A Cross-Cultural Analysis of Explanations for Product Reviews

John O'Donovan
Dept. of Computer Science
University of California, Santa Barbara, CA, USA
jod@cs.ucsb.edu

Shinsuke Nakajima
Faculty of Computer Science and Engineering
Kyoto Sangyo University, Kyoto, Japan
nakajima@cse.kyoto-su.ac.jp

Tobias Höllerer
Dept. of Computer Science
University of California, Santa Barbara, CA, USA
holl@cs.ucsb.edu

Mayumi Ueda
[3]Faculty of Economics
University of Marketing and Distribution Sciences, Kobe, Japan
Mayumi_Ueda@red.umds.ac.jp

Yuuki Matsunami
Faculty of Computer Science and Engineering,
Kyoto Sangyo University,
g1245108@cc.kyoto-su.ac.jp

Byungkyu Kang
Dept. of Computer Science
University of California, Santa Barbara, CA, USA
bkang@cs.ucsb.edu

ABSTRACT

Cosmetic products are inherently personal. Many people rely on product reviews when choosing to purchase cosmetics. However, reviewers can have tastes that vary based on personal, demographic or cultural background. Prior work has discussed methods for generating attribute-based explanations for item ratings on cosmetic products, based on associated text-based reviews. This paper focuses on evaluating explanation interfaces for product reviews and related attributes. We present the results of a cross-cultural user study that evaluates five associated explanation interfaces for cosmetic product reviews across groups of participants from three different cultural backgrounds. We applied a 3 by 2 within subjects experimental design in a user study (N=150) to evaluate effects of UI design and personalization on a range of user experience metrics in a cosmetics shopping scenario. Results of the study show that 1) Korean and Japanese speakers chose the most complex UI more often than English speakers. 2) older participants also preferred more options in cosmetic product selection, regardless of cultural background. 3) personalization of product ratings did not show an effect on user experience. 4) Attribute-based explanations were preferred over star-ratings for all three cultures. 5) Rating propensity evaluation showed that Japanese provided significantly higher ratings than Korean or English participants, and that Females provided higher ratings than Males, regardless of background.

CCS Concepts

•**Human-centered computing** → *HCI design and evaluation methods; User models; User studies;*

Keywords

User Experience, Explanation, Decision Making, User-Centric Evaluation

1 Introduction

Over the last 25 years, recommender systems have attempted to help users find the right information at the right time [15]. More recently, the proliferation of e-commerce applications supports buying and selling products in the global market with relatively little effort. Increasingly, consumers are relying on customer reviews to inform purchasing decisions [8]. In many cases, product reviews are presented in summary form via mechanisms such as star ratings. Such representations, however, typically fail to capture the subtle opinions that exist in the accompanying text-based reviews. In this paper, we build on recent work that automatically extracts attributes and associated ratings from online product reviews [10]. In particular, we focus on understanding how visual representations of various types of extracted item ratings impact user experience and conversion likelihoods in an e-commerce setting, as exemplified in Figure 1. Motivated by recent research that shows the importance of user experience over traditional accuracy metrics in recommender systems [7], we conduct a user experiment to understand how rating display affects user experience. Specifically, we applied a 3 by 2 within subjects design (Table 1) in an online study (N=150) to evaluate effects of UI design and personalization on a user experience metrics in a cosmetics shopping scenario, considering the following research questions:

R1: Do cross-cultural preference differences exist for recommendation interfaces? If so, what are the key predictors of these differences?

R2: Are there cross-cultural preference differences for personalized v/s non-personalized recommender system interfaces?

R3: Are there cross-cultural preference differences between traditional (star-rating) and more granular attribute-based recommender system interfaces?

R4: Are there differences in rating propensities across the three cultures? If so, what are the strongest predictors of observed rating shifts?

The cosmetics domain was used for this study, since they are sold globally and are inherently personal in nature. To explore variances in opinions on the explanation interfaces across different cultural backgrounds, participant groups were sourced from American, Japanese and Korean cultural backgrounds. These particular groups were selected as a representative sample with diverse cultures, and because they are among the fastest growing markets for cosmetics.[1]

2 Related Work

In this study, we focus on explanations and transparency of recommender systems and on the (associated) role of product attributes mined from product reviews. Here, we discuss several related work in these areas.

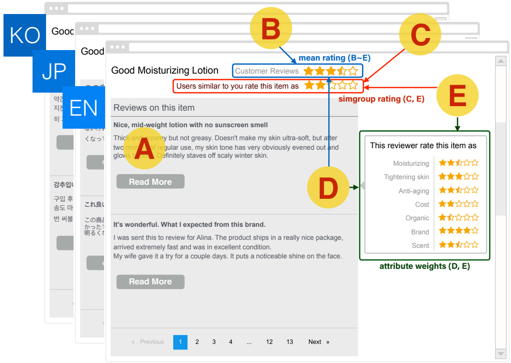
Product Attributes To understand consumer behavior in economics, research has focused on the different attributes and uncertainties that consumers consider when purchasing a product [8, 13]. For buyers, these attributes play important roles when deciding to purchase a product. More importantly, attributes vary widely across product types and users' personal tastes. For example, [3] study the effects of search attributes and provide a comparison between traditional and online supermarkets. A recent study on description and performance uncertainty [4] focused on the difficulty in assessing the product's characteristics. Building on works such as [13] that show advantages of using fine-grained product attributes in the recommendation process, we aim to further our understanding of the role of fine-grained product attribute ratings in consumer decisions.

Explanation and Transparency in Recommendation Within the recommender systems research community, there is an increasing understanding of the need for user-centered evaluations [12]. Recent keynote talks [2] and workshops [14] have helped to highlight the importance of this topic. In this paper, we follow Knijnenburg et al.'s [9] argument for a framework that takes a user-centric approach to recommender system evaluation, beyond the scope of recommendation accuracy. In contrast to that work however, we argue that decision quality is an important evaluation metric that goes beyond the user experience metrics described in [9], and further, that it can be used to explain observed usage patterns for search and recommendation tools. Garcia-Molena [6] described differences and similarities between search and recommendation, and argued that interactive interfaces can help users understand and use these tools in more efficient ways. Along the same vein, it has also been recognized that many recommender systems function as *black boxes*, providing no transparency into the working of the recommendation process, nor offering any additional information to accompany the recommendations beyond the recommendations themselves [7]. To address this issue, static or interactive/conversational explanations can be given to improve the transparency and control of recommender systems. Research on textual explanations in recommender systems to date has been evaluated in wide range of domains (varying from movies to financial advice [5]). From a cross-cultural perspective, Pu and Chen performed a related study that evaluated perceptions of different recommendation interfaces in [1], using subjects from Chinese and Swiss backgrounds. In contrast to their study, which compared a novel UI against a list view and assessed user experience metrics, we focus on the perception of attribute ratings versus traditional less-fine grained ratings, and on the impact of personalization on these perceptions. A second contrast to [1] is that our work explores rating propensity across the different groups.

3 Mining Attribute Ratings

This study builds on a recent work [10] on attribute extraction from online product reviews. Specifically, we posit that more explanations of a given product in the form of multiple attributes with corresponding scores (on five star rating scales), see Figure 1, can provide benefits to potential customers. In the prototype of the proposed recommender system, both personalized information (*Simgroup* ratings: "Users similar to you rate this item as") and multiple product attributes extracted from a review text are added as features. Through an online user study, we apply both novel approaches as controlled variables to the prototype design and investigate the preference of the users to such features across demographic backgrounds, particularly, cultural backgrounds (English, Korean and Japanese).

4 Interface Design



Nice, mid-weight lotion with no sunscreen smell

Thick and creamy but not greasy. Doesn't make my skin ultra-soft, but after two months of regular use, my skin tone has very obviously evened out and glows to boot. Definitely staves off scaly winter skin.

It's wonderful. What I expected from this brand.

I was sent this to review for Alina. The product ships in a really nice package, arrived extremely fast and was in excellent condition.
My wife gave it a try for a couple days. It puts a noticeable shine on the face.

Figure 1: Screenshots of the interface used in the online user study. The annotations A-E show the items that varied in each condition, as shown in Table 1.

We designed a novel user interface for product review pages based on the feedback we received from a preliminary user study (N=100). We performed the study with a simple design layout to test the different visual conditions outlined in Table 1. Participants gave feedback on their preference for each UI in a virtual shopping scenario. They were also required to leave a comment on the interface design. For example, they reported the benefit of the new features, such as "*I like the level of detail it has related to the product*", and suggested preferred features, such as "*More alive colors*" / "*More explanations and ratings*". The collection of 100 comments were manually assessed, and improvements were made to the UI, including shortened review text with "read more" button and breakdown of multiple attributes extracted from the review text on stars. The revised design is shown in Figure 1.

5 Experimental Setup

Figure 1 shows an example of the refactored interface for a sample product review. To test our hypotheses above, a

[1] http://polishcosmetics.pl/Korean-Market-Analysis.pdf

Table 1: Overview of the controlled variables for the online user study.

UI Config	**non-personalized** (no information from similar users)	**personalized** (with social data from similar users)
review text only	**A**: product review text	
review text with star rating	**B**: A + mean rating on stars	**C**: A + mean rating and the rating from active user's *simgroup* on stars
review text, star rating and attributes	**D**: B + attribute weights computed from current review text (on stars)	**E**: C + attribute weights computed from current review text (on stars)

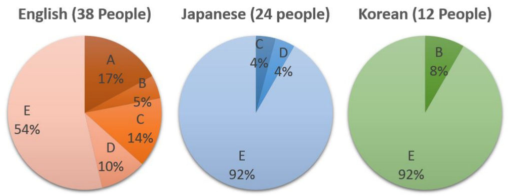


Figure 2: Preferred User Interface by Culture.

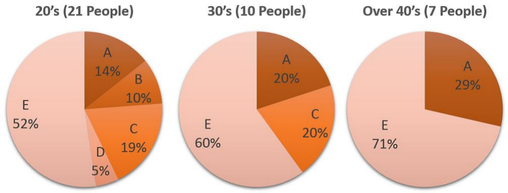


Figure 3: Preferred User Interface by Age.

3x2 within subjects experiment was conducted, controlling for personalization, and rating type, as shown in Table 1. The study (N=150) was performed on the crowdsourcing platform, Amazon Mechanical Turk. Each participant was shown a randomly ordered set of 5 different design layouts corresponding to the treatments in Table 1, and were asked to rank them in order of preference. They were also asked to rate the helpfulness of each. Participants were evenly balanced across cultural backgrounds. All participants were shown with the five interfaces in random order. The content was shown in their primary language based on their cultural background. Overall, participants took between 5-10 minutes doing the study, and were paid $1.50 for their time. Questions were added to test for user attention level and for language proficiency, including identification of differences between UIs and simple math questions written in the appropriate language. After filtering our data based on these metrics, group sizes were 39, 25 and 12 for English, Japanese and Korean, respectively. Participant age ranged between 18-64 with an average of 26. Gender groups were not evenly distributed, as expected for the cosmetics domain, with 70% female and 30% male.

6 Results

Perception and Rating Differences Figure 2 shows the results for the UI ranking task, broken down by age. The result shows a clear preference for design E in all groups, but there is a significant increase in that preference for participants over 40 (shown on the right side). This effect was also seen from 100 participants in the preliminary study. Interface E, shown in Figure 1, shows the most information, and allows users to understand how users similar to them rate individual product attributes. This effect might be a result of specific preferences for cosmetics developing with age, and accordingly, an increased need to explore user ratings on fine grained product attributes (see Figure 3).

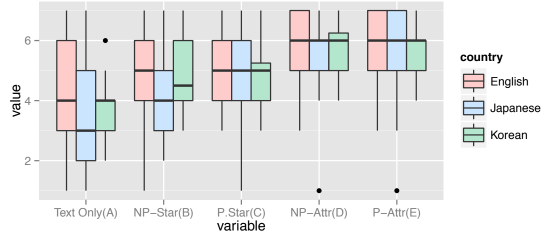


Figure 4: Cross-cultural perspective of helpfulness of the five evaluated interfaces.

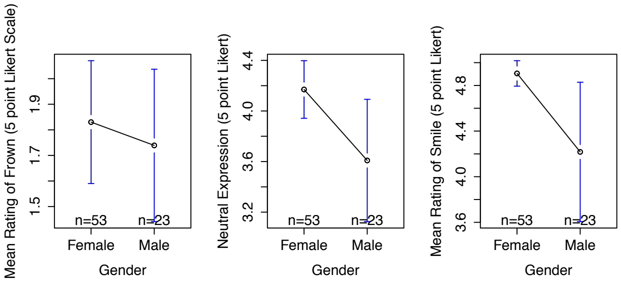


Figure 5: Difference in rating propensity by gender.

Personalization and Rating Type Figure 4 shows the results of perceived usefulness of the interfaces, broken down by cultural groupings. Each UI condition is shown as a group on the x-axis, and each group contains the mean utility score for the three cultural groups. The x-axis groups (UI treatments) are also ranked from left to right based on number of visible features (UI complexity). This graph shows several interesting effects: first, there is a general preference across all groups for the attribute-based representations (groups D and E, on the right side), over less granular, star-ratings or text-based UIs. This is a promising result that indicates that attribute extraction and visualization has a positive effect on Ux. The second interesting result is that within the star-rating group (2nd and 3rd group) and the attribute-rating (4th and 5th) groups there is no notable difference between the personalized and non-personalized treatments. This result tells us that the granularity of presented ratings has more positive impact on user experience than the perception that the ratings come from similar users. To investigate this result in more depth, a followup experiment is planned with a large corpus of product reviews collected from Amazon.com [11] [2] to compute actual similarity scores based on user profiles. This would clearly give better insight into the observed effect. Figure 4 also answers R2, in that there are no significant differences between the cultural groups within

[2] http://jmcauley.ucsd.edu/data/amazon/

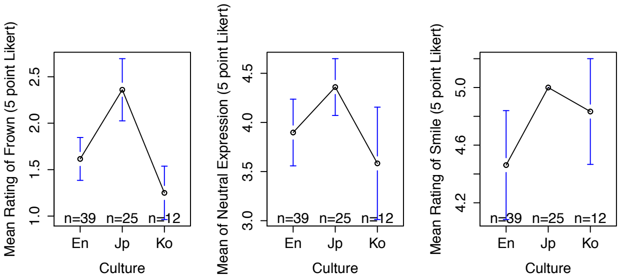


Figure 6: Difference in rating propensity by culture.

each UI treatment, although the Japanese showed a trend towards favoring the more complex UI treatments.

Rating Propensity For some users, rating an item with a specific number of stars can have very different meanings. User ratings on items serve as the basis for most collaborative recommendation techniques, but they tend to ignore such differences when computing neighborhoods for recommendation. Further, little work has been done to understand cross-cultural differences in rating propensity. Since these participant groupings were available our experimental setup, a logical step was to evaluate rating propensities within each of the cultural groups, to serve as both an independent result, and as a weighting factor for the analysis in Figure 4. Each participant was shown three randomly ordered faces, showing expressions with happy, neutral and sad expressions. They were asked to rate the 'happiness' perceived in each on a five point Likert scale. Figure 5 shows the results by gender (for all groups). Interestingly, there is a trend for Females to rate higher than males, and the difference becomes more pronounced for the 'happy' expression, shown on the rightmost plot of Figure 5 with a mean difference of 0.7 (relative increase of 16%, $p<0.005$). Figure 6 shows the results of the rating propensity analysis broken down by cultural group. Again, the graphs represent mean rating for sad, neutral and happy expression ratings from left to right, respectively. Here, we see a clear trend for higher ratings in the Japanese group across all three expressions. While this is only a small-scale initial study, we believe that this is an important result for the study of recommender system performance across different cultures in general, and a follow-up study on propensity of ratings for recommender systems is planned to investigate this further.

7 Discussion and Future Work

This study applied a 3 by 2 within subjects experimental design in a user study (N=150) to evaluate effects of UI design and personalization on a range of user experience metrics in a cosmetics shopping scenario using participant groups from three different cultural backgrounds. Results of the study show that 1) Korean and Japanese speakers chose the most complex UI more often than English speakers. 2) older participants also preferred more options in cosmetic product selection, regardless of cultural background. 3) personalization of product ratings did not show an effect on user experience. 4) attribute-based explanations were preferred over star-ratings for all three cultures. 5) Rating propensity evaluation showed that Japanese had significantly higher ratings than Korean or English, and that Females provided higher ratings than Males, regardless of background. A clear next-step is to evaluate on real product data. The authors plan a follow-up study to compare LDA and dictionary-based approaches to product attribute extraction, and to explore how the resulting attributes can improve explanations, and user profiles for collaborative filtering. Additionally, a more detailed evaluation of the different rating propensities across cultures is underway using a larger number of participants and multiple product domains.

8 References

[1] L. Chen and P. Pu. A cross-cultural user evaluation of product recommender interfaces. In *Proceedings of the 2008 ACM Conference on Recommender Systems*, RecSys '08, pages 75–82, New York, NY, USA, 2008. ACM.

[2] E. H. Chi. Blurring of the boundary between interactive search and recommendation. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pages 2–2. ACM, 2015.

[3] A. M. Degeratu, A. Rangaswamy, and J. Wu. Consumer choice behavior in online and traditional supermarkets: The effects of brand name, price, and other search attributes. *International Journal of research in Marketing*, 17(1):55–78, 2000.

[4] A. Dimoka, Y. Hong, and P. A. Pavlou. On product uncertainty in online markets: Theory and evidence. *Mis Quarterly*, 36, 2012.

[5] A. Felfernig, E. Teppan, and B. Gula. Knowledge-based recommender technologies for marketing and sales. *Int. J. Patt. Recogn. Artif. Intell.*, 21:333–355, 2007.

[6] H. Garcia-Molina. Thoughts on the future of recommender systems. In *Proceedings of the 8th ACM Conference on Recommender Systems*, pages 1–2. ACM, 2014.

[7] J. L. Herlocker, J. A. Konstan, and J. Riedl. Explaining collaborative filtering recommendations. In *ACM conference on Computer supported cooperative work*, pages 241–250, 2000.

[8] Y. Kim and R. Krishnan. On product-level uncertainty and online purchase behavior: An empirical analysis. *Management Science*, 61(10):2449–2467, 2015.

[9] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5):441–504, 2012.

[10] Y. Matsunami, M. Ueda, S. Nakajima, T. Hashikami, S. Iwasaki, J. O'Donovan, and B. Kang. A method for automatic scoring of various aspects of cosmetic item review texts based on evaluation expression dictionary. In *Proceedings of the 24th International MultiConference of Engineers and Computer Scientists*, IMECS '16, pages 392–397. IAENG, 2016.

[11] J. McAuley and A. Yang. Addressing Complex and Subjective Product-Related Queries with Customer Reviews. *ArXiv e-prints*, Dec. 2015.

[12] S. M. McNee, J. Riedl, and J. A. Konstan. Being accurate is not enough: How accuracy metrics have hurt recommender systems. In *Extended Abstracts of the 2006 ACM Conference on Human Factors in Computing Systems (CHI 2006)*, 2006.

[13] J. O'Donovan, B. Smyth, V. Evrim, and D. McLeod. Extracting and visualizing trust relationships from online auction feedback comments. In *IJCAI*, pages 2826–2831, 2007.

[14] J. O'Donovan, N. Tintarev, A. Felfernig, P. Brusilovsky, G. Semeraro, and P. Lops. Joint workshop on interfaces and human decision making for recommender systems (intrs). In H. Werthner, M. Zanker, J. Golbeck, and G. Semeraro, editors, *RecSys*, pages 347–348. ACM, 2015.

[15] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of ACM CSCW'94 Conference on Computer-Supported Cooperative Work*, pages 175–186, 1994.