Case Amazon: Ratings and Reviews as Part of Recommendations

Juha Leino and Kari-Jouko Räihä
University of Tampere, Department of Computer Sciences
Tampere Unit for Computer-Human Interaction
FIN-33014 University of Tampere, Finland
{juha.leino, kari-jouko.raiha}@uta.fi

ABSTRACT

We studied user behavior in a recommender-rich environment, Amazon online store, to see what role the algorithm-based and user-generated recommendations play in finding items of interest. We used applied ethnography, on-location interviewing and observation, to get an accurate picture of user activity. We were especially interested in the role of customer ratings and reviews and what kind of strategies users had developed for such an environment. Our results underline the need to develop recommender systems as a whole. The way the recommendations are shown affects which items get picked, and for improving the interface, it is necessary to study the whole in addition to studying the parts in isolation.

Categories and Subject Descriptors

H.5.2 **[User Interfaces]**

General Terms

Design, Experimentation, Human Factors.

Keywords

Recommender systems, Amazon, rating systems, reviews.

1. INTRODUCTION

"*I'll just peak at the customer reviews quickly. You know, to see what they've said about it. It's not that it's so expensive but it does take time to read it. That's more expensive*." - Participant 4

The growth of e-commerce has witnessed proliferation of various types of recommender systems. Recommender systems, if considered to simply be algorithms that produce hits based on collaborative filtering and other methods, do not exist in a vacuum. The purpose of any recommender system is to direct the users to the items that best satisfy them. As no recommender system can be 100% correct or produce systematically novel and serendipitous results [3], the presentation of the results has to be considered an integral part of the system and so the algorithm cannot be seen as be-all-end-all. Consequently, we need to look at the recommenders as a part of the whole, and consider them in the actual use context.

Moreover, not all recommender systems are based on algorithms. For instance, Amazon's customer reviews and ratings constitute a type of recommendation that is based on human input but that does not go through collaborative filtering before it is presented to the users.

The problem of finding the desired item in an Internet store is compounded if the item cannot be shown in electronic form. Users can be allowed to listen to an mp3 song, but they cannot be given a book to see in its physical form. Under such circumstances, users must make up their minds based on the information given to them in the list and item pages.

Few recommender system studies have focused on the effects of the interface on the use of recommendations, choosing instead to focus on the algorithms. While Cosley et al. [1] studied the effects of the interface on giving opinions for recommender systems, we studied its effects on selecting an item.

We set out to understand user strategies for selecting an item to buy under the conditions where the item cannot be shown as it is in an online store. We were especially interested in how the users selected items from list pages for a closer look and how they chose which customer reviews to read.

We picked books as item and Amazon, the world's biggest online retailer, for our study as it is the archetypical Internet store that has consistently been an early adopter and developer of new e-commerce approaches [5]. In particular, Amazon has used a wide array of recommender approaches, including customer reviews, for years.

We chose applied ethnography, in this case a combination of observation with verbal protocol and interviewing at the participants' homes using their own computers, as our method to get a true view of the real use and to avoid the say-do problem. People have a human tendency to describe what they do differently from what they actually do. [4]

Our method limited us to having only six participants, and so we did not have the necessary mass for studying alternative approaches as the subgroups would have been too small for reliable conclusions. Furthermore, the participants used Amazon only for buying nonfiction books, and some even stated that for lighter reading their behavior might be different. We were, however, able to see some trends and examples of user behavior in a recommender-rich online store.

RecSys'07, October 19–20, 2007, Minneapolis, Minnesota, USA.
Copyright 2007 ACM 978-1-59593-730-8/07/0010…$5.00.

We found that customer reviews and ratings are an important part of the selection process and thus part of the overall recommender system. Amazon's star ratings were commonly used in selecting items for closer scrutiny in the list pages. The effect was especially striking as all six participants indicated both in interviews and observations that not all the reviews from which the ratings were drawn were relevant to them.

Our study further shows how users approach the item page in real life and how it affects their ability to make the decision to purchase or move on. Consequently, we see the listing pages and item pages as integral parts of the recommendation system that either enhance or reduce its effectiveness in allowing the users find the desired items.

After introducing our method in detail, we discuss our results. We look at the role of the recommender systems in finding items, the role of star ratings in selecting items for a closer look, and the role of reviews in the item page.

2. METHOD

We used a combination of interviewing and observation with verbal protocol. The participants were given four tasks and asked questions before, during, and after each task. Care was taken not to direct participants' attention with the questions. After the tasks, a semi-structured interview was conducted.

The tasks were given to the participants in a web site made for the study. Each task was on its own page. Task 1 was to find and buy a book the participant had not chosen beforehand from Amazon.com or .co.uk. The choice of site was given to preserve normal conditions although there are differences between the two interfaces. Three used .co.uk, two used .com, and one used both. On average, 23 minutes were used on Task 1. Each participant was given 15 € towards purchasing the book so that they would select a book they really wanted.

Task 2 was to choose a digital photography guide from a list of seven books. A list page, constructed to look like a list page from Amazon.co.uk, was made to include books with high rating, low rating, no rating and "Search inside" function. A mock-up page was used to make sure that all the different conditions were present. Links led to actual item pages in the .co.uk site. On average, 11 minutes were used on Task 2. The analysis of the other tasks is beyond the scope of this paper.

The participants were six Finnish males, aged between 33 and 44. We refer to Participant 1 as P1, Participant 2 as P2, etc. A book purchase from Amazon was a requirement for recruitment to make sure that all were actual users of Amazon. On average, participants had purchased 10 books (the smallest number purchased being 2 and the largest 30) from Amazon prior to the study. The common reason for using Amazon was the availability of books. Four participants had also bought other items from Amazon. Participants had used Amazon for 4.5 years on average. All were in working life, had at least polytechnic-level education, and were experienced Internet users.

The interview-observation sessions, one per participant, were held where participants typically visited Amazon, using their own computer. In all cases, this turned out to be their home. The sessions lasted 2-3 hours. Because no problems emerged in the pilot study, it was also used in the analysis.

The sessions were videotaped with the camera directed at the computer screen. The camera also recorded the talking. All sessions were transcribed and then contrasted for analyzing. No analysis software was used.

3. RESULTS AND DISCUSSION

3.1 Strategy for finding items

Any recommender system needs some input to generate personalized recommendations that are not simply based on item popularity. Amazon uses two types of input: first, the user's long term engagement with the site and second, the user's current activities. Returning clients are recognized by cookies.

Task 1 was used to study the book-finding strategies. P3 used personalized recommendations. P2 also started with recommendations after signing in but he already owned the only interesting book in them. He continued with keyword search. Three participants used keyword searches directly while the last one started with categories but moved to keyword searching after failing to locate a book to buy.

P1 found a book with keyword search but after putting it into the shopping basket, he saw a recommendation for another book and went to its item page. When seeing an offer to buy the book in the basket with the new book ("Perfect Partner" recommendation) for £40 (slight discount), he decided to buy both even though he had earlier on mentioned wanting to get a book on the subject for about £10. P5 was also interested in the "Perfect Partner" recommendation, but did not buy the second book because he already had it. P4 found the book to buy from "Customers who bought this item also bought" list after a few searches. Thus, of the seven books bought in Task 1, three were found by recommendations and four by keywords.

All participants used one or more recommender feature (Table 1). Interestingly, no feature was clearly more popular than others. All were used in different phases of the item finding process.

Table 1. Recommendation usage in Task 1 by participants.

Task 1	P1	P2	P3	P4	P5	P6
Bought a book offered by algorithmic recommender	X		X	X		
Used keyword search	X	X		X	X	X
Used categories for searching						X
Used "Perfect Partner"	X			X	X	
Used personalized recommendations		X	X			
Used "Customers who bought/viewed this item..."	X			X		X

When asked how normal the setting in Task 1 was, P1 said that in many cases, he already knew quite well which books he was going to buy, but that when he needed a book for a particular topic, he did use Amazon in this way. Both P2 and P3 said that they typically started with personalized recommendations. P3 considered the emailed recommendations a great way to stay up-to-date. P5 normally used his favorite authors as a starting point, as did P2 in addition to recommendations. P2, P4, and P6 said that they typically already knew something about the book they came to buy, which made this way of searching somewhat different from normal. However, all said that after getting on with the task they ended up using Amazon pretty much the way they normally did.

Overall, our study shows that recommender systems can play an important part in item-finding when seeing all items individually is impossible. While keyword searches are the standard way, different recommender systems can complement the searches and potentially replace them as in the case of P3 and partially P2.

3.2 List pages and ratings

"*I must've done it automatically because of looking at those stars there. This one's gotten five stars while the other one's gotten three and a half. But when you go to the page, you see that there are only two reviews. So in that sense it's humbug but that's the reason why I went there.*" – P4

In Amazon, item list pages have recommendations in the form of average number of stars given by customers in the reviews. The list page only showed the stars but not on how many reviews they were based. While all participants felt that they must first read a review to decide if it is relevant for them, the stars on the list page had an impact on which books three participants chose to view at item page level. The other three users claimed not to pay any attention to the stars before the item page. Observation supports this as these participants only mentioned stars when on an item page.

Surprisingly, no participant knew how the books were sorted in the list page. This gave rise to misconceptions. For instance, P1 went to the item page to see how many reviews had been written to see if the book had sold well even though the books were sorted by "Bestselling": "*I checked how many people had commented the book, that is, how well it has sold. If there are no reviews, it probably hasn't sold much.*"

Amazon does show the sorting principle and allows the users to change it with a drop-down list but only P3 had used it. P6 even said that he wished to have a way to organize the items in the list page. In practice, both observation and interviews showed that the participants dealt with a large numbers of hits, most of them not very useful, by making more specific keyword searches.

In addition, the fact that P2 and P4 complained about not being able to see the number of reviews on the list page underlines the need to show the bases of recommendations. Five stars might appear attractive in the list page but if the star rating was based on one review, the participants gave it little weight unless its content was especially helpful. On the other hand, a book with four stars drawn from fifty reviewers would definitely be of interest to the participants. P4: "*I don't think that one Joe Blow's say-so matters. You could say that the strength of Joe Blow reviews lies in the mass. And that's it. If 200 Joe Blows give four stars on average, then that's kind of reliable.*" It appears that the better the users understand how the recommendations are made, the better they can process the recommendations and make logical choices.

In general, it might turn out that users could be better able to choose, say, a movie from the recommended ones if they knew why a certain movie was included on the list. Is it there because of the main actor or director they have liked in the past or because other people who have liked the movies by a certain director have also liked movies by this director? The knowledge of the reasoning might enable users to look at the recommendation from the right perspective and thus improve their chances of finding the right item. Findings by Herlocker, Konstan, and Riedl [2] certainly support this conclusion.

3.3 Item pages and customer reviews

"*You can't take it at face value. ... And that's why I somewhere earlier, when someone had given two stars, I didn't really care because I immediately saw that he didn't know what he was talking about. ... So I knew the guy was stupid and so his two stars were irrelevant. That's what's important; you have to figure out who's this guy who's reviewing.*" – P5

Customer reviews in the item page represented the most direct customer-to-customer recommendation in Amazon that was relevant to the participants. Interestingly, others, such as Listmania! lists, tags, and Search Suggestions, were not used at all. Part of the reason is probably that because the participants did not visit Amazon that often but had used it for a long time, on average four and a half years, they were used to the features that have been available for some time.

All participants said that the customer reviews played a role in deciding whether or not to buy a book but that the reviews were typically not the only factor. The customer reviews tended to have a more pronounced preventive than encouraging effect (Table 2).

Table 2. Importance of customer reviews. (1: Has only bought two books from Amazon. 2: With other contributing factors.)

Importance of customer reviews	P1	P2	P3	P4[1]	P5	P6
Do customer reviews play a part in your buying decisions?	Y	Y	Y	Y	Y	Y
Have you decided not to buy a book due to negative customer reviews?	Y	Y	Y		Y	Y[2]
Have you bought a book based on positive reviews?		Y	Y		Y	

Reviews served the participants in two ways. First, they were seen as sources of information about the contents of the book. For this, the stars and the positive or negative recommendation inherent to the review were irrelevant. The participants wanted to know what the book contained, and customer reviews were seen as a source for this together with "Search inside" and "Synopses".

Second, the participants reflected the reviewers' needs and expertise against their own. A negative review was typically not seen as a deterrent if the reviewer's needs or level of expertise in the field were different from the participant's. Likewise, a positive review did not count for much if the participant did not see the relevance of what the reviewer said in relation to himself.

While four participants claimed to typically read about five reviews when available, observations did not support this. The common start-off strategy seemed to be to pick one to three reviews for more careful reading after a general glance. No participant went to the second page of reviews. This is in contrast to the list page, which could be scanned more quickly, and consequently it was not rare for the participants to view 2–3 pages. Interestingly, the participants did not know in which order the reviews were shown. Three guessed it but admitted not knowing for sure.

The reviews that got read carefully were typically longer—all participants said that the shorter ones cannot give proper reasoning for their conclusions—and were written in a matter-of-fact style. In fact, two participants expressed clear dislike for emotional tone. Bad English and improper tone were also common turn-offs. However, as it appears that the short ones get glanced at, they do affect at least the overall impression.

While the name the reviewer goes by was important to five participants, none of them had noticed Amazon's "Real Name™" badge for verifying the identity of the reviewer. Furthermore, only two had tried to look at other reviews written by the reviewers to get further information on them. P2 tends to seek for reviews by reviewers he knows personally or whose books he has read, but if he does not find any, he tends to read one or two reviews from the top. No participant had ever learned to recognize and trust any reviewers from reading the reviews and other means Amazon provides for getting to know more about the reviewers.

P1, P5 and P6 compare the stars of the reviews to see if the reviewers have reached consensus. P3, P5, and P6 tend to check out the extreme views identified by high or low stars and do not find middle-of-the-road reviews useful. P4 considers negative reviews the most useful because they point out the potential problems. For P1 and P2, the positiveness or negativeness of the reviews does not really affect their choice.

In Amazon, users can click Yes/No buttons to indicate if a review was helpful for them. The number of those who found the review useful is given above the review. However, P2 and P3 had never even noticed the feature, and the four others did not give it much consideration in selecting the reviews for reading.

It appeared that the participants were looking for certain keywords from the reviews when selecting ones for careful reading. For instance, in Task 2, participants looked for words like "beginner" and "professional" from the reviews. Some wanted an elementary book, others a more advanced book, and so these words indicated for whom the book was written or the level of expertise of the reviewer.

The process of selecting reviews for reading appeared only partially conscious, as P2 said: "*It comes from somewhere deep from some random number generator, that's where it pretty often comes from*." The conscious factors were there, but the data from this study only partially illuminated the whole process and further research on this is necessary.

The number of book reviews in Amazon can run in tens and even hundreds and span several pages. As this study indicates, the users can have strategies for selecting which reviews to read, and so the service should provide tools for using these strategies effectively. Such tools could also help others to formulate strategies. Length, star rating, and use of reviewer's real name appear to be among the factors that could be used for filtering or sorting the reviews. In addition, if the reviewer expertise level in the field was set by the reviewer for the review, it could be useful for the readers. Amazon.com has recently started to give the users tools for finding relevant reviews, but Amazon.co.uk has not. Our study underlines the need for such tools.

4. CONCLUSION

Our study stemmed from the wish to understand what kind of user strategies emerge in a complex online shopping environment, such as Amazon. Our data shows how recommender systems are actually used and affords us a glimpse at online shopping reality.

By observing and interviewing six Finnish Internet shoppers using Amazon, we found that while keyword search was still the most common approach to finding products from a large number of possibilities, recommender systems played an important part in helping users find the books they wanted. Algorithmic recommender features helped the participants find three books out of seven while the more direct recommendations helped the users decide which items to view more closely and which items to buy.

One interesting area for further research is how to give users an understanding of the internal workings of the recommender system to allow them to process the result sets more effectively. What information do users need and how can we help them to develop correct mental models even if they only use the services occasionally? Everything in the interface communicates, and we need to make sure that it leads to the right understanding of the system. We are also interested to further study the strategies used in the selection process: which items get picked from listings, be it review lists or item lists, for closer scrutiny.

All in all, while we certainly need to develop recommender algorithms, we also need to keep in mind that any algorithm-based recommender system needs to communicate its results to the users in a meaningful manner, and so we need to study recommenders as parts of the integral whole as well. Studying the whole tells us how to improve its parts. Our study is a step in that direction.

5. ACKNOWLEDGMENT

This work was supported by the Finnish Funding Agency for Technology and Innovation (project 40279/05).

6. REFERENCES

[1] Cosley, D., Lam, S. K., Albert, I., Konstan, J. A., Riedl, J., Is Seeing Believing? How Recommender Interfaces Affect Users' Opinions. In *Proc. of CHI 2003*, ACM Press, 2003, 585-592.

[2] Herlocker, J. L., Konstan, J. A., and Riedl, J., Explaining Collaborative Filtering Recommendations. In *Proc. of CSCW'00*, ACM Press, 2000, 241-250.

[3] Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T., Evaluating Collaborative Filtering Recommender Systems. *ACM Trans. on Information Systems*, *22*, 1, 2004, 5-53.

[4] Jordan, B., and Dalal, B., Persuasive Encounters: Ethnography in the Corporation. *Field Methods*, *18*, 4, 2006, 1-24.

[5] Kotha, S., Competing on the Internet: The Case of Amazon.com. *European Management Journal*, *16*, 2, 1998, 212-222.