions and personal interests, such as self-expression, group commitments or monetary rewards [3]. As such, they are biased and selective in the product features they consider and thus are less helpful for making informed buying decisions. For instance a review praising the high quality of a purchased hand-bag without saying anything about the payed price would not help in finding a hand-bag that is a good compromise between price and quality. Useful feedback should therefore rather be comprehensive, meaning that all aspects that were relevant to a purchase should be considered in the customer judgment. Moreover, the judged aspects should be appropriate, i.e. meaningful in the context of the judged product or service. For instance, it makes sense to judge an aspect "taste" when referring to a wine purchase, but not when assessing the performance of a ticket booking service. Existing rating and reviewing approaches do not support customers in providing meaningful, i.e. comprehensive and appropriate, feedback.

In this paper, we propose a rating elicitation scheme that overcomes the identified shortcomings of existing rating and reviewing schemes.

- It supports consumers in providing meaningful and machine-comprehensible responses in terms of multi-criteria ratings judging several features of a purchased product or used service.
- This is achieved by recommending judgment targets that are both, appropriate and comprehensive, in the context of a judged product or service.
- The approach also considers a customer's willingness to provide ratings by taking his past judgment behavior into account.

The latter is particularly important, since asking a consumer for a number of judgments he is not willing to provide is likely to result in no or bad quality feedback [4], [5]. The suggested mechanism is designed for judgment scenarios where products or services are purchased/used and judged on a single site, such as on Amazon.com or other E-commerce platforms. It relies on the availability of a semantic model of a customers requirements and preferences with respect to the judged product or service. Such a model can for instance be learned by analyzing the consumer's interactions with the platform prior to the judged purchase. Our evaluation results show that the proposed procedure effectively adjusts to a consumer's personal judgment preferences and thus provides helpful support for the elicitation of meaningful multi-criteria feedback. A detailed discussion on how to effectively use multi-criteria consumer feedback is out of the scope of this paper. Refer to [6], [7] for work on this topic.

II. Related Approaches

The need to derive structured feedback information from textual reviews or other web sources, such as social media, to facilitate informed customer decisions has long been recognized in the research community. In particular, text mining techniques have been used to discover review dimensions, such as price or quality, and performing a subsequent sentiment analysis to determine a reviewers attitude towards those aspects (see e.g. [8], [9]). Similar algorithms have been studied to explain why a certain overall judgment was given [10] and thus allowing for effective product recommendations based on customer ratings. Though structured knowledge extracted from user-provided feedback can be helpful, the quality and thus the usefulness of mined information is still limited to some degree.

Another line of research is therefore concerned with directly eliciting structured feedback information. Recently, collaborative mechanisms that leverage multi-criteria ratings, judging separate aspects of a product or service interaction gained attention in the research community (see [6] for an overview). First results indicate that systems that are based on such detailed consumer feedback can produce significantly more accurate product or service recommendations than those relying on traditional single-criterion judgments [6], [7]. However, the focus of this work is on effective usage of judgments at the attribute level, while we are interested in the effective elicitation of this kind of feedback.

The elicitation of multi-criteria feedback is also an issue in an increasing number of commercial recommender systems, such as on Ebay, HRS, Epinions or Yahoo! Music. Typically, the set of aspects that might be judged by a consumer is either the same for all product types or specific per product category. However, in the first case, this set of aspects is either very generic and thus less informative for potential customers, or not appropriate for all products. In the second case, this set has to be specified manually for each new product. Judgments referring to products of different categories are also not comparable. For instance, Epinions allows its users to judge movies with respect to the aspects *Action Factor*, *Special Effects* and *Suspense*. Those might make sense for *Action Movies*, but do not fit when considering other movie categories offered by Epinions like *Children Movies*, *Education* or *Comedy*. Epinions also allows to review online stores. They can be rated in terms of the aspects *Ease of Ordering*, *Customer Service*, *On-Time Delivery* and *Selection*. Those aspects are very general and do not allow to judge shop-specific characteristics. Moreover, single aspect ratings are typically supplementary in the sense that they do not have any influence on a product's overall rating.

Alternatively, some reviewing engines such as those provided by Bazaarvoice or Powerreviews, offer more flexible semi-structured reviewing facilities based on tagging. Those systems allow consumers to create tags describing the pros and contras of a given product. These tags can then be reused by other users. Tagging provides a very intuitive and flexible mechanism that supports customers in providing appropriate and helpful judgments. However, the high flexibility of the approach is at the cost of the judgments' meaningfulness. This is due to the fact that tags do not have a clear semantics. In particular, the relationship between different tags is unknown and thus makes them incomparable. Moreover, those systems do not ensure that all relevant aspects of a product or a service interaction are judged.

To summarize our findings, more flexible and adaptive mechanisms to effectively elicit and describe multi-criteria feedback are required. In particular, the question of how to describe this type of feedback meaningfully and machine-comprehensible has been hardly considered. To the best of our knowledge, the issue of assisting consumers in providing meaningful, i.e. comprehensive and appropriate, feedback has not been addressed at all. Although the question of what motivates consumers to write reviews is an ongoing topic in social psychology and research on electronic word of mouth [3], [11], aspects such as a consumer's willingness to judge certain attributes of a product have been hardly considered in existing solutions.

III. Semantic Requirements Model

As mentioned, our approach relies on the availability of a semantic model of a consumer's requirements and preferences with respect to the judged product or service. Such a model can be learned based on the customer's interactions for instance with an E-Commerce platform or a (Conversational) Recommender System prior to the judged purchase/selection. Since a detailed discussion of this topic would go beyond the scope of this paper, we refer the interested reader to [12], [13].

Although, our solution is not restricted to a specific modeling technique, applicable candidates have to fulfill some basic requirements. In particular, suitable requirements models should hierarchically and more and more fine-grained characterize acceptable products or services by means of their attributes, such as price or color, where higher-level attributes represent categories of product or service properties. Each attribute may be constrained by conditions on its values. As an example, consider the requirements model depicted in Fig. 1. Acceptable products are mobile phones that are cheaper than 50$, are either silver or black, are of bar or slider style and are from either Nokia or Sony Ericsson. While the (category) attribute price describes acceptable prices and currencies, the (category) attribute productType characterizes desirable product types.

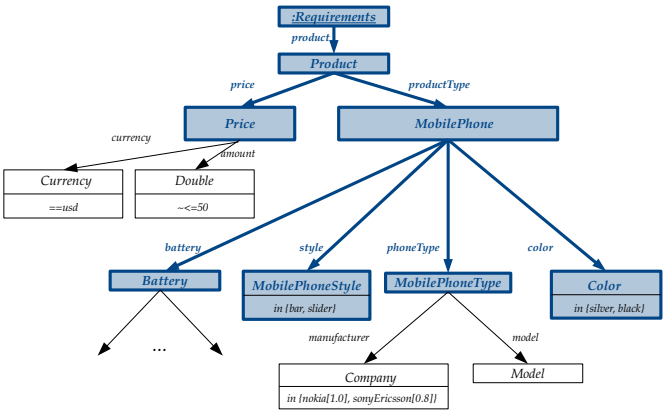

Fig. 1. Semantic requirements model and a possible feedback structure derived from it (blue part)

Attributes referenced in the requirements model should also have a well-defined meaning and should refer to concepts in an ontology, which defines not only valid product/service properties, but also valid relationships among them. Applicable modeling techniques may, but do not necessarily have to, provide means to indicate the relative importance of attributes and to specify preferences over acceptable attribute values.

For the implementation of our approach to feedback elicitation, we have chosen the semantic modeling technique proposed in [14]. It fulfills our requirements and is both, expressive and easy to use. It is not just capable of modeling consumer requirements, but also suitable for modeling attribute-related preferences.

IV. Meaningful Consumer Feedback

In this section, we will demonstrate how semantic requirements models can be leveraged to describe and acquire attribute-level customer ratings that are machine-comprehensible, comprehensive and appropriate for characterizing a product or service. Potentially, all aspects that are considered in a requirements model are relevant to the consumer and thus are appropriate judgment targets when rating a product or service that has been selected based on those requirements. More specifically, a judgment referring to a certain aspect of the model indicates the performance of the purchased product with respect to this aspect as experienced by the judgment provider. Judgments referring to intermediate, i.e. non-leaf, aspects of a requirements model indicate the product's aggregated performance with respect to those attributes' child aspects (and thus with respect to all aspects covered by the model subtree rooted at the judged aspect). Consider for example the requirements model depicted in Figure 1. By providing a rating for the attribute *productType*, a judgment provider indicates the aggregated performance of the purchased product with respect to the aspects *battery*, *style*, *phoneType* and *color*.

Letting judgments reference aspects that are covered in the requirements model is advantageous. By coupling (parts of) the consumer's semantically described requirements with the provided judgments, we provide judgments with a well-defined and commonly agreed upon meaning. This is due to the fact that attributes covered in the requirements model refer to concepts in an ontology, which define not only valid product or service attributes and valid constraints on them, but also valid relationships among them. This ensures that there is an agreed upon machine-comprehensible, meaning of the judgments provided by different consumers and thus allows for the comparability of judgments provided by different users and for different items. This in turn, is essential for their automated processing to select and present consumer feedback in a personalized and easily comprehensible way.

However, we also have to ensure that provided judgments are comprehensive. The challenging question is how the latter can be achieved while at the same time still being flexible in the choice of the aspects to rate. We propose the concept of a *feedback structure* to deal with this issue.

Definition 1: (Feedback Structure) A feedback structure is a subtree of the hierarchical requirements model, whose leaves correspond to the aspects that have to be rated by the user. In contrast to the requirements model from which it is derived, a feedback structure does not contain any preference information.

Consider the requirements model depicted in Fig. 1. The blue part of the tree indicates a possible feedback structure for that model, where the aspects *price*, *battery*, *style*, *phoneType* and *color* have to be rated by the consumer. Restricting the judgment space to feedback structure leaves guarantees, that no aspect is judged twice. This would for instance happen, if a consumer judged both, *productType* and *phoneType*, since the *productType*-judgment contains an indirect judgment of *phoneType*. We prohibit multiple judgments for single attributes, since they are likely to cause inconsistencies due to inconsistent judgments. To assure that the provided feedback is comprehensive, the model subtrees rooted at the feedback structure's leaves should cover all leaves of the model tree. This guarantees that all aspects considered in the requirements model, i.e. all aspects that are relevant to the consumer, are either directly or indirectly (by providing an aggregated rating) judged. The feedback structure depicted in Fig. 1 fulfills this requirement and thus is valid. Omitting, e.g., the aspect *phoneType* would result in an invalid structure, since the aspects *phoneType*, *manufacturer* and *model* would not be judged.

Note, that we are still flexible in the choice of the feedback structure and hence in the choice of the attributes that have to be judged by the consumer. For example, a user might provide a single rating for *productType* instead of judging *battery*, *style*, *phoneType* and *color* separately. Providing just a single overall rating to judge an item as a whole is also a valid option when using this scheme. In this case, the feedback structure comprises of a single node corresponding to the requirements model's root node. Obviously, there is a trade-off between the detailedness and thus the number of the provided judgments and the rating effort. The more detailed the provided judgments are, the more informative and thus the more valuable they are for other customers seeking to make a purchasing decision. However, providing detailed judgments imposes a higher judgment effort on the user. A strength of the proposed solution is that, by letting judgment providers freely choose the feedback structure to judge, it enables them

to make this compromise according to their personal judgment preferences.

V. Recommending Judgment Targets

In the previous section, we discussed how consumer feedback can be described in a way that ensures its comprehensiveness, appropriateness and machine-comprehensibility. However, so far we owe to explicate how humans can be assisted in actually providing the desired judgment information. In particular, we did not explain how feedback quality can be ensured by accounting for a consumer's judgment preferences, i.e. his willingness to provide certain ratings, when eliciting judgment information and how such a process can flexibly and automatically adjust to different judgment preferences. In this section, we will suggest an elicitation mechanism that satisfies those requirements by suggesting suitable judgment targets based on a user's past judgment behavior.

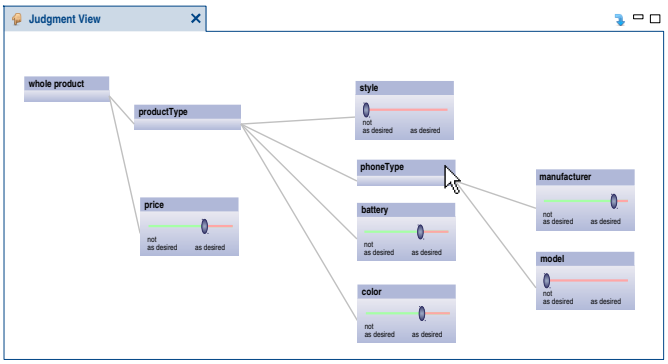

Fig. 2. Judgment view

Assume, that given certain requirements, an appropriate product or service was selected and purchased and now its performance shall be judged by the consumer. In a first step, we utilize the available requirements model to determine valid feedback structures as defined in the previous section. Subsequently, the structure that is most likely to be judged by the user, i.e. in the context of the given customer requirements, fits best to the consumer's personal judgment preferences, is selected and graphically displayed to the user (Fig. 2). The required knowledge about the user's judgment preferences is learned from his behavior in previous judgment sessions. The presented feedback structure represents a careful compromise between the consumer's competing judgment preferences and as such typically cannot perfectly meet all his judgment requirements. It is just a recommendation to the user and thus, if required, might be adjusted by him to match his actual judgment needs. This can be done by expanding and/or collapsing subtrees rooted at the presented structure's leaves. Subtrees are expanded level-wise by simply clicking on the considered attribute node and collapsed completely upon clicking on an expanded node. Thereby, maximally expanding the displayed feedback structure will result in the attribute structure given by the requirements model. As an example consider the feedback structure depicted in Fig. 2. The user might expand the leaf *phoneType* to judge its subaspects *manufacturer* and *model*. He might also collapse the root attribute, to provide just a single overall judgment for the purchased item. After having customized the recommended feedback structure according to his needs, the user finally judges all leaf attributes of the structure by providing a rating. This can for instance be done by simply moving a slider that is shown for those attributes as shown in Fig. 2. However, any other rating scheme, such as a 5-star-rating can be used as well. Once, the consumer submits his judgments, the system takes care of storing all relevant feedback information and session data for future recommendations. In particular, it is recorded which and how many attributes were judged by the consumer and which requirements model led to the judgment. The acquired information are used later on to identify the feedback structure that is most likely to be judged by the user in future judgment sessions.

As already mentioned, there exist typically numerous valid feedback structures for a given requirements model. However, usually they do not fit equally well to a user's judgment preferences. Subsequently, we will therefore introduce a mechanism that assesses the likelihood of a certain structure to be judged by the user and thus enables the identification of the/a feedback structure among the possible structures that best fits to the user's judgment preferences.

A. Assumptions

Users might have various reasons for preferring to judge certain item aspects to others, e.g. one might be willing to judge aspects that are important and might not be willing to judge aspects that are unimportant or private. Though those issues might be interesting from a psychological perspective, we do not and cannot consider them here, since there is simply little known on this topic [3]. In fact, our approach to feedback structure recommendation is purely behavioral, i.e. it considers the effect of a consumer's judgment preferences, namely the resulting judgment behavior, but does not try to explain it. In particular, we assess the likelihood of a certain feedback structure to be judged by a user by considering his judgment behavior in past judgment sessions. The likelihood is estimated based on two behavioral indicators, namely the number and kind of aspects that have been judged in the past. So far, we do not consider other factors that might influence the likelihood of a feedback structure to be judged. As already argued, a consumer's judgment preferences might vary when having different product or service requirements, e.g. a user's judgment preferences after having booked a flight might differ from those after having purchased a book. However, the basic assumption underlying our approach is that users will select a similar number and kind of judgment targets when having similar requirements towards the judged items. More specifically, we make the following two assumptions.

Assumption 1: (Number of judged aspects) Let r' be the requirements model that underlies a judgment of m' aspects and let r be a requirements model that is semantically similar to r'. Then we assume that the user is willing to judge a similar number m of aspects referring to an item selected based on the requirements r.

Assumption 2: (Type of judged aspects) Let r' be the requirements model that underlies a judgment of the set A' of aspects and let r be a requirements model that is semantically similar to r'. Then we assume that the user is willing to judge a set A of aspects, that refers to an item selected based on

the requirements r and is similar with respect to the kind of aspects contained.

We think that those assumptions are reasonable and valid in the context of our scenario. This particularly holds for the latter assumption, since similar requirements induce similar item attributes of interest and thus a similar set of aspects that might be potentially judged. As we will see, our evaluation results will support this hypothesis. In the remainder of this section, we will provide details on how the likelihood of a given feedback structure to be judged by the user is determined. We start with a discussion on how the probability of a user being willing to judge as many aspects as required by a given feedback structure can be assessed, followed by remarks on how the likelihood a user being willing to judge the kinds of aspects as required by a given feedback structure can be estimated. Finally, we will explicate how those probabilities are leveraged to decide upon a feedback structure for recommendation.

B. Probability to Judge a Certain Number of Aspects

Consider a user having had the requirements given by the model r and having selected and purchased an appropriate item or service based on r. Let FS_r be the set of valid feedback structures that can be derived from r and that might be used to judge the performance of the selected item with respect to r. Let further $E = \{(r',fs')|\ fs' \text{ was judged based on } r'\}$ be the set of information on the user's past judgment behavior, each characterized by the feedback structure fs' that was judged and the requirements model r' that led to the judgment. We refer to the probability distribution, indicating for each feedback structure $fs \in FS_r$ the likelihood of the user being willing to judge as many aspects as is required by fs as $p_{num}(r,fs)$. It is determined by Bayesian inference using the evidence provided in E, i.e. the information about the user's behavior in past judgment sessions. We start with a uniform prior distribution, i.e. when having no past judgment information, we simply assume that all feedback structures $fs \in FS_r$ are equally likely to be judged. The posterior probability distribution $p_{num}(r,fs|r',fs')$, taking the past judgment information $(r',fs') \in E$ into account, is given by

$$p_{num}(r,fs|r',fs') = c_{num} \cdot P_{num}(r',fs'|r,fs) \cdot p_{num}(r,fs),$$

where $p_{num}(r,fs)$ is the prior distribution before taking the observation (r',fs') into account and c_{num} is a normalizing constant chosen in a way ensuring that $p_{num}(r,fs|r',fs')$ is in fact a probability distribution. The probability $P_{num}(r',fs'|r,fs)$ indicates the likelihood of the user being willing to judge as many aspects as required by the feedback structure fs' derived from r', if fs derived from r would have already been judged. It can be estimated using Assumption 1.

Let fs be a feedback structure that was judged by a user in a past judgment session based on the requirements r. Let further r' being the user's requirements underlying the current judgment session and fs' being a valid feedback structure that might be judged by the user. Let $sim_{req}(r',r) \in [0,1]$ denote the semantic similarity of the requirements models r' and r, indicating how similar the item requirements encoded in the model r' are to those in the model r. A similarity value of 1 indicates that the two models are semantically identical, while values lower than 1 indicate decreasing semantic similarity. A detailed discussion of how to compute such a similarity is out of the scope of this paper and can be found for instance in [15]. Let further $sim_{num}(fs',fs) \in [0,1]$ be the similarity of the feedback structures fs' and fs with respect to the numbers m' and m of aspects they require to be judged. A similarity value of 1 indicates that the values m' and m are identical, while values lower than 1 indicate increasing distance between the two values. Then, according to Assumption 1, the following holds. If r is semantically identical to r', i.e. $sim_{req}(r',r) = 1$, and the number m' of aspects that has to be judged according to fs' is equal to the number of aspects m that has been judged based on fs, i.e. if $sim_{num}(fs',fs) = 1$, then the likelihood $P_{num}(r',fs'|r,fs)$ of the user being willing to judge as many attributes as required by the feedback structure fs' is 1. If the numbers m and m' maximally differ, i.e. $sim_{num}(fs',fs) = 0$, then $P_{num}(r',fs'|r,fs) = 0$. This does not hold, if the requirements r and r' totally differ, i.e. $sim_{req}(r',r) = 0$. In this case, we cannot draw any conclusions about the user's willingness to judge as many aspects as required by fs' from the fact that he judged fs. Hence, $P_{num}(r',fs'|r,fs)$ is set to 0.5 to indicate that both, judging fs' and not judging fs' is equally likely. The more similar the requirements r and r' are, the more similar the user's feedback structure selection behavior will be to that exhibited in the past judgment session, i.e. when judging fs. The following estimation of the probability $P_{num}(r',fs'|r,fs)$ is in compliance with those considerations.

$$P_{num}(r',fs'|r,fs) = 0.5 + ((sim_{num}(fs',fs) - 0.5) \cdot sim_{req}(r',r))$$

A natural measure for the similarity $sim_{num}(fs',fs)$ of two feedback structures with respect to the numbers m' and m of aspects they require a user to judge is based on the distance $|m' - m|$ between m' and m. The similarity should be 1 for $|m'-m| = 0$ and should decrease to 0 with increasing distance, i.e. $\lim_{|m'-m|\to\infty} sim_{num}(fs',fs) = 0$. In our implementation, we use the following similarity measure that fulfills those requirements:

$$sim_{num}(fs',fs) = \frac{1}{a\sqrt{|m'-m|}}.$$

The real number a can be freely chosen and determines how fast the similarity value decreases with increasing distance.

Note, that we do not necessarily have to consider all information in E to determine the probability distribution $p_{num}(r,fs|r',fs')$. Instead, we might restrict ourselves to the past judgment experience(s) that is (are) most similar to the current judgment situation in terms of the requirements the user had, i.e. to $\{(r',fs') \in E|\neg\exists(r'',fs'') \in E \text{ with } sim_{req}(r,r') < sim_{req}(r,r'')\}$. In our evaluation, we considered the latter variant, since it delivered more accurate recommendations than the former.

C. Probability to Judge Certain Kinds of Aspects

Consider again a consumer having had the requirements given by the model r and having selected and purchased an appropriate item or service based on r. Let FS_r be the set of valid feedback structures that can be derived from r and let E be the set of information on the user's past judgment behavior. We refer to the probability distribution, that indicates for each feedback structure $fs \in FS_r$ the likelihood of the user being willing to judge the kinds of aspects as required by the feedback structure fs, as $p_{attr}(r,fs)$. It is also

determined by Bayesian inference using the evidence provided in E and initialized as a uniform distribution assigning equal probabilities to all valid feedback structures $fs \in FS_r$. The posterior probability distribution $p_{attr}(r,fs|r',fs')$, taking the past judgment information $(r',fs') \in E$ into account, is given by

$$p_{attr}(r,fs|r',fs') = c_{attr} \cdot P_{attr}(r',fs'|r,fs) \cdot p_{attr}(r,fs),$$

where $p_{attr}(r,fs)$ is the prior distribution before taking the observation (r',fs') into account and c_{attr} is a normalizing constant chosen in a way ensuring that $p_{attr}(r,fs|r',fs')$ is in fact a probability distribution. The probability $P_{attr}(r',fs'|r,fs)$ indicates the likelihood of the user being willing to judge the kinds of aspects as required by the feedback structure fs' derived from r', if fs derived from r would have already been judged. It can be estimated similarly to $P_{num}(r',fs'|r,fs)$ by making use of Assumption 2.

$$P_{attr}(r',fs'|r,fs) = 0.5 + ((sim_{attr}(fs',fs) - 0.5) \cdot sim_{req}(r',r)).$$

The value $sim_{attr}(fs',fs) \in [0,1]$ indicates the similarity of the set of attributes $A_{fs'}^{judged}$ that has to be judged according to the feedback structure fs' and the set of attributes A_{fs}^{judged} that has to be judged according to the feedback structure fs. As a measure for $sim_{attr}(fs',fs)$, we use Jaccard's similarity coefficient [16] that is often used for comparing sample sets with respect to their elements. In particular, we define

$$sim_{attr}(fs',fs) = \frac{|A_{fs'}^{judged} \cap A_{fs}^{judged}|}{|A_{fs'}^{judged} \cup A_{fs}^{judged}|}.$$

The similarity value is 0, if the two attribute sets $A_{fs'}^{judged}$ and A_{fs}^{judged} do not share any attributes, and increases with increasing number of shared attributes up to 1 for sets that contain the same attributes.

Again, we do not necessarily have to consider all information in E to determine the probability distribution $p_{attr}(r,fs|r',fs')$, but might restrict ourselves to the past judgment experience(s) that is (are) most similar to the current judgment situation in terms of the requirements the user had. In our evaluation, we considered the latter variant.

D. Deciding which Structure to Recommend

Let r be the requirements model encoding the user's requirements and preferences towards the item to be judged. We define the measure $suit(r,fs;\alpha)$, indicating for each feedback structure $fs \in FS_r$ how well it fits to the user's judgment preferences, to be

$$suit(r,fs;\alpha) = \alpha \cdot p_{num}(r,fs) + (1-\alpha) \cdot p_{attr}(r,fs). \quad (1)$$

The parameter α with $\alpha \in [0,1]$ determines the influence of the probabilities $p_{num}(r,fs)$ and $p_{attr}(r,fs)$, respectively. An appropriate value for α might vary from one user to another. In the subsequent section, we will demonstrate how it can be learned from a consumer's past judgment behavior.

Valid feedback structures $fs \in FS_r$ of r can now be compared with respect to $suit(r,fs;\alpha)$, i.e. suitability for a given user. The most suitable feedback structure, i.e. the one with the highest value of $suit(r,fs;\alpha)$ is selected and presented to the user.

E. Determining α

As discussed earlier, the parameter α that weights the influence of the probabilities $p_{num}(r,fs)$ and $p_{attr}(r,fs)$ on the suitability $suit(r,fs;\alpha)$ of a feedback structure for a judging person, might vary from user to user. In this section, we will demonstrate how this value can be learned from a consumer's past judgment behavior.

Initially, i.e. without having information about a user's previous judgment behavior, we do not know anything about the parameter's value, so α could be any value from the interval $[0,1]$. Hence, for the purpose of computing the value $suit(r,fs;\alpha)$ which indicates how well the valid feedback structure fs of r fits to the user's judgment preferences, we equally weight the probabilities $p_{num}(r,fs)$ and $p_{attr}(r,fs)$, i.e. we set $\alpha = 1-\alpha = 0.5$. Once having determined the feedback structure fs_{rec} that is most suitable, we present it to the consumer, who has the opportunity to change it by expanding and/or collapsing attributes. Finally, the consumer provides judgments for the resulting feedback structure's leaf attributes. Let the feedback structure that was finally judged by the user be fs_{judged}. Inspired by the idea of maximum-likelihood estimation, we determine how α should have been chosen to make the judged feedback structure fs_{judged} most suitable according to $suit(r,fs;\alpha)$. This can be easily done using knowledge about the set of valid feedback structures FS_r that can be derived from r and knowledge about the feedback structure fs_{judged} that was actually judged. More specifically, we know that for each unjudged feedback structure $fs \in FS_r \setminus \{fs_{judged}\}$, the inequality $suit(r,fs;\alpha) \leq suit(r,fs_{judged};\alpha)$ must hold. Using Formula 1, we find that

$$\alpha \leq \frac{suit(r,fs_{judged};\alpha) - p_{attr}(r,fs)}{p_{num}(r,fs) - p_{attr}(r,fs)}$$

for $p_{num}(r,fs) > p_{attr}(r,fs)$ and

$$\alpha > \frac{suit(r,fs_{judged};\alpha) - p_{attr}(r,fs)}{p_{num}(r,fs) - p_{attr}(r,fs)}$$

for $p_{num}(r,fs) < p_{attr}(r,fs)$. Using those information, we can adjust, i.e. shrink the range of α correspondingly. For example, if $\alpha \leq 0.8$ holds, we adjust the interval to $[0,0.8]$. Redundant information, such as if $\alpha \leq 0.8$ holds, when already having $\alpha \in (0.5,0.7]$, are simply ignored.

Information about the range that has been determined for α is stored with the other information about the judgment session and can be leveraged to recommend a suitable feedback structure in a future session. This is done as follows. Given the requirements model r_{curr}, for which a feedback structure shall be recommended, we retrieve information about the past judgment session that is based on the requirements model r_{past} that is most similar to r_{curr}, i.e. for all requirements models r that led to past judgment sessions, the inequality $sim_{req}(r,r_{curr}) \leq sim_{req}(r_{past},r_{curr})$ holds. If more than one past judgment session with this property exists, we select the most recent one. The parameter α that has been determined based on the user's judgment behavior in the selected session (the midpoint of the determined range) is used in the recommendation process of the current judgment session.

VI. EVALUATION

In this section, we will analyze the quality of the recommendations produced by our algorithm for judgment target recommendation, i.e. we will investigate how well the suggested feedback structures fit to the users' judgment preferences. Thereby, we will first introduce the test data that have been used to evaluate our approach and will then describe the tests that have been performed and detailedly discuss their results.

A. Test Data

Test data comprising of pairs of consumer requirements and judged item aspects were acquired by providing test users with a set of predefined requirements profiles and asking them to indicate the item aspects they would like to judge. They were obtained using a number of 20 test users aged between 25 and 30 years (10 male and 10 female), of which 9 had a computer science background and 11 not. The elicitation procedure was as follows. Each test person was provided with a questionnaire comprising of 12 requirements profiles covering typical preferences and requirements of consumers looking for computer items. 6 of the profiles referred to desktop PCs and another 6 to digital watches. Requirements profiles of a single type varied with respect to the attributes that were important to the user, the attribute values that were acceptable and with respect to the consumer's preferences related to these values and attributes. We chose the indicated requirements domains, since the test users were familiar with buying and rating products online as well as with the considered product types. Moreover, the two product domains share common attributes, e.g. for both an attribute manufacturer is defined, and thus allow to demonstrate the recommendation procedures ability to infer about a judgment providers rating preferences across item domains. Fig. 3 shows a sample requirements profile from the questionnaire. As can be seen, each requirements

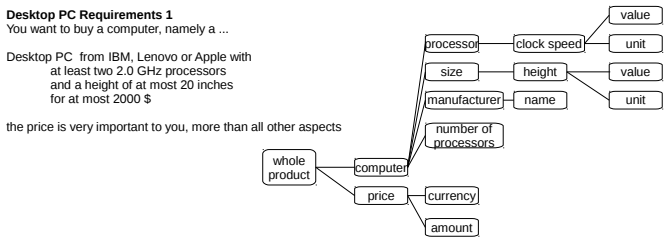

Fig. 3. Excerpt of the questionnaire

profile was given in terms of a textual description (Fig. 3 left). In addition to that, a tree comprising of product attributes that might potentially be judged was provided (Fig. 3 right). Before being asked to complete the questionnaire, the test persons were given a short introduction to how to read and how to proceed with the form. Moreover, they were instructed about valid selections of judgment targets. For each user and requirements profile, it was recorded which and how many aspects were judged.

B. Tests and Results

To evaluate the accuracy of our feedback structure recommendation algorithm, we ran a number of tests with different subsets of the test data elicited from each test user. The basic procedure for each test was as follows. Starting with no information about a user's previous judgment behavior, several judgment sessions were performed. During each session, one of the requirements profiles in the test data was selected. After that, the system proposed a feedback structure using the suggested recommendation algorithm provided with knowledge about the user's judgment behavior in the previous judgment sessions of the test. After being provided with the recommended feedback structure, the suggested structure was adjusted according to the structure that has been actually judged by the test user (as indicated in the questionnaire response). This was done by expanding/collapsing feedback structure nodes.

The quality of the proposed feedback structure was measured as the edit distance between the recommended feedback structure and the feedback structure that was judged by the test user. More formally, we counted the number of expand/collapse operations the user would have to perform to create the structure whose leaves he finally judged. The rationale behind this measure is, that the edit distance is a direct measure of the user's effort to produce the desired structure and thus, in our opinion, is a good measure for the quality of the recommendation.

For each test user, we performed 3 separate test runs using different subsets of the test data elicited from the user. Those were two tests *HomDesk* and *HomDigi*, each based on a set of requirements profiles that were homogeneous in the sense that the involved profiles referred to the same type of computer item, and a test *Het*, that was based on a set of heterogeneous requirements profiles that referred to different types of computer items. While test *HomDesk* referred to the set of desktopPC related profiles, test *HomDigi* referred to the set of digital watch related profile for which judgment data have been elicited. During the tests, each requirements profile was considered twice, i.e. the processing sequence was *requirements profile 1, . . ., requirements profile 6, requirements profile 1, . . ., requirements profile 6*. Test *Het* considered both, the desktop PC and digital watch profiles, where requirements profiles referring to desktop PCs and profiles referring to digital watches were processed alternately. The resulting processing sequence was *desktop PC profile 1, digital watch profile 1, . . ., desktop PC profile 6, digital watch profile 6*.

The Figs. 4 and 5 show the results of our tests. While the blue curve(s) of each figure indicate(s) the mean edit distance over all test users for each judgment session of a test, the red curve(s) refer(s) to the mean edit distance that would result, if the system would always suggest to provide a single overall judgment for the considered item (default, if no experiences about past judgment sessions are available). The error bars indicate the 95% confidence interval for the mean edit distance.

Fig. 4 depicts the evaluation results for test *HomDesk*. The first half of the curves depicted in this figure refers to sessions were the judgments based on a certain requirements profile are given for the first time, while the second half shows the edit distance resulting from sessions that refer to requirements profiles on whose basis judgments have been already been provided, i.e. where the user's judgment behavior based on that requirements profile is known to the system. Consider the first part of the curves presented in Fig. 4. As can be seen,

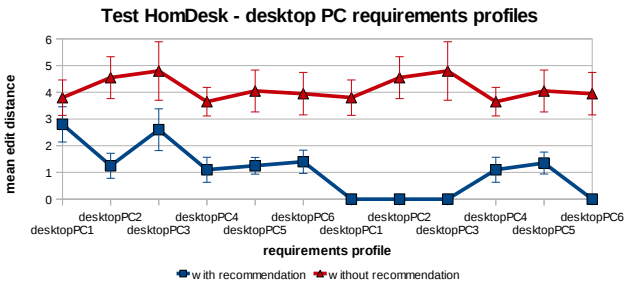

Fig. 4. Results of test *HomDesk*

the edit distances resulting from the recommended feedback structures are lower than that resulting from always suggesting the default structure and decrease as the knowledge acquired in previous judgment sessions increases. This indicates that the test users' judgment behavior in a future judgment sessions was successfully inferred from their past judgment behavior that was based on similar requirements models. The mean number of editing operations required to be performed by the users to create the desired feedback structure was 1.4 ± 0.4 (after the 6th judgment session). This is a reduction by 1.4 ± 0.3 operations. As expected, the edit distance further decreases to 0 during the second part of the test. This is due to the fact that knowledge about the users' judgment preferences with respect to the involved requirements models is already available. The results show that the proposed recommendation mechanism typically correctly references this knowledge, i.e. by comparing the similarities of the involved requirements profiles chooses the right judgment experience to generate a recommendation. This does not hold for the *desktop PC profiles 4 and 5*, which are not correctly referenced. Instead, judgment experiences that are based on similar requirements profiles were leveraged for the recommendation. As a result, we observe an edit distance larger than 0 for these models. Similar results have been observed for test *HomDigi*, but due to space restrictions are omitted.

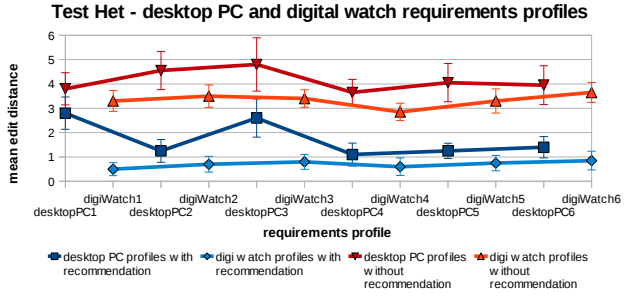

Fig. 5. Results of test *Het*

Fig. 5 depicts the evaluation results for test *Het* involving both types of requirements profiles. As can be seen, the curves depicted in this figure coincide with the first half of the corresponding curves observed for the homogeneous tests. Hence, the fact that judgment experiences that refer to different kinds of consumer requirements are available to the recommendation mechanism, has no negative impact on its accuracy, i.e. judgment experiences referring to different types of requirements profiles are referenced correctly. Instead, we observe that the mean edit distance observed for *digital watch profile 1* is even lower than that during the homogeneous test (not depicted). This implies that, in contrast to the first session of test *HomDigi*, where we cannot draw on previous judgment experiences, the knowledge acquired from the first desktop PC session was successfully exploited to improve the recommendation accuracy for *digital watch profile 1*.

VII. Conclusion

In this paper, we demonstrated how multi-criteria judgments, that are machine-comprehensible and meaningful, can be elicited and how users can be supported in that process. Our main contribution is an algorithm that suggests aspects that might be judged by a consumer. Our evaluation results show, that the proposed procedure effectively adjusts to a user's personal judgment preferences and thus provides helpful support for rating provision.

References

[1] T. N. Company, "Global trends in online shopping a nielsen global consumer report," June 2010.

[2] M. Anderson, "Study: 72recommendations," http://searchengineland.com/study-72-of-consumers-trust-online-reviews-as-much-as-personal-recommendations-114152, 3 2012.

[3] F. Bronner and R. de Hoog, "Vacationers and ewom: Who posts, and why, where, and what?" *Journal of Travel Research*, vol. 50, no. 1, pp. 15–26, 2011.

[4] A. Josang, R. Ismail, and C. Boyd, "A survey of trust and reputation systems for online service provision," *Decis. Support Syst.*, vol. 43, no. 2, pp. 618–644, 2007.

[5] T. A. Lakiotaki K., Matsatsinis N., "Multi-criteria user modeling in recommender systems," *IEEE Intelligent Systems*, 2010.

[6] G. Adomavicius, N. Manouselis, and Y. Kwon, "Multi-criteria recommender systems," in *Recommender Systems Handbook*, 2011, pp. 769–803.

[7] D. Jannach, Z. Karakaya, and F. Gedikli, "Accuracy improvements for multi-criteria recommender systems." in *ACM Conference on Electronic Commerce*, B. Faltings, K. Leyton-Brown, and P. Ipeirotis, Eds. ACM, 2012, pp. 674–689.

[8] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *In Proceedings of KDD '04*. New York, NY, USA: ACM, 2004, pp. 168–177. [Online]. Available: http://portal.acm.org/citation.cfm?id=1014073

[9] S. Brody and N. Elhadad, "An unsupervised aspect-sentiment model for online reviews." in *HLT-NAACL*. The Association for Computational Linguistics, 2010, pp. 804–812.

[10] J. J. McAuley and J. Leskovec, "Hidden factors and hidden topics: understanding rating dimensions with review text," in *Recommender Systems*, 2013.

[11] C. M. K. Cheung and M. K. O. Lee, "What drives consumers to spread electronic word of mouth in online consumer-opinion platforms." *Decision Support Systems*, vol. 53, no. 1, pp. 218–225, 2012.

[12] F. Klan and B. König-Ries, "A conversational approach to semantic web service selection," in *EC-Web*, 2011, pp. 1–12.

[13] L. Chen and P. Pu, "Hybrid critiquing-based recommender systems," in *IUI*, 2007, pp. 22–31.

[14] U. Küster, B. König-Ries, M. Klein, and M. Stern, "Diane - a matchmaking-centered framework for automated service discovery, composition, binding and invocation," in *Proceedings of WWW2007*, Banff, Alberta, Canada, May 2007.

[15] B. K.-R. Friederike Klan, "Supporting consumers in providing meaningful multi-criteria judgments," in *In Proceedings of PRSAT 2010 in conjunction with RecSys2010, Barcelona, Spain*, 2010.

[16] P. Jaccard, "Étude comparative de la distribution florale dans une portion des alpes et des jura," *Bulletin de la Société Vaudoise des Sciences Naturelles*, vol. 37, pp. 547–579, 1901.