How Users Perceive and Appraise Personalized Recommendations

Nicolas Jones[1], Pearl Pu[1], and Li Chen[2]

[1] Human Computer Interaction Group, Swiss Federal Institute of Technology
(nicolas.jones, pearl.pu)@epfl.ch
[2] Department of Computer Science, Hong Kong Baptist University
lichen@comp.hkbu.edu.hk

Abstract. Traditional websites have long relied on users revealing their preferences explicitly through direct manipulation interfaces. However recent recommender systems have gone as far as using implicit feedback indicators to *understand* users' interests. More than a decade after the emergence of recommender systems, the question whether users prefer them compared to stating their preferences explicitly, largely remains a subject of study. Even though some studies were found on users' acceptance and perceptions of this technology, these were general marketing-oriented surveys. In this paper we report an in-depth user study comparing Amazon's implicit book recommender with a baseline model of explicit search and browse. We address not only the question "do people accept recommender systems" but also how or under what circumstances they do and more importantly, what can still be improved.

1 Introduction and Related Work

Twenty years ago, the classical buying-scheme was that when a user entered a shop, a knowledgeable seller would be available to advise and inform him/her on products. With the emergence of the Internet, online shops started to appear, proposing interfaces where the users had a high level of control, and where actions triggered predictable results. Classical interface have allowed people to express their preferences by browsing along a set of well defined categories. For Books these might be poems, romance or thriller. In addition search tools rapidly appeared allowing users to more quickly navigate to their target items. Later on, recommender systems (RS) were introduced, often relying on explicitly expressed ratings of items. More recently, there has been a lot of research on indirect ways for users to reveal their preferences (e.g. through their purchase history), paving the way to behavioral recommenders. This difference from search & browse to today's behavioral recommenders follows very well a more general and long standing debate, central to the UM community, about automation and direct manipulation which was voiced in [11]: to what extent should users give up control of their interaction with interfaces in favor of depending on intelligent "agents" that learn the likes and dislikes of a user?

In this paper we compare traditional user-controlled interfaces with more recent personalized systems using recommendations. A lot of research has been done on ways for users to reveal their preferences, and experiments such as [10] suggest that when users implicitly give feedback, the performance of the RS can be close to the more traditional ones using explicit feedback. But the work is highly incremental and there are no studies directly comparing both extremes. For these reasons, we decided to evaluate how recommendations based on *implicit preference feedback* compare with results provided to users who explicitly reveal their preferences in a traditional *user controlled* way. We chose to conduct this study on Amazon.com [3] and set up a comparative between-group user-study where users were instructed to search for five books. One group of users tested Amazon without the benefit of the RS, by searching and browsing. This represented the baseline measure for the experiment. Two other groups tested Amazon's recommendations which were based on their past purchase history. One group had a small purchase history whereas the second group had a larger profile. The experiment was conducted online and users' opinions were collected through a post-study assessment questionnaire, evaluating multiple dimensions from satisfaction to intention to return.

2 Background and Related Work

In content-based recommenders, users specify their needs explicitly in terms of content or features [8]. Similarly, in user involved RS, ratings are used to determine like-minded users through collaborative filtering. More recently, unit or compound critiquing techniques, rather than single valued ratings, were proposed to improve accuracy [2]. Such direct feedback is the most common interest indicator, offering a fairly precise way to measure users' preferences, but suffers from several drawbacks [3]. These include the fact that a user must stop to enter explicit ratings, which alters browsing and reading patterns. Users may not be very motivated to provide ratings unless this effort is perceived to be beneficial [9], or because the user might not yet know his preferences as he just started to use the system, and often changes them in different contexts [6, 8].

In behavior based RS, a user's purchase history or his reading time on a page can be used to infer interests and preferences. In Nichols' seminal paper on implicit rating and filtering [7], he identifies several types of data that can implicitly capture a user's interest, including past purchases, repeated uses, and decisive actions (printing, marking, examining). Since then, several of these indicators have been used like in [10] where Shapira et al. showed that mouse movements normalized by reading time were a good preference indicator, or as in [3] where the time spent on a page is shown as a potentially good indicator. Unfortunately, research work measuring the progress of RS, with few exceptions, has concentrated on improving the accuracy of algorithms, the most common metric being the mean average error (MAE) [5]. The earliest paper evaluating six RS in depth, with real users is [12] where the central concern was to compare the performance and acceptance of such systems against human recommenders (friends). A recent marketing survey [1] reported that consumers strongly preferred sites that provide personalized product recommendations, with 45% claiming that they are more likely to shop at sites with personalized recommendations than at sites without them.

Our work is the first significant in-depth user study that reports on the users' perceptions of today's behavioral recommender systems compared to classical search & browse patterns.

[3] We chose Amazon because it has a well-established RS; we have no affiliation with Amazon.

3 Hypotheses

We established three simple hypotheses. First, we expected that, when a user just starts using a website, a user-controlled solution would be more effective at supporting his information needs than an indirect one. If a user has a small purchase history, for example, there is perhaps not enough information to infer his preferences, most certainly resulting in an inadequate recommendation. We thus propose hypothesis H1. Second, we considered how recommendation quality might evolve. When a user controls a search, he may only cover a specific subset of all his preferences, whereas information gathered over time gives a much broader view of these preferences. We highlight this with hypothesis H2 where we fix an arbitrary cut-off level of twenty books. Finally, since an indirect profile should cover multiple aspects of a user's real profile, we hypothesized H3.

H1 for users with a small profile size, search & browse should provide higher recommendation accuracy than indirect feedback.
H2 there exists a profile size as of which indirect feedback should propose a better accuracy than the baseline explicit elicitation.
H3 non-expert users are likely, overall, to significantly benefit from recommendations based on indirect feedback.

4 Real-User Evaluation

The experiment was limited to the domain of *books*. We designed a between-group experiment of three user groups, with 20 users in each: the baseline search & browse group, and two recommendation-receiving groups with small and big purchase profiles respectively. All users were told to find five books to purchase, similar to what they would do on the real website.

4.1 Evaluation Setup & Procedure

We implemented a user study with a wizard-like online web application containing all the instructions, interfaces and questionnaires so that subjects could remotely participate in the in-depth evaluation. The general procedure consists of the following steps.

Step 1. Based on how many books a participant bought in the past on *Amazon* (profile size), he is oriented to the adequate experiment (baseline or recommendations).

Step 2. Basic background information is collected (gender, age, etc.)

Step 3. After reading a brief scenario, the user is given detailed instructions: The tester of the *search & browse* interface is instructed to go to *Amazon.com*, make sure he is not logged in, and then to browse through the available categories of literature, until he finds a book which he likes. The tester of the *implicit RS* system is asked to head to *Amazon.com* and log in to his account. He is then asked to go the "my recommendations" section and to navigate through the book section of the recommendations until he finds a book that he likes.

Step 4. The user starts the experiment. He is asked to select *five* books; for each one, he must fill in a template-questionnaire allowing him to rate the book on the spot.

Step 5. To conclude the study, the user is asked to complete a nine questions assessment questionnaire to evaluate the system he has just tested.

4.2 Measured Variables

All questions in this study are statements to which a user can indicate his level of agreement on a five-point Likert scale, ranging from -2 (strongly disagree) to $+2$ (strongly agree); 0 is neutral. Not having access to Amazon's interaction logs, we recorded users' opinion about the recommendation quality through a template, immediately after selecting each book (novelty, appreciation, intention to buy). Then, once five books had been selected, an overall appreciation was recorded through a set of nine questions, measuring *experience* (satisfaction, effort, trust, confidence, novelty, diversity) and *decision* (acceptance of a recommended book, future usage, sharing with friends). Because of the setup of the experiment, each question was adapted into two variants such as to differentiate between the baseline and recommendation experiments, but tested identical dimensions.

4.3 Participants

The user study was carried out over a period of three weeks and an incentive was proposed. The study was taken by off-campus users (half of the participants), students (7%) and academic researchers in Switzerland. The study collected 60 users, resulting in a sample size of twenty participants per group. There were 17 female and 43 male, with 66% being aged between 25 and 30; 18% were younger, and 15% older. The group of *baseline* users showed slightly less familiarity with Amazon as 25% more users disagreed that they "read a lot of books", and 30% of them had never surfed on Amazon before. We accepted this potential bias as such users have a fresh view of Amazon, less influenced by the evolution of the site.

5 Evaluation Results

Results are reported in Figure 1. An Anova analysis showed that five questions conveyed statistically significant different averages across all three groups of users. The question S2 shows an increase in results from *baseline* elicitation to *recommendation* users with a large profile, who found that the system required less effort (with an average of 1). The *recommendation* users with a small profile scored 0.6 on average. The difference between all three groups is significant ($p = 0.02$). S5, the question on trust, shows the same general tendency, albeit a smaller increase between the first two groups (significant, $p = 0.05$). S3, the confidence about making the best choice, presents a *baseline* average around 0.5 and one of -0.5 for the *recommendation* small group, with the *recommendation* big being amid (significant, p = 0.02). Diversity S4 shows a very similar pattern, but with an increased score from the *baseline* users, around 1 (significant, $p < 0.01$). One of the template questions also shows a significant difference: T3, the intention to buy where the *recommendation* small is much lower than both other groups ($p = 0.04$).

For S1, satisfaction, the 0.5 difference between the first two groups is significant ($p = 0.02$). T2, on perceived accuracy, gives much higher averages around 1.0, with a significant difference (t-test, $p = 0.02$) between the two *recommendation* groups. Finally, the special question for *recommendation* users about them having "already used"

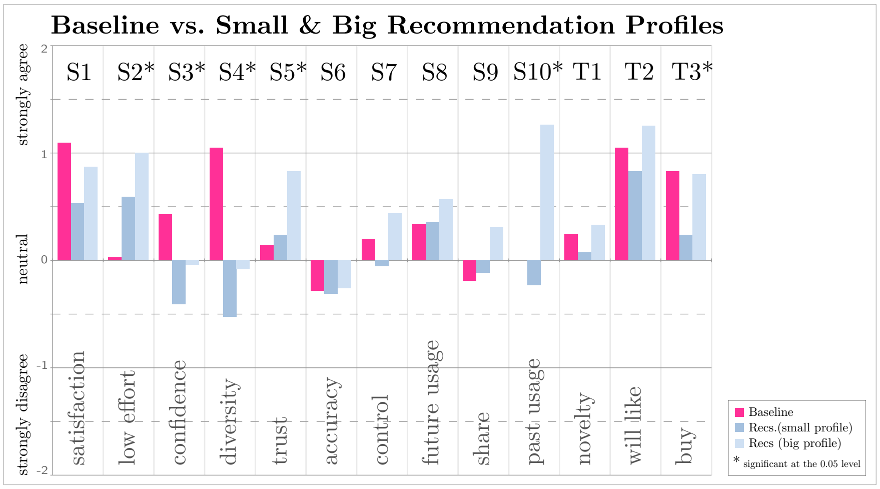

Fig. 1. Detailed Graph of Users' Preferences

this recommendation feature showed S10 the expected trend with a score close to 0 for the *recommendation* small group, and above 1 for the *recommendation* big group (significant, $p < 0.01$). These results reveal that although a recommender interface provides users with an overall satisfaction and perceived benefits like a lower effort required, most users have to wait until their profile reaches a certain size to enjoy the full benefits.

6 Discussion and Conclusions

Through our hypotheses H1 and H2, we predicted that at first a controlled search would be more accurate but that this would rapidly change, seeing the accuracy of recommendations increase with the profile size. The direct assessments of perceived accuracy, S6 and T2, are not strongly conclusive. This twist-and-turn between hypotheses and results is surprising. However, we would like to point out that if "accuracy" does *not* reveal itself as imagined, other dimensions *do* demonstrate some parallels with the predictions. Elements like confidence and diversity, show us that search & browse methods are more efficient at the beginning, but that larger recommendation profiles actually start to catch up. Nevertheless, and this brings us to H3, there are not many measures where an implicit large profile strongly beats an explicit one (only trust S5 and low effort S2).

The results point out that the two types of interface mechanisms being compared can provide quite similar overall satisfaction for the users. The difference in the amount of effort required to operate in both systems is highly noticed by users, and clearly in favor of the RS (which required lower effort). On the other hand, users clearly found the *baseline* as proposing a much more diverse set of books, which is problematic for the recommender engine. It is also disappointing to see such low scores for the novelty

(T1) from the recommender. Measures of confidence show that users are more confident about their choices in the search & browse scheme. However, people are trusting the system's implicitly generated recommendations, as soon as their profile reaches a certain size, which is encouraging. This was further reflected in users' comments. When compared to books that friends might have proposed, neither methods were perceived as being very accurate; nevertheless users' opinions were positive as in all groups they thought they would like the five selected books. Contrary to purchase intentions, decision variables about future usage of the system or introduction to a friend, were not very high on average, but all three showed good correlations with satisfaction.

A decade has passed since the recommender technology was invented [4]. Today's systems based on this technology are in the mainstream practice of e-commerce and social websites. Even though some surveys demonstrate that acceptance and perception of this technology are showing good sings, we should not take them for granted. Our paper demonstrates that investigating users issues pays off, and that several traditional problems remain unsolved. It gives a clear idea how to improve the current technology and points out design guidelines. Additionally, the challenge of motivating initial users until they build a large profile (hence user loyalty) remains.

References

1. Choicestream personalization survey, 2007, http://www.choicestream.com/.
2. CHEN, L., AND PU, P. Evaluating critiquing-based recommender agents. In *AAAI* (2006).
3. CLAYPOOL, M., CLAYPOOL, M., BROWN, D., BROWN, D., LE, P., LE, P., WASEDA, M., AND WASEDA, M. Inferring user interest. *IEEE Internet Computing 5* (2001), 32–39.
4. GOLDBERG, D., NICHOLS, D., OKI, B. M., AND TERRY, D. Using collaborative filtering to weave an information tapestry. *Communications of the ACM 35* (1992), 61–70.
5. HERLOCKER, J. L., KONSTAN, J. A., TERVEEN, L. G., AND RIEDL, J. T. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst. 22*, 1 (2004), 5–53.
6. KEENEY, R. L., AND RAIFFA, H. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Wiley series in probability and mathematical statistics. John Wiley & Sons, Inc, New York, 1976.
7. NICHOLS, D. M. Implicit rating and filtering. In *In Proceedings of the Fifth DELOS Workshop on Filtering and Collaborative Filtering* (1997), pp. 31–36.
8. PU, P., AND CHEN, L. User-involved preference elicitation for product search and recommender systems. *AI Magazine 29(4)* (2008), pp. 93–103.
9. RESNICK, P., IACOVOU, N., SUCHAK, M., BERGSTORM, P., AND RIEDL, J. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In *Proc. of ACM CSCW'04*, ACM, pp. 175–186.
10. SHAPIRA, B., TAIEB-MAIMON, M., AND MOSKOWITZ, A. Study of the usefulness of known and new implicit indicators and their optimal combination for accurate inference of users interests. In *SAC '06* (New York, NY, USA, 2006), ACM.
11. SHNEIDERMAN, B., AND MAES, P. Direct manipulation vs. interface agents. *interactions 4*, 6 (1997).
12. SWEARINGEN, K., AND SINHA, R. Beyond algorithms: An hci perspective on recommender systems, 2001.