Enhancing Recommendation Diversity with Organization Interfaces

Rong Hu
Human Computer Interaction Group
Swiss Federal Institute of Technology (EPFL)
CH-1015, Lausanne, Switzerland
rong.hu@epfl.ch

Pearl Pu
Human Computer Interaction Group
Swiss Federal Institute of Technology (EPFL)
CH-1015, Lausanne, Switzerland
pearl.pu@epfl.ch

ABSTRACT
Research increasingly indicates that accuracy cannot be the sole criteria in creating a satisfying recommender from the users' point of view. Other criteria, such as diversity, are emerging as important characteristics for consideration as well. In this paper, we try to address the problem of augmenting users' perception of recommendation diversity by applying an organization interface design method to the commonly used list interface. An in-depth user study was conducted to compare an organization interface with a standard list interface. Our results show that the organization interface indeed effectively increased users' perceived diversity of recommendations, especially perceived categorical diversity. Furthermore, 65% of users preferred the organization interface, versus 20% for the list interface. 70% of users thought the organization interface is better at helping them perceive recommendation diversity versus only 15% for the list interface.

Author Keywords
Diversity, recommender system, user study, interface.

ACM Classification Keywords
H.5.2 [Information Interfaces and Presentation]: User Interfaces - Evaluation / methodology.

General Terms
Design, Experimentation, Human Factors.

INTRODUCTION
Recommender systems have been popularly developed in recent years to help users overcome information overload in various information systems. In the literature, more attention is traditionally placed on the improvement of algorithm accuracy, with an underlying assumption that high recommendation accuracy will essentially lead to more user satisfaction. However, recent studies have shown that accuracy is not the only dimension to be taken into account when measuring the quality of recommender systems from the users' point of view [2, 4]. Other criteria, such as diversity, are emerging as important characteristics for consideration. The recommendations that users need should not only be accurate (close to their preferences), but also *useful* (helpful for their decision making processes) [2].

Many researchers have explored the issue of augmenting users' satisfaction via diversifying the items proposed in a recommended list [3, 4, 9]. For example, Ziegler et al. [9] introduced a diversification approach to decrease the topical similarity of a recommendation list. Their online user study demonstrated that going beyond accuracy, users' perceived list diversity could also influence users' overall satisfaction. In contrast to overspecialized recommendation lists, a balanced and diversified recommendation list broadens its coverage of users' spectrum of interests, so that it has the capability of helping users to discover novel items of interest or supporting them to find the ideal item quickly [3, 4]. However, one open area of study is whether a system's interface can change the way users perceive the diversity of recommended results by re-arranging the recommended items in a novel layout.

Chen and Pu [1] analyzed users' eye movements with one group of users viewing recommendations in a conventional list interface, versus another two groups viewing the same set of items in two organization interfaces respectively. Their findings reveal that users tend to focus on the top of the list interface. Items that are located farther down in the list are likely to be ignored. It seems that the number of retrieved results in a list interface is quickly beyond the extent of human cognitive processing capability. On the other hand, the organization interfaces could attract users' attention to more items and enhance the final decision quality. Furthermore, prior studies have demonstrated that organization interfaces have the capability of explanation [5]. It is, therefore, reasonable to ask the research question whether the organization interface design has the potential to enhance users' perception of diversity.

In this study, we conducted a within-subject user study to compare a conventional list interface and an organization interface in terms of users' perception of the list diversity

IUI 2011, February 13–16, 2011, Palo Alto, California, USA.

and users' satisfaction to the recommendations in a well-known commercial website. Our results reveal that the organization interface could effectively increase users' perceived diversity, especially in the context of items of diverse kinds (henceforth called categorical diversity). Most importantly, 65% of users preferred the organization interface, versus 20% for the list interface. 70% of users found the organization interface better at helping them perceive recommendation diversity. Furthermore, users using the organization interface are more likely to use the system again, tell their friends about it and buy the recommended items.

ORAGNIZATION-BASED INTERFACE

The idea of organization-based interfaces was first proposed as an explanation interface, with the aim of inspiring users' trust in recommender systems [5]. In an organization interface, recommendations are grouped into categories and each category is annotated with a title representing the common tradeoff properties among all items in that category compared to the current selected item (e.g., "these products are cheaper and lighter, but have slower processor speed"). In our experiment, we utilized the Editorial Picked Critiques (EPC) technique to generate categories for our organization-based interface.

Editorial Picked Critiques (EPC)

EPC was originally developed in the context of applying critiquing-based recommendation technology to public taste products such as music, films, perfumes, fashion goods or wine [7]. In contrast to high-involvement products such as PCs, digital cameras, users tend to spend less time choosing public taste goods and are more likely to rely on public opinions or experts' advice to make decisions [8]. EPC was designed to take into account the public opinions, popularity information and editorial suggestions, as well as the needs for personalization and diversity.

We adopted the compound categories proposed for perfume in [7] as our classification categories for the organization interface. These categories are "more popular and cheaper", or "more popular but more expensive" in the case that the former category does not contain any products, "same brand and cheaper" or "same brand but more expensive", "just as popular and cheaper", "same price range and just as popular", and finally "people who like this also like". We remapped the items in a recommendation list into these five categories, and used the bestselling order and customers' ratings as a popularity measure. The items which cannot be categorized into any of the first four categories are put into the category "people who like this also like".

USER STUDY

Materials

A well-known commercial website was used as our experimental platform due to its high reputation. Its standard list interface was used as the baseline. The organization version was achieved with the help of an open-source filtering HTTP proxy program, PAW (http://paw-project.sourceforge.net). The recommendation list we focused on was "Customer Who Viewed This Item Also Viewed" in the detail page for each product (perfume in our experiment). Unlike the organization-based interface designs in [1, 5, 7], the categories in this study were organized in a tab-based structure to better conform to the horizontal list style in that website. By clicking on each tab, users could see the recommendations in the corresponding category. The categories which had no products classified into were not presented. For both interfaces, the number of displayed recommendations was restricted to be the same (five in the current study) and the "next" and "previous" buttons were used to explore more items in a list. In order to avoid confusion, we removed the recommendation list of "Customers Who Bought This Also Bought" in the same page. In addition, we placed the section "Customers Who Viewed This Item Also Viewed" just beneath the selected product so that users could easily find it.

Dataset and Participants

The dataset of perfumes used was crawled from the well-known website and updated just before launching the study to ensure that we had a dataset containing the most recent and popular perfumes available on the market. In our experiment, 21,071 items were accessible, covering 13,246 items for women (6,281 Eau de Toilette, 689 Cologne and 6,276 Eau de Parfum) and 7,825 items for men (6,066 Eau de Toilette, 1,474 Cologne and 285 Eau de Parfum).

A total of 20 participants (10 females) were recruited. The incentive for the participants was a lottery: one out of the 20 users could win a 100 CHF gift voucher to purchase one of the perfumes the winner put in the basket during the study. These participants were from 8 different countries with various professions (student, research assistant, engineer, interface designer); their age mainly ranged from 20-30 (19 participants), and they represented various educational backgrounds (including bachelor, master and Ph.D.). In addition, all participants said that they were regular computer users and used the Internet frequently. 11 participants thought they have knowledge about perfume, and 8 participants had neutral opinions. 17 of the participants had used the tested commercial website before.

Evaluation Criteria

In order to evaluate users' perceived qualities of a recommender, we used a simplified version of a user-centric recommender evaluation model (ResQue) [6]. More specifically, two questions were designed to measure users' diversity perception of the recommendation lists. One referred to the difference among categories, querying whether "the items recommended to me are of various kinds" (called categorical diversity). The other considers the difference among each item, asking whether "the items recommended to me are similar to each other" (also called item-to-item diversity). We also tried to investigate the influence of perceived diversity on users' acceptance of a recommender system. In our evaluation, we took into

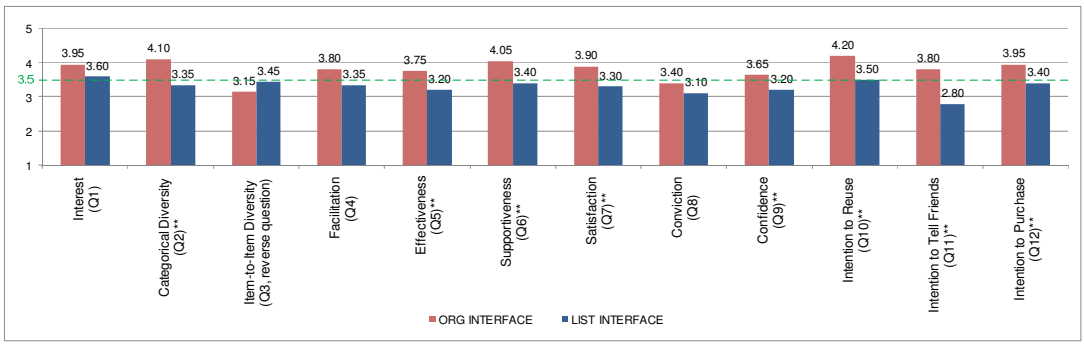

Figure 1. Usability and user satisfaction assessment results. A cut off value at 3.5 represents agreement on the Liker scale. ** is marked for significant difference at 5% (p-value < 0.05).

account perceived ease of use and usefulness of a system (facilitation, effectiveness, and supportiveness), users' attitudes towards the system (satisfaction, conviction, and confidence), and behavioral intentions to use it (intention to reuse, intention to tell friends, and intention to purchase). Besides, we measured users' perception on recommendation quality. Each question was required to respond on a 5-point Likert scale from "strongly disagree" (1) to "strongly agree" (5).

Experiment Design and Procedure

Our user study was conducted in a within-subjects design. All participants used both interfaces, and then filled in the assessment questionnaire for the respective interface. In the end, they were asked to answer 5 questions about their preferences on these two interfaces in terms of the following aspects: general preference, informative, useful, good at recommending, and good at helping perceived diversity. All participants were randomly assigned to two experimental conditions, with a differing order in using the two interfaces. That is, 10 users in one condition evaluated the list view interface first and then the organization view interface; the other condition had the reverse sequence. Counterbalance measures were taken to eliminate order and learning effects as much as possible.

Participants were given specific tasks when using each interface. In the first interface, we asked a user to find up to three perfumes that he/she has never heard of or used before and would be willing to purchase for himself/herself given the opportunity and put them into the shopping cart. When using the second interface, the user was asked to search for three perfumes which he/she would be willing to purchase for someone of the opposite gender as a gift, in order to reduce the potential influence of users' familiarity with the product domain after using the first interface.

RESULTS ANALYSIS

Users' Perception of Usability, Satisfaction and Behavioral Intentions

In this experiment, the recommendation contents were kept identical in both interfaces. Thus, the differences in users' perceptions of the systems only resulted from the interface design. All responses for each measured variable were analyzed using paired sample t-tests. The results are shown in Figure 1. For a 5-point Likert-scale question (from 1 to 5), a score greater than 3.5 is considered to be agreement. Overall, users indicated more favorable scores for the organization interface (ORG) over the list interface (LIST) with respect to all questions. All of their responses for ORG were above 3.5 except for Q3 (the reverse question, 2.85 (6 - 3.15)) and Q8 (3.4), whereas the scores for LIST were under 3.5 on all questions except for Q1 and Q10. Furthermore, the variable marked with (**) denotes that a significant difference was observed among users' responses (i.e., $p < 0.05$). The detailed analysis is as follows.

Users found the recommended items from both interfaces to be interesting (Q1) with a slight advantage for ORG ($p = 0.07$). To assess users' perceived diversity in both interfaces, we asked two questions. One emphasizes the categorical difference (Q2). The other simply considers the general differences between each item (Q3). Interestingly, we could see from the results that the difference between the two interfaces was only significant with respect to the question Q2. That is, the level of perceived categorical diversity in the organization interface was significantly higher than that of the list interface (mean = 4.1, SD = 0.788 for ORG, vs. mean = 3.35, SD = 0.988 for LIST, $p < 0.05$, $t = 3.68$). However, no significant difference was measured on item-to-item diversity ($p = 0.186$). Users seemed to disagree that items were similar to each other in both interfaces (reverse scale of item-to-item diversity). Therefore, we conclude that the organization-based interface helped users' awareness of the diversity present by variety differences.

Perceived ease of use and usefulness of the system were evaluated in terms of three aspects: facilitation (Q4), effectiveness (Q5), and supportiveness (Q6). While users found ORG is more easy to use (Q4), the difference between ORG and LIST was slightly significant ($p = 0.09$). On the other hand, users thought that the recommended items were significantly more effective in helping them find the ideal product (Q5) in ORG (mean = 3.75, SD = 0.851,

vs. mean = 3.2, SD = 1.005 for LIST, $p < 0.05$, $t = 2.773$). They also felt more supported in selecting the items to buy with the help of ORG (Q6, mean = 4.05, SD = 0.686, vs. mean = 3.4, SD = 1.095 for LIST, $p < 0.05$, $t = 2.371$).

In order to evaluate users' attitude towards the tested interfaces, three evaluation measures were considered: satisfaction (Q7), conviction (Q8), and confidence (Q9). Users expressed significantly higher satisfaction for ORG (Q7, mean = 3.9, SD = 0.912, vs. mean = 3.3, SD = 0.923 for LIST, $p < 0.05$, $t = 3.559$). In addition, they seemed to be more confident that they would like the recommended items in ORG (Q9, mean = 3.65, SD = 0.875, vs. mean = 3.2, SD = 0.894 for LIST, $p < 0.05$, $t = 2.269$)

Significant differences were also revealed on measures of users' behavioral intentions to use a system. More specifically, users scored significantly higher for ORG on reusing the system (Q10, mean = 4.2, SD = 0.834, vs. mean = 3.5, SD = 0.827 for LIST, $p < 0.001$, $t = 4.273$), telling friends about it (Q11, mean = 3.8, SD = 0.894, vs. mean = 2.8, SD = 0.894 for LIST, $p < 0.001$, $t = 4.359$) and purchasing the recommended items given the opportunity (Q12, mean = 3.95, SD = 0.686, vs. mean = 3.4, SD = 0.940 for LIST, $p < 0.05$, $t = 2.463$).

Final Preference

After evaluating two interfaces, users were asked to answer five questions regarding their preferences for these two interfaces. ORG got dominant preferences with more than 50% votes on all of the five questions. Particularly, 65% of users preferred ORG versus only 20% for LIST, while 5% of them prefer both interfaces. More users thought that ORG was more informative (70% vs. only 10%), more useful (60% vs. 15%) and better at recommending items (50% vs. 10%). More importantly, 70% (vs. 15%) of users thought that ORG is better at helping them perceive the diversity of recommendations in contrast to LIST.

CONCLUSION AND FUTURE WORK

In this paper, we conducted an in-depth user study to compare an organization-based interface with the standard list-based interface. Experimental results reveal that even though there is no significant difference between users' perception of item-to-item diversity in both interfaces, users more *strongly* perceived categorical diversity of the recommendation results in the organization view interface compared to the list view interface. There is a 22.4% increase. Most importantly, ORG users are more likely to use the system again, tell their friends about it and buy the recommended items.

The within-subject study also revealed interesting outcomes regarding users' direct preferences of the two interface formats. In contrast to the list interface, three times more participants preferred the organization interface (65% vs. 20% in favor of ORG and LIST interfaces respectively). Together with the significantly positive scores for the organization interface on measures of perceived usefulness of the system, attitudes towards the system and behavioral intentions to use it, this suggests that the organization-based interface is a more advantageous interface design for displaying a list of recommendations.

Our future work will be to explore the in-depth relations between users' perceived diversity and their satisfaction of a recommender system.

ACKNOWLEDGMENTS

We thank EPFL, the Swiss National Science foundation, and the ministry of education of the People's Republic of China for supporting the reported research work. We are grateful to the participants of our user studies for their patience and time.

REFERENCES

1. Chen, L. and Pu, P. Eye-Tracking Study of User Behavior in Recommender Interfaces. In*: P. De Bra, A. Kobsa, and D. Chin (Eds.): UMAP 2010*, LNCS 6075, pp. 375–380, 2010.
2. Herlocker, J., Konstan, J., Terveen, L., and Riedl, J. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22, 1 (2004), 5–53.
3. McGinty, L., and Smyth, B. On the role of diversity in conversational recommender systems. In *Workshop Proceedings of the Fifth International Conference on Case-Based Reasoning (ICCBR 2003)*, page 1065, 2003.
4. McNee, M., Sean, J. R., and Konstan, J. A. Accurate is not always good: How accuracy metrics have hurt recommender systems. In *extended abstracts on Human factors in computing systems (CHI'06)*, pages 1097-1101, 2006.
5. Pu, P., and Chen, L. Trust Building with Explanation Interfaces. In *Proceedings of the 11th International Conference on Intelligent User Interface (IUI'06)*, pages 93-100, 2006, Sydney, Australia.
6. Pu, P., and Chen, L. A User-Centric Evaluation Framework of Recommender Systems. In *Proceedings of UCERSTI Workshop of* RecSys'10, pages 14-21, Sept. 26-30, 2010, Barcelona, Spain.
7. Pu, P., Zhou, M., and Castagnos, S. Critiquing Recommenders for Public Taste Products. In *Proc. of the 3rd ACM Conference on Recommender Systems*, New-York City, NY, USA, October 2009.
8. Smith, D., Menon, S., and Sivakumar, K. Online Peer and Editorial Recommendations, Trust, and Choice in Virtual Markets. *Journal of Interactive Marketing 19(3)*, 15-37. 2005.
9. Ziegler, C., McNee, S. M., Konstan, J. A., and Lausen, G. Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web*, pages 22–32, 2005.