Explaining Recommendations Based on Feature Sentiments in Product Reviews

Li Chen
Department of Computer Science, Hong Kong Baptist University
Hong Kong, China
lichen@comp.hkbu.edu.hk

Feng Wang
Department of Computer Science, Hong Kong Baptist University
Hong Kong, China
fwang@comp.hkbu.edu.hk

ABSTRACT
The explanation interface has been recognized important in recommender systems as it can help users evaluate recommendations in a more informed way for deciding which ones are relevant to their interests. In different decision environments, the specific aim of explanation can be different. In high-investment product domains (e.g., digital cameras, laptops) for which users usually attempt to avoid financial risk, how to support users to construct stable preferences and make better decisions is particularly crucial. In this paper, we propose a novel explanation interface that emphasizes explaining the tradeoff properties within a set of recommendations in terms of both their static specifications and feature sentiments extracted from product reviews. The objective is to assist users in more effectively exploring and understanding product space, and being able to better formulate their preferences for products by learning from other customers' experiences. Through two user studies (in form of both before-after and within-subjects experiments), we empirically identify the practical role of feature sentiments in combination with static specifications in producing tradeoff-oriented explanations. Specifically, we find that our explanation interface can be more effective to increase users' product knowledge, preference certainty, perceived information usefulness, recommendation transparency and quality, and purchase intention.

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous.

Author Keywords
Recommender systems; explanation interfaces; sentiment analysis; product reviews; user study.

INTRODUCTION
Explanation, by definition, refers to "making clear by giving a detailed description" [38]. In recommender systems, it either serves to explain the recommendation process (i.e., the logic of underlying algorithm) [16, 36, 2], or justify why the recommendation might be good for a user [37, 41, 15]. According to [14, 15, 39], the types and aims of explanation can be different in different recommender systems. For instance, in collaborative filtering (CF) systems, the explanation can be in form of a histogram with grouping of neighbors in different rating categories (high, medium and low) for increasing system transparency and persuasiveness [16]. Keywords or user-tags based approaches can help users to make qualified and good decisions in CF or content-based recommender systems [37, 41]. However, most of existing explanation interfaces emphasize explaining a single item, rather than a set of multiple recommendations by revealing their tradeoff relationship (i.e., relative *pros* and *cons*). As claimed in [31, 32], the advantage of tradeoff-oriented explanation is that it can support users to more effectively compare different items and better judge which ones are relevant to their needs.

Therefore, in this work, we focus on generating the tradeoff-oriented explanation for multiple recommendations. Considering that the incorporation of item features into explanation can be helpful to support users' decision making in general [15], we target to extract informative features from item descriptions, such as product reviews, to disclose their tradeoff properties. To be specific, we develop a novel explanation interface for preference-based recommender systems that have been mainly applied in complex, high-investment product domains[1] in e-commerce environments. Because of the high financial risk users will bear in such domains if they make a wrong choice, they usually demand high decision accuracy. On the other hand, as they may possess little product knowledge at the start because of infrequent experiences with buying those products, their initial preferences for products are likely to be uncertain and incomplete [29, 30, 40]. Preference-based recommender systems hence aim to elicit users' attribute preferences on site and adapt recommendations to their preference refinement behavior [26, 12, 19, 7]. The goal of explanation in such system should thus be to educate users about product knowledge by explaining what products do exist, help them to resolve preference conflict and construct stable preferences, and eventually enable them to make informed and confident decisions.

[1] Here, the definition of "investment" rests on purchase price (for example, movies and music are of low investment, and digital cameras and laptops are of high investment) [23, 38].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
IUI 2017, March 13–16, 2017, Limassol, Cyprus.
Copyright © 2017 ACM ISBN 978-1-4503-4348-0/17/03 ...$15.00.
http://dx.doi.org/10.1145/3025171.3025173

In a related work [32, 6], they propose a category structure, called Preference-based Organization (Pref-ORG), to organize recommendations in preference-based recommender systems. They use category title to explain items within the same category in terms of their tradeoff relationship with the top candidate, based on static specifications (e.g., "*This group of digital cameras have better values at effective pixels, weight, price, but worse value at screen size*") (see Figure 1).

Our current work can be considered as an extension of Pref-ORG, with particular emphasis on exploiting product reviews to expose tradeoff opportunities. As a matter of fact, product reviews have become an essential part of product description along with its static specifications for a new buyer to assess the product's quality, especially if the product like a digital camera can not be directly sensed in the online environment. Many buyers tend to seek for advices from other customers' opinions to reduce decision uncertainty [22, 43]. In order to save their efforts in manually processing textual reviews, some review summarization interfaces have been developed [3, 44, 9], but little work has incorporated review information into recommendation explanations [8, 28]. In our view, feature sentiments as embedded in product reviews (such as other customers' opinions towards a camera's *effective pixels*, *image quality*, and *ease of use*) can be beneficial to augment the system's explanatory power by better explaining products' tradeoff properties and helping users to construct more stable preferences via learning from other customers' experiences.

In short, relative to related work, our contributions are fourfold: 1) We combine feature sentiments in product reviews with static specifications to build both product profile and users' multi-attribute preference model. 2) We develop an algorithm to generate sentiment-based recommendations and group products with similar tradeoff properties into categories. 3) We conduct user studies to compare our prototype with the original Pref-ORG, which demonstrate its significant performance in increasing users' product knowledge, preference certainty, perceived information usefulness, recommendation transparency and quality, and purchase intention. 4) Based on the findings, we conclude several design guidelines for enhancing tradeoff-oriented explanations in preference-based recommender systems.

In the following, we first introduce related work on explanation interfaces in recommender systems. We then present our methodology of producing tradeoff-oriented explanations for multiple recommendations, which integrates both products' static specifications and feature sentiments. The next section gives the details of two experiments, including materials, participants, experimental procedure, evaluation metrics and hypotheses, followed by results of each experiment and the summary of their major findings. Finally, we discuss the results' practical implications and draw the conclusion.

RELATED WORK

The importance of explanation interfaces in providing system transparency and thus increasing user acceptance has been recognized in a number of fields, such as expert systems, medical decision support systems, intelligent tutoring systems, and data exploration systems [45, 31]. In recent years, being able

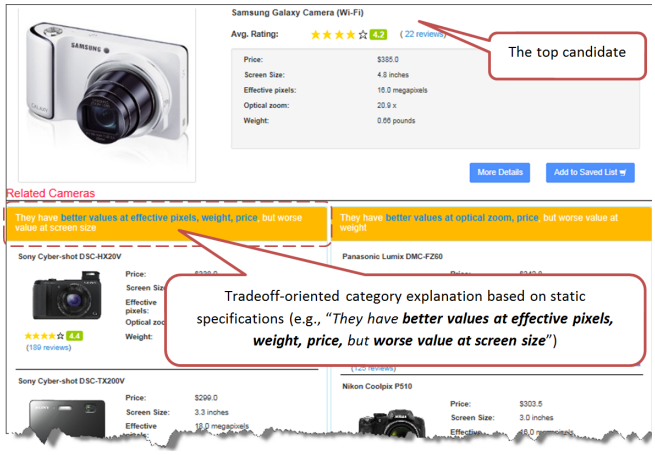


Figure 1. Preference-based organization interface (Pref-ORG) based on products' static specifications (where recommended products are grouped into categories, with the category title to explain multiple products' shared tradeoff properties).

to effectively explain recommendations has been regarded important in recommender systems [16, 36, 38]. Some researchers have explored the impact of explanation interfaces in different types of recommender systems. For example, Herlocker *et al.* studied explanation for collaborative filtering (CF) systems and found that the histogram with grouping of neighbor ratings is the best in terms of convincing users to try the recommended item [16]. In CF and content-based systems, keywords or user-tag based approaches have also been proposed to explain the content similarity between the current recommendation and the target user's previously liked items (e.g., "*Item A is suggested because it contains features X and Y that are also included in items B, C, and D, which you have liked*") [37, 41, 15]. In case-based reasoning systems, they stressed the need of helping users to explore product space by explaining what alternative products exist in the data set [35]. They concretely developed compound critiques (e.g., "*Less Optical Zoom & More Digital Zoom & A Different Storage Type*"), each representing a set of available products, for users to pick as improvement on the current recommendation. In preference-based recommender systems that focus on performing products' ranking according to users' multi-attribute preference model as explicitly acquired on site, the explanation has mainly served to educate users about product properties (such as their tradeoff relationship) and assist them in constructing and refining preferences [31, 32]. A typical work is the Preference-based Organization interface (Pref-ORG) [6], which differs from standard single-item explanation in that it can expose shared tradeoff opportunities (*pros* and *cons* in respect of products' static specifications) of a group of products relative to the user's top recommendation.

A survey on explanation studies in recommender systems has been made in [39], which summarizes several popular explanation styles (e.g., case-based, content-based, collaborative-based, knowledge-based) from various aspects: presentation type (e.g., histogram, tag cloud, text), displayed content (e.g., ratings, neighbors, keywords), and explanatory aim (e.g., in-

creasing system *transparency*, system *persuasiveness*, user *trust*, user *satisfaction*, users' *decision efficiency*, and users' *decision effectiveness*). Some earlier rating-based explanations, such as the histogram with grouping of neighbor ratings [16], mainly focused on providing system transparency and persuasiveness, which however may cause users to overestimate item quality [2]. Therefore, recent studies have put more emphasis on achieving other explanatory aims. In [2], the keyword-style explanation was shown significantly more effective at allowing users to make accurate decisions. In another work [15], the authors compared ten different explanation types and found that the content-based tag cloud explanations are more effective and helpful to increase users' satisfaction with the explanation, though decision efficiency is not optimized. Trust, as a long-term relationship that can be established between users and recommender systems, has been investigated in [31, 32]. They showed the Preference-based Organization interface (Pref-ORG), which is mainly based on products' specification content, is more competent in inducing users' trust in recommendations as users possess higher intention to return to use the system.

Thus, it can be seen that the content-based explanation, in general, performs more effectively in terms of supporting users' decision making. However, so far few studies have exploited *product reviews* to enhance such explanation. Given that reviews normally include the reason why people like or dislike a product based on their usage experiences, we can capture multi-faceted nature of user opinions towards the product's attributes from their reviews and then incorporate the opinion information into explaining recommendations.

To the best of our knowledge, one work is most related to ours [27, 28]. They have also conducted feature-based sentiment analysis in product reviews for generating explanations. Concretely, they associated each discovered feature with an importance degree (according to its occurring frequency) and a sentiment score for building both item's and user's profiles. Each recommended product was then explained in terms of *pros* and *cons* features it contains that matter to the target user, as well as being compelling relative to alternative recommendations. Their user study showed that this interface can improve the explanation's overall clarity and helpfulness [28]. However, similar to traditional explanation design, their explanation is for a single item, rather than stressing the tradeoff relationship within multiple recommendations. Moreover, they only consider reviews, but ignore users' preferences for products' static specifications. Another limitation is that they primarily serve users who have written reviews before, not *new users* who have zero or few review histories (the popular cold-start problem in high-investment product domains).

In contrast, our explanation is targeted to highlight attributes that are able to distinguish a group of recommended products from others in terms of both feature sentiments and static specifications. Therefore, each explanation is applied to multiple products, instead of a single one, for facilitating users to make product comparison. The similarity to their work is that we also divide feature sentiments into *pros* and *cons*, but our objective is to expose the inherent tradeoff relationship between products and use explanations to group and select products, so that the system can assist *new users* in constructing and stabilizing preferences.

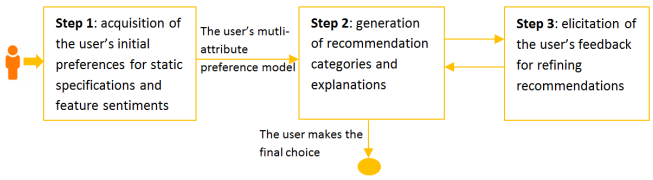


Figure 2. The three major steps of generating recommendations and tradeoff-oriented explanations in our method.

INTERFACE DESIGN AND IMPLEMENTATION

As shown in [31, 32], the advantage of categorical structure of explanations in the Preference-based Organization interface (Pref-ORG) is that it enables users to compare products more efficiently and inspires their trusting intention, relative to the traditional list design where recommended products are explained separately (the single-item explanation). Therefore, our interface design basically follows the principles established for Pref-ORG [31]: *Principle 1*: Categorizing remaining recommendations according to their similar tradeoff properties relative to the top candidate; *Principle 2*: Proposing improvements and compromises in the category title using conversational language, and keeping the number of tradeoff attributes under five in each title to avoid information overload; *Principle 3*: Diversifying the categories in terms of their titles and contained recommendations; *Principle 4*: Displaying actual products in each category.

The novelty of our current work is that, in addition to static specifications, we extract feature sentiments from product reviews to produce tradeoff-oriented explanations. For example, one category's explanation is "*They have better value at optical zoom, better opinions at effective pixels, weight, but worse value at price*", where "*better opinions at effective pixels, weight*" reflects sentiment improvements of customers on "effective pixels" and "weight" of this category's products in comparison to the top candidate. In this section, we first explain how we identify features and their associated sentiments from product reviews, and then describe three major algorithmic steps of implementing our explanation interface (see Figure 2).

Extraction of feature sentiments from textual reviews. Feature-based sentiment analysis (also called opinion mining) has become an established subject in the field of natural language processing [24]. There have been various approaches developed to capturing reviewers' opinions towards specific features that were mentioned in their textual reviews, such as statistical methods [17, 18], machine learning methods like those based on lexicalized Hidden Markov Model (L-HMMs) [20] and Conditional Random Fields (CRFs) [34], and Latent Dirichlet Allocation (LDA) based methods for identifying features directly [25].

In our work, we adopt a popular statistical approach proposed in [17, 18], because it is more domain-independent without the need of manual labeling and model training. The experiment also showed this approach can achieve reasonable accuracy [17]. Concretely, given that a feature is normally expressed as a noun or noun phrase (e.g., "display", "size", "image quality") in raw reviews, we first perform association rule mining to discover all frequent noun words/phrases [2] as feature candidates (that exceed certain frequency threshold, i.e., with minimum support value 1%). Those feature candidates are then mapped to a pre-defined set of attributes (e.g., the digital camera's price, screen size, effective pixels, optical zoom, weight, image quality, video quality, ease of use), by computing the lexical similarity via WordNet [13] (for example, "resolution", "pixels", and "megapixels" are mapped to "effective pixels").

The sentiment associated with each identified feature is further extracted by looking for adjacent adjective words (within 3-word distance to the feature) in a review sentence (e.g., "*vivid*" in "*The LCD display screen provides very vivid previews of photos or video*"). The polarity value of an adjective word w is formally determined via SentiWordNet [11] as follows: $polarity(w) = neg(w) * r_{min} + pos(w) * r_{max} + obj(w) * \frac{r_{min}+r_{max}}{2}$, where $neg(w)$, $pos(w)$, and $obj(w)$ respectively denote three polarity scores, i.e., negativity, positivity, and objectivity, as defined in SentiWordNet ($pos(w) + neg(w) + obj(w) = 1$), and r_{min} and r_{max} are respectively set to 1 and 5 for restricting the value of $polarity(w)$ into the range [1,5] (from "least negative" to "most positive"). In the case there appear odd number of negation words (e.g., "not", "don't", "no", "didn't") in the same review sentence, the adjective word's polarity value will be reversed.

Next, an attribute's sentiment score is calculated by aggregating all polarity values of features that are mapped to that attribute of a product:

$$senti_i(p) = \frac{1}{|R(a_i,p)|} \sum_{r \in R(a_i,p)} senti_i(r) \tag{1}$$

where $R(a_i,p)$ is the set of reviews w.r.t. product p that contain opinions on attribute a_i, and $senti_i(r)$ is defined as:

$$senti_i(r) = \frac{\sum_{w \in SW(a_i,r)} polarity(w)^2}{\sum_{w \in SW(a_i,r)} polarity(w)} \tag{2}$$

where $SW(a_i,r)$ is the set of sentiment words that are associated with all features mapped to attribute a_i in a review r, and $polarity(w)$ is the polarity value of word w.

In consequence, each product can be defined as $\{(a_i, speci_i, senti_i)_{1:m}, (a_j, senti_j)_{m+1:n}\}$, where a_i indicates the attribute (such as "effective pixels") that has both static specification value $speci_i$ (e.g., 18 megapixels) and sentiment score $senti_i$ (e.g., 4 out of 5), and a_j refers to the attribute that only has sentiment score (such as "image quality", "video quality", and "ease of use").

[2]through Core-NLP package for Part-of-Speech (POS) tagging: http://nlp.stanford.edu/software/corenlp.shtml

Step 1. Modeling of users' initial preferences

A user's preferences are then modeled on the above-defined attributes, for which we revise the traditional weighted additive form of value functions [21, 6] to incorporate the user u's preferences for feature sentiments:

$$Utility_u(p) = \sum_{i=1}^{m} W_i * [\alpha * V(speci_i(p)) + (1-\alpha) * V(senti_i(p))] + \sum_{j=m+1}^{n} W_j * V(senti_j(p)) \tag{3}$$

where $Utility_u(p)$ represents the utility of a product p in terms of its matching degree to the user's preferences. $V(speci_i(p))$ denotes the user's preference for attribute a_i's specification value $speci_i$, $V(senti_i(p))$ gives her/his preference for the attribute's sentiment score $senti_i$, and α indicates the relative importance between $speci_i$ and $senti_i$ for the user (default set as 0.5). W_i (also W_j) is the attribute's importance relative to the other attributes (default set as 3 out of 5). Theoretically, this model is grounded on Multi-Attribute Utility Theory (MAUT) [21], which can explicitly resolve users' preference conflict by accommodating attribute tradeoffs.

In our system, each user will be initially asked to specify her/his value preferences and weights for product attributes. For example, for the camera's "effective pixels", the user's preference for its specification may be ">= 16 megapixels" [3] and that for its sentiment score be ">= 4" (be positive), and the weight preference is 5 (indicating the highest importance). Default functions will be assigned to attributes that the user does not state any actual preferences (e.g., for price, "the cheaper, the better"; for feature sentiments, "the higher, the better"). Then, according to her/his preferences, all products will be ranked by their utilities as computed via Equation (3), and the top k products with higher utilities will be recommendation candidates.

Step 2. Generation of recommendation categories and explanations

Recommendation categories. Among those top k recommendation candidates, except the ranked 1*st* one that is taken as the top candidate, each product is converted into a tradeoff vector $\{(a_i, tradeoff_i)\}$, where $tradeoff_i$ is either *improved* ($\uparrow$) or *compromised* ($\downarrow$), indicating the attribute's specification value or sentiment is better or worse than that of the top candidate. Formally, it is defined in the following way:

$$tradeoff_i(p',p) = \begin{cases} \uparrow_v & \text{if } 1 \le i \le m \text{ and } V(speci_i(p')) > V(speci_i(p)) \\ \downarrow_v & \text{if } 1 \le i \le m \text{ and } V(speci_i(p')) < V(speci_i(p)) \\ \uparrow_o & \text{if } 1 \le i \le n \text{ and } senti_i(p') > senti_i(p) \\ \downarrow_o & \text{if } 1 \le i \le n \text{ and } senti_i(p') < senti_i(p) \end{cases} \tag{4}$$

where p is the top candidate and p' is the currently considered product. For attribute a_i ($1 \le i \le m$), there can be two possible tradeoff values assigned to it, i.e., better or worse specification value ($\uparrow_v$ or $\downarrow_v$), and better or worse sentiment ($\uparrow_o$ or $\downarrow_o$); for attribute a_j ($m+1 \le j \le n$), there is only one tradeoff value in respect of its sentiment, $\uparrow_o$ or $\downarrow_o$.

[3]In this case, $V(x) = \begin{cases} 1 & \text{if } x \ge 16 \\ 1 - \frac{|x-16|}{max(a)-min(a)} & \text{otherwise} \end{cases}$

We then run Apriori algorithm [1] over all candidates' tradeoff vectors in order to retrieve frequently occurring subsets of $(a_i, tradeoff_i)$ pairs. Products containing the same subset of tradeoff pairs are then grouped into one category. For example, "$(optical\ zoom, \uparrow_v), (effective\ pixels, \uparrow_o), (weight, \uparrow_o), (price, \downarrow_v)$" refers to a category of products that all have higher optical zoom, better opinions on effective pixels and weight, but higher price, than the top candidate. Since a product might belong to multiple categories if it shares different tradeoff properties with different groups of products, we further make category selection in order to find those that are not only with higher tradeoff benefits but also diverse among each other.

Selection of categories. Each category is assigned with a score $F(C)$ to reflect two characteristics [6]: *tradeoff benefit* (*pros* against *cons*) relative to the top candidate, and *diversity degree* with the other selected categories SC.

$$F(C) = TradeoffBenefit(C) \times Diversity(C, SC) \quad (5)$$

Specifically, the tradeoff benefit is calculated as:

$$TradeoffBenefit(C) = (\sum_{i=1}^{|C|} W_i \times tradeoff_i) \times (\frac{1}{|SR(C)|} \sum_{p \in SR(C)} Utility_u(p)) \quad (6)$$

where C denotes the currently concerned category represented by a set of $(a_i, tradeoff_i)$ pairs, and $SR(C)$ denotes the set of products within C (ranked by their utilities). The tradeoff value $tradeoff_i$ is default set as 0.75 if *improved* ($\uparrow$), or 0.25 if *compromised* ($\downarrow$).

The *diversity degree* of C is defined in terms of both the category title (i.e., the set of $(a_i, tradeoff_i)$ pairs) and its contained products:

$$Diversity(C, SC) = \min_{C_i \in SC} ((1 - \frac{|C \bigcap C_i|}{|C|}) \times (1 - \frac{|SR(C) \bigcap SR(C_i)|}{|SR(C)|})) \quad (7)$$

Thus, the first selected category is that with the highest tradeoff benefit, and the subsequent category is selected if it has the highest score $F(C)$ among the remaining non-selected categories. Such selection process ends when the desired N categories are selected.

Category explanations. In the recommendation interface, each category has a title containing the explanation of *pros* and *cons* in its associated $(a_i, tradeoff_i)$ pairs in natural language. For example, "$(optical\ zoom, \uparrow_v), (effective\ pixels, \uparrow_o), (weight, \uparrow_o), (price, \downarrow_v)$" is automatically translated into the explanation "*They have better value at optical zoom, better opinions at effective pixels, weight, but worse value at price*", because $\uparrow_v$ is converted to "*better value at*", $\uparrow_o$ to "*better opinion at*", and $\downarrow_v$ to "*worse value at*". Moreover, in the explanation, attributes with "*better*" properties are displayed in front of those with "*worse*" properties. If several attributes are with the same tradeoff property, the one with higher weight is shown first.

Figure 3 shows an example of explanation interface generated by the above steps, which we call "Mixture View" in our prototype since it is based on both static specifications and feature sentiments.

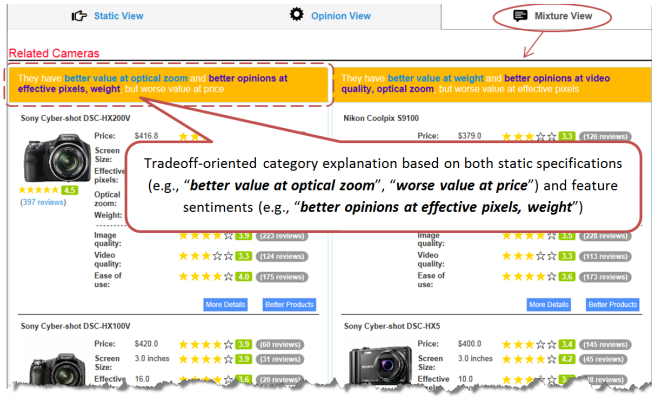


Figure 3. Preference-based organization interface based on both static specifications and feature sentiments ("Mixture View" in our prototype).

Step 3. Elicitation of user feedback for refining recommendations

If the user cannot make a choice in current recommendations, we will elicit her/his feedback for generating new recommendations in the next round. Along with every recommended product, there is a button "Better Products" that the user can click. It actually acts as the "feedback" we aim to elicit. To be specific, in case that the user is interested in one product but not satisfied with all of its attributes, s/he can click this button to see some related products that possess better specification values and/or sentiments. To retrieve those products, we first update the user's preference model. As the product that the user is interested in belongs to a certain category C, we refer to the category's explanation for adjusting the involved attributes' weights (W_i and α in Equation (3)), as follows:

$$W_i = \begin{cases} 5 & \text{if } (a_i, \uparrow_v) \in C \text{ or } (a_i, \uparrow_o) \in C \\ 1 & \text{else if } (a_i, \downarrow_v) \in C \text{ or } (a_i, \downarrow_o) \in C \end{cases} \quad (8)$$

$$\alpha = \begin{cases} 0.8 & \text{if } (a_i, \uparrow_v) \in C \text{ and } (a_i, \uparrow_o) \notin C \\ 0.2 & \text{else if } (a_i, \uparrow_o) \in C \text{ and } (a_i, \uparrow_v) \notin C \end{cases} \quad (9)$$

Then, we are able to rank products by their new utilities and select a new set of candidates. Step 2 will be performed again over those candidates to produce recommendation categories and explanations, and return them on the interface. As shown in Figure 2, the user's interaction with our system will end when s/he makes the final choice.

USER STUDY

We conducted two experiments in order to identify whether our tradeoff-oriented explanations as incorporated with feature sentiments can be more effective to increase users' preference certainty, perceived system competence, and behavioral intention. They used the same materials (the compared interfaces) and evaluation criteria, while the major difference occurs in the evaluation procedure (*before-after* and *within-subjects*).

Materials

We implemented three explanation interfaces (also called explanatory views in the following content): 1) the original

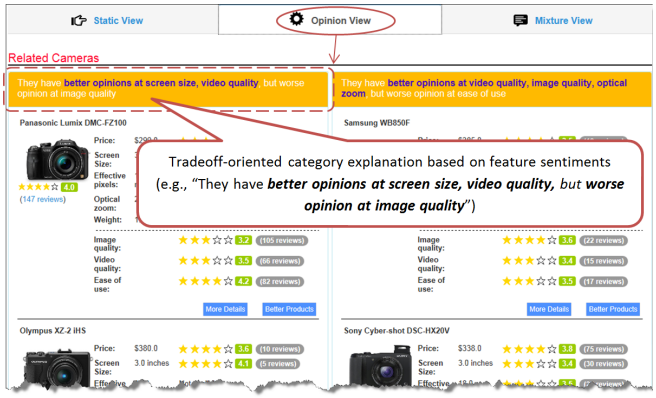


Figure 4. Preference-based organization interface purely based on feature sentiments ("Opinion View" in our prototype).

Pref-ORG [32, 6] that is only based on products' static specifications (see Equation (10) with its definition of tradeoff vector, and Figure 1 with its sample interface that is called "Static View" in our implementation); 2) the one that is purely based on feature sentiments from product reviews, for which the tradeoff vector is defined in Equation (11) and the sample interface is in Figure 4 (called "Opinion View"); 3) the mixture of static specifications and feature sentiments in tradeoff-oriented explanations, as we described in the previous section ("Mixture View" in Figure 3).

$$tradeoff_i(p', p) = \begin{cases} \uparrow_v & \text{if } 1 \le i \le m \text{ and } V(speci_i(p')) > V(speci_i(p)) \\ \downarrow_v & \text{if } 1 \le i \le m \text{ and } V(speci_i(p')) < V(speci_i(p)) \end{cases} \quad (10)$$

$$tradeoff_i(p', p) = \begin{cases} \uparrow_o & \text{if } 1 \le i \le n \text{ and } senti_i(p') > senti_i(p) \\ \downarrow_o & \text{if } 1 \le i \le n \text{ and } senti_i(p') < senti_i(p) \end{cases} \quad (11)$$

In each explanation interface, except the top candidate, the remaining recommendations are grouped into N categories ($N = 4$ in our prototype). The category title explains shared *pros* and *cons* (i.e., tradeoff properties) of products that belong to it. Along with every product, users can view not only its static specifications (e.g., a camera's price, screen size, effective pixels, optical zoom, weight), but also reviewers' average rating (and inferred sentiment scores on major attributes in "Opinion View" and "Mixture View"). Moreover, there are two buttons attached to the product in a category, i.e., "More Details" (to see a full list of specifications and raw reviews) and "Better Products" (to initiate a new round of recommendation; see Step 3 in the previous section). In case a user is interested in "buying" one product, s/he can add it to "Saved List" and go to "Checkout" page to make the final choice.

Our experimental goal was to let each participant evaluate all of the three explanation interfaces, so as to reduce the amount of errors arising from natural variance between individuals. However, if the three interfaces are implemented in three different systems, there will be 18 comparative conditions (3 product types x 6 evaluation orders) according to the within-subjects experimental setup. Therefore, in order to reduce experiment complexity, we implemented the three interfaces in one prototype system, acting as three optional views for users to freely choose and use. Then, through comparing this prototype (hereafter named as Senti-ORG) with the original Pref-ORG system (only with the "Static View"), we are able to verify not only the benefits that feature sentiments bring (i.e., "Opinion View"/"Mixture View" vs. "Static View"), but also the effect of combining static specifications and feature sentiments in "Mixture View" on users' decision behavior and subjective perceptions by means of analyzing their activities (clicking actions) and self-reported preference for this view.

We implemented both Senti-ORG and Pref-ORG in two product catalogs, digital camera and laptop. The digital camera catalog consists of 346 products crawled from *Amazon.com*. Each product is described by 6 primary attributes (i.e., brand, price, screen size, effective pixels, optical zoom, weight) and 54 full specifications. The total amount of reviews posted to these products (till February, 2014) is 46,434 (mean = 134.2, min = 10), where we extracted 8 features (5 mapped to the above-mentioned 6 primary attributes, and 3 only having sentiment scores that are *image quality*, *video quality*, and *ease of use*). As for laptop, there are 303 products (with totally 16,047 reviews and average 52.96 reviews per product). Each laptop is described by 11 attributes, 4 of which only have sentiment scores (brand, price, processor speed, RAM, hard drive, screen size, weight, *display*, *battery life*, *portability*, and *quietness*).

Two Experiments

Before-After Experiment

One experiment is based on the *before-after* setup, which asks users to make a choice first in Pref-ORG (with the task of "*finding a product that you would purchase if given the opportunity*"), and then evaluate Senti-ORG (to "*decide whether the product you have chosen in the previous step is truly the best for you, or you prefer another one.*"). The two compared systems are with the same product type (e.g., digital camera). When users come to use the second system, the choice they made in the first system is displayed at the top as the reference product, with the three explanatory views shown below (the default is Static View).

Therefore, this experiment requires each participant to evaluate both systems in a fixed order, which is with the main focus on analyzing her/his switching behavior: *would s/he be motivated to make a better choice in Senti-ORG or not?*

Within-Subjects Experiment

The second experiment follows *within-subjects* design [4], for which we developed four conditions (2 product types x 2 evaluation orders) so as to avoid any carryover effects [4]. Specifically, the manipulated factors are systems' order (Pref-ORG first or Senti-ORG first) and product catalogs' order (digital camera first or laptop first). For example, one condition requests a user to evaluate Senti-ORG first for finding a digital camera to "buy" (in this setting, the default explanatory view in Senti-ORG is randomly chosen), and then use

[4]Generally, *before-after* is also a type of within-subjects design. Here we use *within-subjects* to refer to the particular setup that the two evaluated systems are in two different product catalogs and their evaluation order is altered among participants.

Pref-ORG to find a laptop to "buy". The advantage of this experiment is that the two compared systems are independent (with different types of product), so that the product knowledge users have gained when they used the first system would not affect their behavior and perceptions in using the second one. We can hence compare the two systems under equivalent circumstances.

Online Procedure

Each participant was randomly assigned to either *before-after* or *within-subjects* experiments. We developed an online procedure, which contains instructions, evaluated interfaces and questionnaires, for the participant to carry out the experiment at her/his convenience. At the beginning, s/he was debriefed on the experiment's objective and upcoming tasks. In particular, s/he was asked to compare two product finder systems and determine which one is more effective in terms of supporting her/him to make a purchase decision. Thereafter, a short questionnaire was to be filled out about her/his demographics, e-commerce experiences, and product knowledge. The participant then started evaluating the two systems one by one according to the assigned order. In order for the user to be familiar with each evaluated system, a demo video (lasting around 2 minutes) was first played before s/he formally starts the evaluation task. After evaluating each system, the participant was prompted to answer a questionnaire about her/his overall opinions and comments on the used interfaces. At the end of the experiment, s/he was additionally asked to compare the two evaluated systems and indicate which one s/he prefers to use for searching for products.

Participants

We launched the two experiments through various channels (e.g., campus email, advertisement in forums, crowdsourcing via Amazon Mechanical Turk). Each participant was rewarded with around US$2 as incentive. At last, we collected 118 users' data. We filtered out records which were with incomplete/invalid answers to our questionnaires or lasted less than 5 minutes. As a result, 94 users' records remain (43 females). 42 of them took part in the *before-after* experiment, and 52 were involved in the *within-subjects* experiment (with each condition performed by 13 users). Those users cover different age ranges ("<= 20" (3), "21-30" (52), "31-40" (27), "41-50" (7), and ">=51" (5)), and are from over 6 different countries (India, USA, China, France, Japan, Nigeria). They also have different education majors (IT, education, fiance, sociology, literature, mathematics, biology, etc.), degrees (high school, associate degree, Bachelor, Master, PhD), and professions (banker, homemaker, librarian, student, engineer, accountant, manager, etc.). They all visited e-commerce stores before, and 94.7% of them have purchased products online.

Evaluation Criteria and Hypotheses

As indicated in "Introduction", the goal of explanation in preference-based recommender systems is mainly to explain why the recommended products are presented, help users to construct stable preferences, and enable them to make informed and confident decisions. Therefore, in our experiments, we mainly measured four aspects of users' subjective perception with the used system: *decision effectiveness* (product knowledge, preference certainty), *perceived system competence* (transparency, information usefulness, recommendation quality, recommendation novelty), *system trustworthiness*, and *behavioral intention* (intention to purchase). Most of them are in accordance with the definitions of explanatory aims in [39] (i.e., transparency, effectiveness, persuasiveness) and trust construct in [31, 32]. It is worth mentioning that for *decision effectiveness*, we measured not only users' perceived effectiveness, but also their actual effectiveness through the before-after experiment (i.e., whether the user would switch to a better choice that more suits her/his preferences when the explanations are based on feature sentiments).

Table 1 lists questions for assessing these subjective variables (each question was responded on a 5-point Likert scale, e.g., from "strongly disagree" to "strongly agree"), most of which were proven with strong content validity in related literatures [33]. Moreover, in the post-task questionnaire after using Senti-ORG, we added one question for knowing the user's favorite view among the three options, i.e., Static View, Opinion View, and Mixture View ("*Which view of showing the related products do you prefer?*"). In the final questionnaire after the user evaluated both systems (Pref-ORG and Senti-ORG), we asked her/his overall preference ("*Which system do you prefer to use for searching for products (like cameras or laptops)?*").

Besides, we also measured participants' objective behavior, including the time they consumed in using each system, the session length (the number of interaction cycles they were involved in refining recommendations), clicking actions (e.g., the chosen explanatory view(s), the selected product(s), the examined product specification(s)/review(s)), and the final choice. The time and session length are particularly related to *decision efficiency*, which indicates whether explanation can make it faster for users to decide the best choice or not [39].

Our main hypothesis was that the incorporation of feature sentiments into tradeoff-oriented explanations would be more helpful for enhancing users' *decision effectiveness*, especially *product knowledge* and *preference certainty*, as they would be able to learn from other customers' experiences to stabilize their attribute preferences. Moreover, their perception with the system's *recommendation quality* would be strengthened, since the recommended products can be justified based on other users' opinions towards them, rather than simply based on static specifications in the original Pref-ORG. In consequence, users would be more inclined to purchase the product that they choose in Senti-ORG (w.r.t. *purchase intention*). As for *decision efficiency*, given prior studies showed that task completion time is not significantly correlated with user satisfaction and trust [31, 15], we postulate it would still not be optimized in our system and users would tend to spend time in analyzing explanations for making good decisions.

RESULTS

Before-After Experiment Results

We first report the before-after experiment's results. By analyzing 42 users' objective behavior in using the two compared systems, we find 47.6% (20 out of 42) users made new choices after they were presented with explanations based on feature

Subjective variable	Question (each responded on a 5-point Likert scale)	*Before-after*		*Within-subjects*	
		Pref-ORG	Senti-ORG	Pref-ORG	Senti-ORG
Product knowledge	*How would you rate your knowledge about xxx?*	3.74 vs. 3.64	**3.88**** vs. 3.64	3.62 vs. 3.52	**3.52**** vs. 3.29
Preference certainty	*I am very certain about what I need in respect of each attribute.*	4.02	**4.24****	3.73	**4***
Info. usefulness	*This system helped me discover some useful info.*	4	**4.33****	3.90	**4.29****
Explanatory ability	*The system explained to me why the products were recommended.*	4.21	4.21	3.88	**4.13****
Recom. transparency	*I understood why the items were returned to me.*	3.86	**4.07****	3.69	**4.04****
Recom. quality	*The system returned to me some good suggestions.*	4.095	**4.33***	3.83	**4.04***
Recom. novelty	*The system helped me discover new products.*	4.14	**4.40****	3.96	**4.23***
Trust	*The system can be trusted.*	3.93	4.07	3.69	**3.87***
Satisfaction	*Overall, I am satisfied with the system.*	4.02	4.19	3.88	**4.12***
Purchase intention	*I would purchase the product I just chose if given the opportunity.*	3.86	**4.12****	3.75	**3.96***

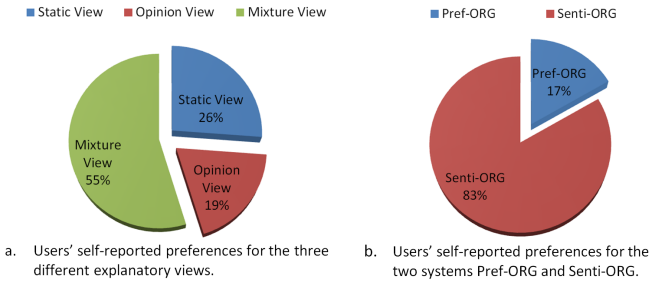
Note: the number is mean value; ** $p < 0.05$, * $0.5 < p < 1$ via Paired Samples t-Test (the product knowledge is relative to the user's initial knowledge level).

Table 1. Assessment of users' subjective perceptions and comparison results.

sentiments in Senti-ORG. More specifically, in Senti-ORG, 23 and 22 users clicked Opinion View and Mixture View respectively, among whom 5 and 10 evaluated products (by clicking "More Details") in the two views. The percents of users who switched to better choices are correspondingly 20% (1 out of 5) in Opinion View and 70% (7 out of 10) in Mixture View. Some users (5) were even motivated to read feature-specific reviews (the highlighted parts in raw review texts that commented a specific feature). Regarding the other objective measures, it shows users took significantly less time and interaction cycles in using Senti-ORG (3.71 mins vs. 5.67 mins in Pref-ORG, $t = 5.19$, $p < 0.01$; 0.79 vs. 1.74 in Pref-ORG, $t = 3.91$, $p < 0.01$), which should be because they already gained certain familiarity with the products when they used the first system Pref-ORG, so the time and cycles consumed in revising their preferences in Senti-ORG were shortened.

We then compared those participants' subjective perceptions (see Table 1). It shows their product knowledge is significantly increased after using Senti-ORG (mean = 3.88 vs. 3.64 initially, $t = -2.68$, $p = 0.01$), but not after Pref-ORG (3.74 vs. 3.64 initially, $t = -1.43$, $p = 0.16$). Moreover, users' preference certainty is significantly improved owning to the experience with Senti-ORG (4.24 vs. 4.02 in Pref-ORG, $t = -2.29$, $p = 0.027$). Other significant differences between Senti-ORG and Pref-ORG are found with respect to users' perceived information usefulness, i.e., whether the system helps them discover useful product information (4.33 vs. 4, $t = -2.396$, $p = 0.02$), recommendation transparency (4.07 vs. 3.86, $t = -2.04$, $p = 0.048$), recommendation quality (4.33 vs. 4.095, $t = -1.95$, $p = 0.058$), and recommendation novelty (4.40 vs. 4.14, $t = -2.55$, $p = 0.01$). In terms of users' purchase intention, it is also significantly higher in Senti-ORG than Pref-ORG (4.12 vs. 3.86, $t = -2.05$, $p = 0.047$).

Figure 5 additionally shows the distribution of users' expressed preferences for the three explanatory views in Senti-ORG, i.e., Static View, Opinion View, and Mixture View, as well as their overall preferences for the two compared systems Pref-ORG and Senti-ORG. It can be seen that around 54.8% (23) users prefer Mixture View that integrates both static specifications and feature sentiments into explanations, followed by 26.2% (11) users who prefer Static View that is only based on static specifications, and 19% (8) preferring Opinion View purely based on feature sentiments. As for their overall preferences, 83.3% (35 out of 42) participants are in favor of Senti-ORG.



a. Users' self-reported preferences for the three different explanatory views.

b. Users' self-reported preferences for the two systems Pref-ORG and Senti-ORG.

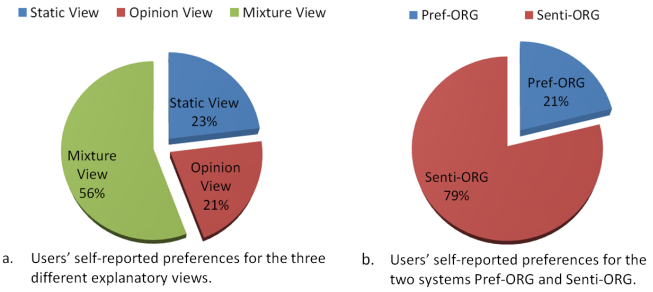
Figure 5. Users' self-reported preferences for the used interfaces in the before-after experiment.

Within-Subjects Experiment Results

The within-subjects experiment's results further consolidate the before-after findings. As mentioned before, this setup can avoid possible carryover effects, because there are different comparative conditions that enable system evaluations to be conducted in different orders and product catalogs.

The results indicate that participants still prefer Senti-ORG in general (41 vs. 11 users who like Pref-ORG, covering 78.8% of all 52 participants). Among the three explanatory views in Senti-ORG, still, more users prefer Mixture View (29 out of 52, 55.8%), which is largely higher than the numbers of users who favor Static View (12, 23.1%) and Opinion View (11, 21.1%). Figure 6 illustrates the distributions.

As for objective measures, in Senti-ORG, users consumed significantly longer time (7.26 mins vs. 5.03 mins in Pref-ORG, $t = -2.83$, $p < 0.01$), evaluated more products in details (1.83 vs. 1.38, $t = -2.02$, $p = 0.048$), and more frequently read products' raw reviews (0.85 vs. 0.44, $t = -1.96$, $p = 0.055$). In-depth analysis of their review-reading behavior reveals that 23 users read raw reviews (6 of them read raw feature-specific review texts), which is higher than the number of users (13) who read raw reviews in Pref-ORG. Another interesting observation is that in Senti-ORG fewer users (37) examined products' full specifications (relative to 41 in Pref-ORG). It hence suggests that incorporating feature sentiments into explanations might motivate users to consult more detailed review information, while less counting on static specifications. Further analysis of users' choices made in Senti-ORG shows 27 participants located their final choices in sentiment-based



a. Users' self-reported preferences for the three different explanatory views.

b. Users' self-reported preferences for the two systems Pref-ORG and Senti-ORG.

Figure 6. Users' self-reported preferences for the used interfaces in the within-subjects experiment.

explanation interfaces (12 in Opinion View and 15 in Mixture View), and 25 participants made choices in Static View.

Regarding users' subjective perceptions, most of responses are consistent with those of the before-after comparison (see Table 1). In particular, their product knowledge level is significantly increased after using Senti-ORG (3.52 vs. 3.29 initially, $t = -2.198$, $p = 0.03$), but there is no significant improvement in Pref-ORG (3.62 vs. 3.52 initially, $t = -1.22$, $p = 0.23$). Users also perceived the product information provided in Senti-ORG more useful than in Pref-ORG (4.29 vs. 3.90, $t = -3.05$, $p < 0.01$), and Senti-ORG more competent in terms of recommendation transparency (4.04 vs. 3.69, $t = -2.76$, $p < 0.01$) and explanatory ability (4.13 vs. 3.88, $t = -2.095$, $p = 0.04$). Moreover, there are some marginally significant enhancements ($0.05 < p < 0.1$) obtained by Senti-ORG, as for users' preference certainty (4 vs. 3.73 in Pref-ORG, $t = -1.82$, $p = 0.075$), system trustworthiness (3.87 vs. 3.69, $t = -1.92$, $p = 0.0598$), perceived recommendation quality (4.04 vs. 3.83, $t = -1.80$, $p = 0.078$), perceived recommendation novelty (4.23 vs. 3.96, $t = -1.85$, $p = 0.07$), overall satisfaction (4.12 vs. 3.88, $t = -1.73$, $p = 0.0898$), and purchase intention (3.96 vs. 3.75, $t = -1.75$, $p = 0.086$).

User Comments

After each experiment, we asked the participant to freely comment the systems they just used and explain why they prefer one over another. Their free comments made the reasons more explicit as to why Senti-ORG was preferred to the original Pref-ORG by the majority of users.

Most of users expressed favorable appraisal of the explanations incorporated with feature sentiments in Senti-ORG. For instance, "*The interface makes it possible to see what other buyers think of the product.*" (before-after) "*I feel that it provided more detailed information in regards to each attribute listed for the products as well as providing user opinions which I prefer to take into consideration when purchasing an item such as this.*" (before-after) "*It gave reviewers/customers more weight in helping potential customers choose a product.*" (within-subjects) "*Because the system combines customers' comment and recommended items together.*"(within-subjects)

Some users even mentioned the impact of feature sentiments on increasing their confidence about what they want: "*The ratings really helped reassure me that the product is a good option.*" (before-after) "*It gives the reviews of the products, so we are more confident for our choosing product.*" (within-subjects) "*The user interfaces in-cooperates user review and that is from using the device thus creating my trust in wanting to purchase the product.*" (within-subjects) "*I prefer this system because if a user doesn't have a big knowledge about the product it will provide very easy detailed score on different attributes of the product. A user will be able to evaluate on which attribute is most suited for his/her need and will be able to compare and judge the product by the score and number of reviews it has received.*" (within-subjects)

Offering the three explanatory views (Static View, Opinion View, and Mixture View) in the same system were also favored by some users: "*I like the options of the static, mixed, or sentiment view.*" (before-after) "*The user interface gave more options for viewing the results: static, sentiment, mixed.*" (within-subjects) "*It gave three different ways to look at the competing products. Overall, it allowed the user to evaluate each item in different ways to determine the best option for them.*" (within-subjects)

Moreover, from users' comments, we can identify the main reason behind their reported preference for Mixture View: explaining products' *pros* and *cons* in terms of both reviewer opinions and specifications makes them better understand products and more quickly compare them. For instance, "*The system had more information on various attributes in the mixed version. The reviews had highlighted the pros and cons which makes understanding the product better.*" (before-after) "*The ease of finding both public opinion and specifications on the various attributes of the different cameras without having to look at the reviews section of each is very convenient.*" (before-after) "*I liked how you could have mixed stats and opinions.*" (within-subjects) "*It allowed me to quickly compare what other users had thought of the product's attributes and the product in general, as well as showing me the attribute values of the product.*" (within-subjects) "*I think item specifics alone are very important, but customer reviews/opinions about a product and how well a product performs based on each feature are also very important.*" (within-subjects)

As for the reason why some participants like Pref-ORG, we find it is mainly owing to its simplicity, i.e., only providing product specification information these users care, which is particularly common among those who are knowledgable about the products initially. For instance, "*It was less cluttered. It provided me with more information pertaining to specifications of the product. It was more efficient to find a desirable product.*" (before-after) "*I can directly see the specifications that I want to know.*" (within-subjects) "*Easy to use and simple interface.*" (within-subjects) "*I felt like the interface was easier to view and not cluttered.*" (within-subjects)

PRACTICAL IMPLICATIONS

Explanation Interface Design

Previous work on Pref-ORG demonstrated its advantage over standard single-item explanations, in terms of inspiring users' trust in recommendations [31, 32]. In this paper, results of our experiments show that incorporating reviewers' feature

sentiments into Pref-ORG can further improve its ability with respect to not only user trust, but also users' product knowledge, preference certainty, perceived information usefulness, recommendation transparency and quality, and purchase intention. The significant statistical findings in combination with users' free comments enable us to derive some design principles, which we believe can be constructive for developing more effective explanation interfaces in preference-based recommender systems:

- ***Principle* [5] 1**: Categorizing recommendations according to their similar tradeoff properties relative to the top candidate, in terms of both feature sentiments and static specifications, which can be helpful for users to increase product knowledge and preference certainty.*
- ***Principle* 2**: Highlighting attributes' pros and cons that distinguish a group of products from others in the category title, which can facilitate users to easily compare products.*
- ***Principle* 3**: Providing different explanatory views (i.e., Static View, Opinion View, and Mixture View) in the system, which can allow users to examine products from different angles. However, the interface's simplicity and ease of use may be compromised.*

Recommender Algorithm Development

As described in the section "Interface Design and Implementation", the feature sentiments extracted from product reviews are not just used to produce explanations in our system. Actually, they are fused into the whole procedure of user preference modeling and recommendation generation. Given that users perceive our system's recommendation quality significantly higher than that of the original Pref-ORG, we believe similar benefits could be obtained by other sentiment-based recommender algorithms that also emphasize the role of feature sentiments in ranking products [5]. For instance, Wang *et al.* have aimed to predict new users' unstated attribute preferences by means of locating similar reviewers whose value preferences can be recovered from their feature sentiments [42]. Dong *et al.* have focused on retrieving products similar to the current user's query product according to their reviews' feature popularity and relative sentiment improvement [10].

In addition, we believe our work can offer some new insights into improving current sentiment-based recommender methods: 1) Feature sentiments can be combined with static specifications for eliciting and constructing users' multi-attribute preference model. 2) Recommendations can be organized into categories, rather than in a ranked list, to display diverse groups of items with representative tradeoff benefits (*pros* vs. *cons*). 3) The sentiment-based explanation can be utilized to elicit users' feedback, so as for the system to incrementally refine its recommendations.

CONCLUSION

In this paper, we present a novel method of implementing tradeoff-oriented explanations in preference-based recommender systems. In comparison to related work on Preference-based Organization (Pref-ORG) [31, 32, 6], we particularly take into account feature sentiments as extracted from product reviews to generate recommendations and explanations in a category structure. Through measuring users' objective behavior and subjective perceptions as well as collecting their free comments in both before-after and within-subjects experiments, we have several interesting findings: 1) Incorporating feature sentiments into Pref-ORG can be effective to increase users' product knowledge, preference certainty, perceived information usefulness, recommendation transparency and quality, and purchase intention. 2) The explanation interface's actual effectiveness was also measured, which indicates almost half of users made better choices after using Senti-ORG. 3) As for decision efficiency, it shows users spent more time in making decisions in Senti-ORG, which is consistent with related literatures' observation [31, 15] that efficiency is not necessarily correlated to users' decision effectiveness and perceived system competence. 4) Three design principles are derived from our experiment results. In particular, given that Mixture View is preferred by the majority of users, we recommend explaining products' tradeoff properties (*pros* and *cons*) in terms of both feature sentiments and static specifications.

As mentioned in "Introduction", our work is under assumption that users are mostly new to the product domain (such as high-investment digital cameras and laptops). Therefore, our target has been to exploit other customers' review data to serve the current new user. In the future, it will be interesting to extend the work in other scenarios where users are not only information seekers but also contributors (e.g., review writers for low-investment items like movies, music). In such case, user reviews might be utilizable by the system to infer their initial attribute preferences and hence generate more relevant explanations at the start. Another direction is to conduct more experiments to test whether our explanation interface could be helpful for establishing *long-term* trust relationship between users and recommender systems [15], in addition to one-time experience in the current evaluation. The inherent tradeoff between the two explanatory aims *decision efficiency* and *effectiveness* will also be verified in our future studies.

[5]The asterisk is for differentiating from the original principles [31].

ACKNOWLEDGMENTS

We thank all participants who took part in our experiments. We also thank Hong Kong Research Grants Council (RGC) for sponsoring the research work (under projects ECS/HKBU211912 and RGC/HKBU12200415).

REFERENCES

1. Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. 1993. Mining Association Rules Between Sets of Items in Large Databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data (SIGMOD '93)*. ACM, NY, USA, 207–216. DOI: http://dx.doi.org/10.1145/170035.170072
2. Mustafa Bilgic and Raymond J. Mooney. 2005. Explaining Recommendations: Satisfaction vs. Promotion. In *Proceedings of the Workshop Beyond Personalization, in conjunction with IUI'05*.
3. Giuseppe Carenini and Lucas Rizoli. 2009. A Multimedia Interface for Facilitating Comparisons of Opinions. In

Proceedings of the 14th International Conference on Intelligent User Interfaces (IUI '09). ACM, NY, USA, 325–334. DOI: http://dx.doi.org/10.1145/1502650.1502696

4. Gary Charness, Uri Gneezy, and Michael A. Kuhn. 2012. Experimental Methods: Between-Subject and Within-Subject Design. *Journal of Economic Behavior & Organization* 81, 1 (2012), 1 – 8. DOI: http://dx.doi.org/10.1016/j.jebo.2011.08.009
5. Li Chen, Guanliang Chen, and Feng Wang. 2015. Recommender Systems Based on User Reviews: The State of the Art. *User Modeling and User-Adapted Interaction* 25, 2 (June 2015), 99–154. DOI: http://dx.doi.org/10.1007/s11257-015-9155-5
6. Li Chen and Pearl Pu. 2010. Experiments on the Preference-based Organization Interface in Recommender Systems. *ACM Transactions on Computer-Human Interaction* 17, 1, Article 5 (April 2010), 5:1–5:33 pages. DOI:http://dx.doi.org/10.1145/1721831.1721836
7. Li Chen and Pearl Pu. 2012. Critiquing-based Recommenders: Survey and Emerging Trends. *User Modeling and User-Adapted Interaction* 22, 1-2 (April 2012), 125–150. DOI: http://dx.doi.org/10.1007/s11257-011-9108-6
8. Li Chen and Feng Wang. 2014. Sentiment-enhanced Explanation of Product Recommendations. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14 Companion)*. ACM, NY, USA, 239–240. DOI: http://dx.doi.org/10.1145/2567948.2577276
9. Li Chen, Feng Wang, Luole Qi, and Fengfeng Liang. 2014. Experiment on Sentiment Embedded Comparison Interface. *Knowledge-Based Systems* 64 (2014), 44 – 58. DOI:http://dx.doi.org/10.1016/j.knosys.2014.03.020
10. Ruihai Dong, Michael P. O'Mahony, Markus Schaal, Kevin McCarthy, and Barry Smyth. 2013. Sentimental Product Recommendation. In *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys '13)*. ACM, NY, USA, 411–414. DOI: http://dx.doi.org/10.1145/2507157.2507199
11. Andrea Esuli and Fabrizio Sebastiani. 2006. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*. 417–422.
12. Alexander Felfernig, Gerhard Friedrich, Dietmar Jannach, and Markus Zanker. 2006. An Integrated Environment for the Development of Knowledge-Based Recommender Applications. *International Journal of Electronic Commerce* 11, 2 (Dec. 2006), 11–34.
13. Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.
14. Gerhard Friedrich and Markus Zanker. 2011. A Taxonomy for Generating Explanations in Recommender Systems. *AI Magazine* 32, 3 (2011), 90–98. http://dblp.uni-trier.de/rec/bibtex/journals/aim/FriedrichZ11
15. Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. 2014. How Should I Explain? A Comparison of Different Explanation Types for Recommender Systems. *International Journal of Human-Computer Studies* 72, 4 (April 2014), 367–382. DOI: http://dx.doi.org/10.1016/j.ijhcs.2013.12.007
16. Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining Collaborative Filtering Recommendations. In *Proceedings of ACM Conference on Computer Supported Cooperative Work (CSCW '00)*. ACM, NY, USA, 241–250. DOI: http://dx.doi.org/10.1145/358916.358995
17. Minqing Hu and Bing Liu. 2004a. Mining and Summarizing Customer Reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04)*. ACM, NY, USA, 168–177. DOI: http://dx.doi.org/10.1145/1014052.1014073
18. Minqing Hu and Bing Liu. 2004b. Mining Opinion Features in Customer Reviews. In *Proceedings of the 19th National Conference on Artifical Intelligence (AAAI'04)*. AAAI Press, 755–760. http://dl.acm.org/citation.cfm?id=1597148.1597269
19. Shiu-Li Huang. 2011. Designing Utility-based Recommender Systems for e-Commerce: Evaluation of Preference-Elicitation Methods. *Electronic Commerce Research and Applications* 10, 4 (July 2011), 398–407.
20. Wei Jin, Hung Hay Ho, and Rohini K. Srihari. 2009. OpinionMiner: A Novel Machine Learning System for Web Opinion Mining and Extraction. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '09)*. ACM, NY, USA, 1195–1204.
21. Ralph L. Keeney and Howard Raiffa. 1976. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Wiley, New York.
22. Young Ae Kim and Jaideep Srivastava. 2007. Impact of Social Influence in e-Commerce Decision Making. In *Proceedings of the Ninth International Conference on Electronic Commerce (ICEC '07)*. ACM, NY, USA, 293–302. DOI: http://dx.doi.org/10.1145/1282100.1282157
23. David N. Laband. 1991. An Objective Measure of Search versus Experience Goods. *Economic Inquiry* 29, 3 (1991), 497–509. DOI: http://dx.doi.org/10.1111/j.1465-7295.1991.tb00842.x
24. Bing Liu. 2010. Sentiment Analysis and Subjectivity. In *Handbook of Natural Language Processing, Second Edition*, Nitin Indurkhya and Fred J. Damerau (Eds.). CRC Press, Taylor and Francis Group, Boca Raton, FL. ISBN 978-1420085921.

25. Julian McAuley and Jure Leskovec. 2013. Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text. In *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys '13)*. ACM, NY, USA, 165–172. DOI: http://dx.doi.org/10.1145/2507157.2507163

26. Kevin McCarthy, James Reilly, Lorraine McGinty, and Barry Smyth. 2005. Experiments in Dynamic Critiquing. In *Proceedings of the 10th International Conference on Intelligent User Interfaces (IUI '05)*. ACM, NY, USA, 175–182. DOI: http://dx.doi.org/10.1145/1040830.1040871

27. Khalil Muhammad, Aonghus Lawlor, Rachael Rafter, and Barry Smyth. 2015. Great Explanations: Opinionated Explanations for Recommendations. In *Proceedings of the 23rd International Conference on Case-Based Reasoning Research and Development (ICCBR'15)*. Springer International Publishing, 244–258. DOI: http://dx.doi.org/10.1007/978-3-319-24586-7_17

28. Khalil Ibrahim Muhammad, Aonghus Lawlor, and Barry Smyth. 2016. A Live-User Study of Opinionated Explanations for Recommender Systems. In *Proceedings of the 21st International Conference on Intelligent User Interfaces (IUI '16)*. ACM, NY, USA, 256–260. DOI: http://dx.doi.org/10.1145/2856767.2856813

29. John W. Payne, James R. Bettman, and Eric J. Johnson. 1993. *The Adaptive Decision Maker*. Cambridge University Press.

30. John W. Payne, James R. Bettman, and David A. Schkade. 1999. Measuring Constructed Preferences: Towards a Building Code. *Journal of Risk and Uncertainty* 19, 1 (1999), 243–270. DOI: http://dx.doi.org/10.1023/A:1007843931054

31. Pearl Pu and Li Chen. 2006. Trust Building with Explanation Interfaces. In *Proceedings of the 11th International Conference on Intelligent User Interfaces (IUI '06)*. ACM, NY, USA, 93–100. DOI: http://dx.doi.org/10.1145/1111449.1111475

32. Pearl Pu and Li Chen. 2007. Trust-inspiring Explanation Interfaces for Recommender Systems. *Knowledge-Based Systems* 20, 6 (Aug. 2007), 542–556. DOI: http://dx.doi.org/10.1016/j.knosys.2007.04.004

33. Pearl Pu, Li Chen, and Rong Hu. 2011. A User-centric Evaluation Framework for Recommender Systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems (RecSys '11)*. ACM, NY, USA, 157–164. DOI: http://dx.doi.org/10.1145/2043932.2043962

34. Luole Qi and Li Chen. 2010. A Linear-chain CRF-based Learning Approach for Web Opinion Mining. In *Proceedings of the 11th International Conference on Web Information Systems Engineering (WISE'10)*. Springer-Verlag, 128–141.

35. James Reilly, Kevin McCarthy, Lorraine McGinty, and Barry Smyth. 2005. Explaining Compound Critiques. *Artificial Intelligence Review* 24, 2 (Oct. 2005), 199–220. DOI:http://dx.doi.org/10.1007/s10462-005-4614-8

36. Rashmi Sinha and Kirsten Swearingen. 2002. The Role of Transparency in Recommender Systems. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems (CHI EA '02)*. ACM, NY, USA, 830–831. DOI: http://dx.doi.org/10.1145/506443.506619

37. Panagiotis Symeonidis, Ros Nanopoulos, and Yannis Manolopoulos. 2008. Providing Justifications in Recommender Systems. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 38, 6 (Nov 2008), 1262–1272. DOI: http://dx.doi.org/10.1109/TSMCA.2008.2003969

38. Nava Tintarev and Judith Masthoff. 2012. Evaluating the Effectiveness of Explanations for Recommender Systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (Oct. 2012), 399–439. DOI: http://dx.doi.org/10.1007/s11257-011-9117-5

39. Nava Tintarev and Judith Masthoff. 2015. Explaining Recommendations: Design and Evaluation. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer US, Boston, MA, 353–382. DOI: http://dx.doi.org/10.1007/978-1-4899-7637-6_10

40. Amos Tversky and Itamar Simonson. 1993. Context-dependent Preferences. *Management Science* 39, 10 (Oct. 1993), 1179–1189. DOI: http://dx.doi.org/10.1287/mnsc.39.10.1179

41. Jesse Vig, Shilad Sen, and John Riedl. 2009. Tagsplanations: Explaining Recommendations Using Tags. In *Proceedings of the 14th International Conference on Intelligent User Interfaces (IUI '09)*. ACM, NY, USA, 47–56. DOI:http://dx.doi.org/10.1145/1502650.1502661

42. Feng Wang, Weike Pan, and Li Chen. 2013. Recommendation for New Users with Partial Preferences by Integrating Product Reviews with Static Specifications. In *Proceedings of the 21th International Conference on User Modeling, Adaptation, and Personalization (UMAP'13)*. Springer Berlin Heidelberg, 281–288. DOI: http://dx.doi.org/10.1007/978-3-642-38844-6_24

43. Jianan Wu, Yinglu Wu, Jie Sun, and Zhilin Yang. 2013. User Reviews and Uncertainty Assessment: A Two Stage Model of Consumers' Willingness-to-pay in Online Markets. *Decision Support Systems* 55, 1 (2013), 175–185.

44. Koji Yatani, Michael Novati, Andrew Trusty, and Khai N. Truong. 2011. Review Spotlight: A User Interface for Summarizing User-generated Reviews Using Adjective-noun Word Pairs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, NY, USA, 1541–1550. DOI: http://dx.doi.org/10.1145/1978942.1979167

45. L. Richard Ye and Paul E. Johnson. 1995. The Impact of Explanation Facilities on User Acceptance of Expert Systems Advice. *MIS Quarterly* 19, 2 (June 1995), 157–172. DOI:http://dx.doi.org/10.2307/249686