Comparing Recommendations Made by Online Systems and Friends

Rashmi Sinha and Kirsten Swearingen
SIMS, University of California
Berkeley, CA 94720
{sinha, kirstens}@sims.berkeley.edu

Abstract: The quality of recommendations and usability of six online recommender systems (RS) was examined. Three book RS (Amazon.com, RatingZone & Sleeper) and three movie RS (Amazon.com, MovieCritic, Reel.com) were evaluated. Quality of recommendations was explored by comparing recommendations made by RS to recommendations made by the user's friends. Results showed that the user's friends consistently provided better recommendations than RS. However, users did find items recommended by online RS useful: recommended items were often "new" and "unexpected", while the items recommended by friends mostly served as reminders of previously identified interests. Usability evaluation of the RS showed that users did not mind providing more input to the system in order to get better recommendations. Also users trusted a system more if it recommended items that they had previously liked.

A common way for people to decide what books to read is to ask friends and acquaintances for recommendations. The logic behind this time-tested method is that one shares tastes in books, movies, music etc., with one's friends. As such, items that appeal to them (friends) might appeal to me. Online Recommender Systems (RS) attempt to create a technological proxy for this social filtering process. The assumption behind many RS is that a good way to personalize recommendations for a user is to identify people with similar interests and recommend items that have interested these like-minded people (Resnick & Varian (1997), Goldberg, Nichols, Oki & Terry (1992)). This premise forms the statistical basis of most collaborative filtering algorithms. Since the goal of most RS is to replace (or at least augment) what is essentially a social process, we decided to directly compare the two ways of receiving recommendations (friends & online RS). Do users like receiving recommendations from an online system? How do the recommendations provided by online systems differ from those provided by the users' friends? Our hypothesis was that friends would make superior recommendations since they know the user well, and have intimate knowledge of his / her tastes in a number of domains. In contrast, RS only have limited, domain-specific knowledge about the users. Also, information retrieval systems do not yet match the sophistication of human judgment processes.

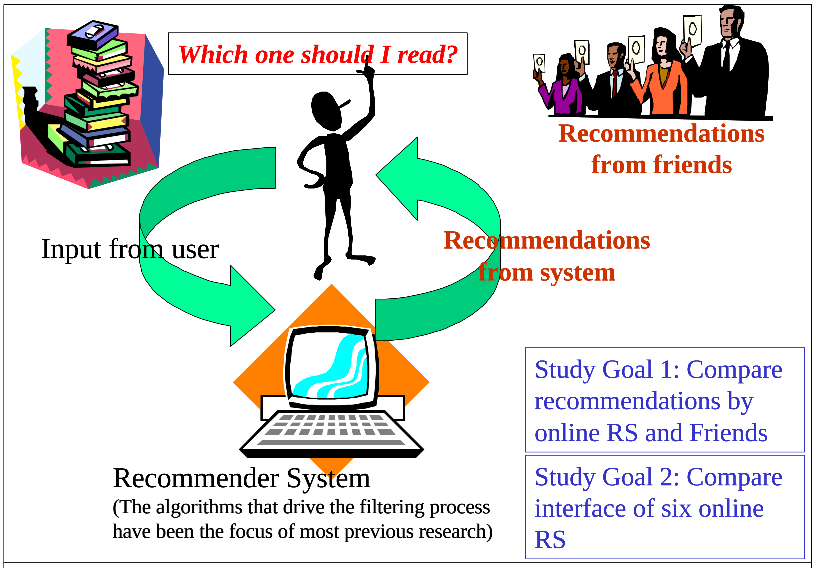
Which one should I read?
Recommendations from friends
Input from user
Recommendations from system
Study Goal 1: Compare recommendations by online RS and Friends
Study Goal 2: Compare interface of six online RS
Recommender System
(The algorithms that drive the filtering process have been the focus of most previous research)

Figure 1: Study Design

STUDY DESIGN

We conducted an empirical study to (a) compare recommendations made by users' friends to those made by online RS, and (b) to evaluate the interface of online RS. (See Figure 1 for Study Design.) We chose a variety of online RS based on differences in interfaces, number of ratings required, and results display (number of items returned, amount of description). RS can typically take explicit or implicit input or a combination of the two (Schafer et al. 1999). In this study, we only examined systems that relied upon explicit input. Systems studied were three book RS

(Amazon's Recommendation Wizard, Sleeper and RatingZone's Quick Picks) and three movie RS (Amazon's Recommendation Wizard, Reel.com's Movie Matches, and MovieCritic).

Independent Variables: (a) Source of Recommendations: Friend or online RS (b) Item Domain: Books or Movies (c) System itself.

Dependent Measures: Some of our measures were designed to evaluate the quality of the recommendations, while other measures focused on interface issues. The dependent measures are described below.

(a) Quality of Recommendations: To evaluate the recommendations provided by online RS and by friends, we computed three metrics: **Good Recommendations**: This was a measure of the system's ability to provide recommendations that interest the user. From the perspective of someone designing a RS, it is important to score highly on this metric. This metric can be broken down further into two categories: (i) **Useful Recommendations**: These are recommendations that the user is interested in, and has not experienced before. This is the sum total of useful information a user gets from the system**--**ideas for books to read / movies to watch in the future. (ii) **Trust-Generating Recommendations**: These are recommendations that the user has had positive experiences with previously. These are not useful in the traditional sense, but they index the degree of confidence a user can feel in the system. If a system recommends a lot of "old" items that the user has liked previously, chances are, the user will also like "new" recommended items.

(b) Overall Satisfaction with recommendations and with online RS: We asked users to rate their overall satisfaction with the recommendations.

(c) Time Measures: We measured time spent registering and receiving recommendations from the system.

(d) Interface Issues: Some of the interface issues we examined were specific to RS: (a) amount of input required from user, (b) Presentation of recommendations to the user (i.e., what information was presented for each item). Other interface issues were more general: screen layout, graphics, navigation, use of color and instructions.

METHOD

Participants: A total of 19 people participated in the experiment. Study participants were mostly students at the University of California, Berkeley. (*Participant Details*: Age range: 20 to 35 years; Gender Ratio: 6 males and 13 females; Technical Background: 9 worked in or were students in technology-related fields; the other 10 were studying or working in non technical fields). Participants were given the choice to explore books or movies. Each participant tested either three book or three movie systems, as well as evaluated recommendations made by three friends.

Procedure: For each of the three book / movie recommendation systems (presented in random order), participants completed the following tasks: (a) Completed online registration process (if any) using a false e-mail address, so that any existing buying/browsing history would not color the recommendations provided during the experiment. (b) Rated items on each RS in order to get recommendations. (Some systems required users to complete a second step, where they were asked for more ratings to refine recommendations.) (c) Reviewed list of recommendations. (d) If the initial set of recommendations did not provide anything that was both new and interesting, participants were asked to look at additional items. They were to stop looking when they found at least one book/movie they would be willing to try, or they grew tired of searching. (e) Completed satisfaction and usability questionnaire for each RS.

The second part of the experiment involved the human recommenders. Participants gave us e-mail addresses for three friends familiar enough with their tastes to be able to recommend 3 books or movies. The only constraint was that the friends could not name a book or movie that the user had discussed with them. (We included this constraint because we did not want friends to recall items they knew the user liked, and simply recommend those.) For each item recommended by a friend, users reviewed a plot synopsis and a cover image. They evaluated the friends' recommendations on the same dimensions as recommendations made by online RS.

RESULTS

Comparing Online Recommender Systems

Number of items in initial set: The first metric by which we compared the RS was the number of items that the system initially recommended to the user (See Table 1). Because each friend recommended exactly three items for the target user, the comparison between RS and friends was not interesting, and we focused on the differences between the various RS. The average number of items recommended by the various systems ranged between 7 & 20.

		Book Recommender Systems			Movie Recommender Systems		
		Amazon (B)	**Sleeper**	**RatingZone Quick Picks**	**Amazon (M)**	**Reel**	**MovieCritic**
How many ratings did users have to provide?		1 favorite item in 4 categories, 16 more items in refinement step	15 items to rate (minimum)	50 items to review, all optional to rate	1 favorite item in 4 categories, 16 more items in refinement step	1 item	12 items to rate (minimum)
User ratings:	Not enough	20%	20%	20%	56%	44%	
	Just right	70%	50%	40%	44%	22%	56%
	Too many		20%	30%			44%
How many results were users given?		15	1 at a time (10 in all)	8	15	5 to 10	20
User ratings:	Not enough	40%	10%	70%		11%	
	Just right	50%	50%	30%	89%	78%	56%
	Too many	10%	10%		11%	11%	44%

* Note: The totals are less than 100% in cases where individuals checked the "no opinion" option

Table 1: Comparison of Recommender Systems: Number of Ratings Required and Recommendations Given

Time to	Books			Movies		
	Amazon	Sleeper	RatingZone	Amazon	MovieCritic	Reel
Register	0.24	0.37	0.17	0.24	0.76	0.00
Receive Recs.	2.56	1.31	0.53	0.88	4.30	0.63

Table 2: Comparing Recommender Systems: Time Measures

Input and Output for the Recommender Systems: We examined the interface for RS from a number of perspectives (See Table 1). First we examined how people felt about the number of ratings they had to provide to receive recommendations, and the number of results they received. Of the book sites, only at Amazon (books) did a majority of users feel that the number of ratings required by the system was "just right." For the other two book sites, the opinions diverged in an interesting way- at Sleeper and RatingZone, about the same number of users felt that the system asked for too much information as felt that it required "not enough." At the movie sites, the ratings are more consistent: almost half the users felt that Amazon and Reel did not require enough information, while nearly half felt MovieCritic asked for too much. All others rated the movie system to be "just right" in the number of ratings required. As for results received, the book sites once again had mixed ratings, with the majority at RatingZone and nearly half at Amazon claiming there were too few results. On the other hand, a majority of users at all the movie sites felt that the number of results was just right.

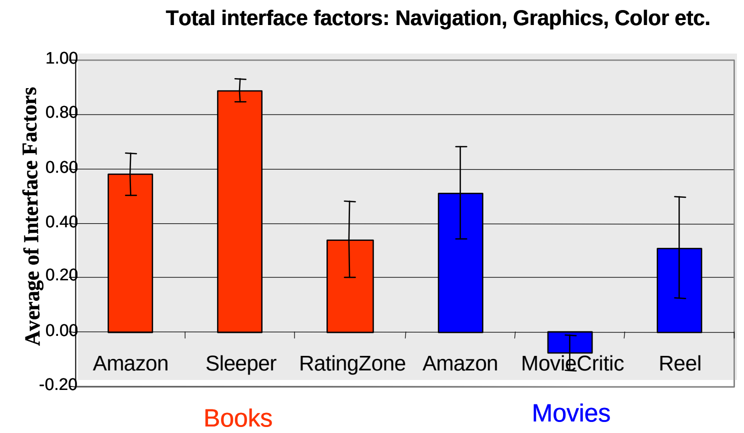

Figure 2: Comparing Recommender Systems on User Interface Factors

Time Measures: Next, we compared the RS on time taken to register at the website and to receive recommendations (See Table 2). Reel was the only website that did not ask people to register, while MovieCritic took the longest to both register and to receive recommendations. The two systems that took the least time to register and get recommendations (Reel and RatingZone) were the only systems not named as

Interface	Books			Movies		
	Amazon	Sleeper	RatingZone	Amazon	MovieCritic	Reel
Item Description	0.84	0.84	0.95	0.93	0.97	0.87
Instructions	0.63	0.67	0.67	0.71	0.67	0.50
Page Layout	0.00	0.84	0.79	0.73	0.53	0.73
Navigation	0.71	0.52	0.71	0.88	0.50	0.67
Graphics	0.42	0.63	0.88	0.50	0.50	0.67
Use of Color	0.52	0.67	0.63	0.53	0.50	0.71
Interface	0.24	0.44	0.44	0.50	0.18	0.57

Table 3: Comparing Recommender Systems: Interface Issues

favorites in post-test interviews.

Interface Factors: Users were asked to indicate whether the system's navigation, layout, color, graphics or user instructions had a positive / negative impact on their experience (See Table 3 and Figure 2). Sleeper performed best overall, followed closely by Amazon. Reel and RatingZone performed at about the same level, with a fairly wide range of ratings from users. MovieCritic was the only system to receive consistently negative interface ratings, primarily in the areas of layout and navigation. Our analysis shows no correlation between graphics, color etc. and perceived ease of use / satisfaction.

Overall Usefulness of System

We also asked users to rate the overall usefulness and ease of use of each RS. Table 4 (below) shows the correlations between the rated usefulness and ease of use of a system with the other metrics we created. The table shows that the overall usefulness of a system correlated highly with % Good and Useful Recommendations. It also correlated with % Previously Experienced and % Trust-Generating Recommendations.

Ease of use correlates with aspects of the interface such as User Instructions and Navigation. Ease of use does not correlate with the number of ratings required to receive recommendations. This is interesting because it indicates that people do not mind spending a few minutes indicating their choices to receive quality recommendations.

The Description of Item ratings indicate whether users felt the system provided enough information for them to make a decision as to interested or not interested. This metric correlates highly both with overall usefulness of the system and ease of use.

Correlation with Overall Usefulness		Correlations with Overall Ease of Use	
Ease of Use	.25 *	Description	.42 **
% Previously Experienced	.37 **	Instructions	.34 **
% Trust-Generating Recs	.32 *	Layout	.47 **
% Useful Recs.	.41 **	Navigation	.51 **
% Good Recs.	.55 **	How many ratings needed	-0.6
Description of Item	.39 *	Overall Usefulness	.26 *
		Description of item	.42 **

* significant at .05 level, ** significant at .01 level

Table4. RS Correlations: Usefulness and Ease of Use

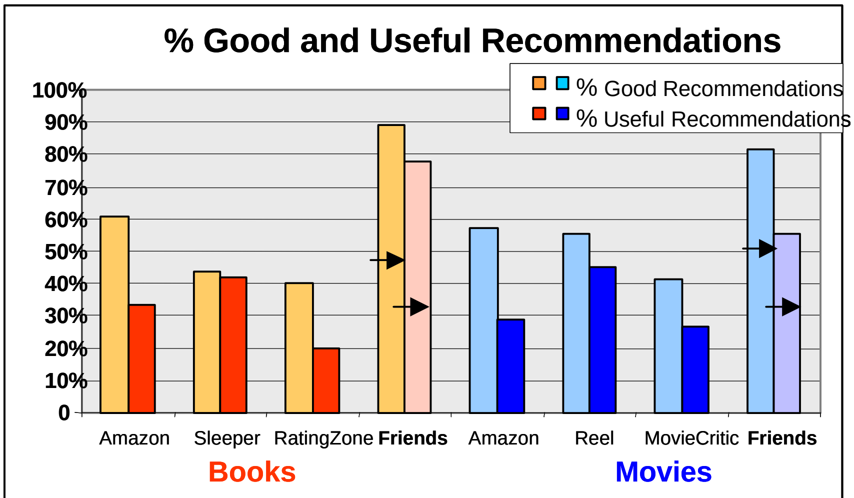

Figure 3: Comparing Human Recommenders to RS

Comparing Recommendations made by Online RS and by Friends

The bulk of our analysis focused on comparing the quality of the recommendations made by friends and by RS on three metrics (Good, Useful, and Trust-Generating Recommendations).

Good and Useful Recommendations: Next we examined the differences in the quality of the recommendations provided by both friends and RS. As the Figure 3 shows, for Good recommendations, friends performed at significantly higher levels than

RS—again (the arrows on the friend bars indicate the RS averages). (Friends=85.44, RS=45.99, t=5.17, p<.000). The same pattern was repeated for Useful recommendations (Friends=67., RS=32.57, t=5.26, p<.000). During a post-test interview we also asked users to indicate which gave them the best overall set of recommendations--one of the 3 online RS or their friends. Despite the friends' strong performance, 11 of the 19 users said they preferred an online RS: Amazon-Books (3), Amazon-Movies (3), Sleeper (3) and MovieCritic (2). This finding does not support our hypothesis that users would prefer recommendations made by friends over those made by online RS. We propose possible explanations in the Discussion section.

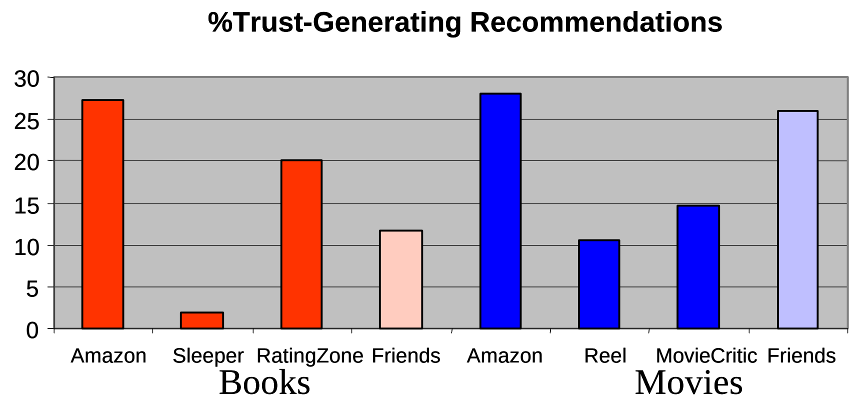

Figure 4: Comparing Human Recommenders to RS for Trust-Generating Recommendations

Previously Experienced Recommendations: On average, the percentage of items previously experienced is higher for movies than for books (Movies = 37.4, Books = 17.79, t=3.89, p<.000). This suggests that it is easier for both RS and friends to tap into movies previously experienced by a user. This could be indicative of greater accuracy in movie predictions, or it could be indicative of a smaller universe of items for movies than for books. Of the items previously experienced, a larger percentage of books fell in the Trust-Generating category (Movies =54.70, Books =89.88, t=3.88, p<.000).

Trust-Generating Recommendations: Recommended items that had been previously liked by users play a unique role in establishing the credibility of the Recommender Systems. Such items are not useful in the traditional sense (i.e., recommendations which the user can use in the future), but help generate trust in the system. Figure 4 shows that Amazon had the highest % of trust-generating recommendations. (It should be noted that we had asked friends not to recommend items that they were aware that their friend had watched/read. This might have placed friends at a disadvantage for this metric.) In post-test interviews, 7 users cited the RS' ability to suggest items they had not heard of as a key advantage over recommendations offered by friends.

DISCUSSION AND DESIGN RECOMMENDATIONS

The quantitative results of our experiment indicate that users prefer recommendations made by their friends to those made by the set of online RS we tested in our study. Though users preferred recommendations made by friends, they expressed a high level of overall satisfaction with online RS. Their qualitative responses in the post-test questionnaire indicated that they found the RS useful and intended to use the systems again. This seemed to be due in part to the ability of RS to suggest items that users had not previously heard of. However, not all RS performed equally well. Therefore, we analyzed both the quantitative and qualitative data gathered in the study to isolate design elements of the RS that contributed to their success. Based on our analysis we offer several design recommendations for RS.

1. **Users don't mind rating more items initially to receive quality recommendations**. Our results indicate that an increase in number of ratings required does not negatively affect ease of use. Some of the systems that required the user to make many ratings (e.g. Amazon, Sleeper) were rated high on satisfaction and usefulness of system. Ultimately what matters to users is whether they get what they came for: useful recommendations. Users appear to be willing to put in a little time and effort if that outcome seems likely.
2. **Allow users to provide initial ratings on a continuous rather than binary choice scale**. Several participants commented favorably on the Sleeper interface that allowed them to express gradations of interest level, rather than forcing them into making ratings on a binary choice or a 4-5 item scale.

3. **Provide enough information about the recommended item for user to make a decision. Make this information readily available.** The presence of longer descriptions of individual items correlated positively with both the usefulness and ease of use of RS. This indicates that users like to have detailed information about the recommended item, so that they can evaluate whether the recommendation is indeed useful. This finding is reinforced by the difference between the two versions of Rating Zone. About midway through our study, RatingZone underwent a redesign. One of the main changes was that a lot more information was provided about recommended items. The first version of RatingZone's Quick Picks did not provide enough information and user evaluations were almost wholly negative as a result. User evaluations were more positive for the second version. A different problem occurred at MovieCritic, where detailed information was offered but users had trouble finding it, due to poor navigation design.
4. **Provide easy ways to generate new recommendation sets.** RatingZone's Quick Picks initially generated a very short list of items but did not offer the means to see more recommendations-users found themselves at a dead end in the system. For this reason, 3 of the 10 users found no useful recommendations at RatingZone.
5. **Interface matters, mostly when it gets in the way.** In designing the interface, navigation and layout seem to be the important factors (i.e., they correlate with ease of use and perceived usefulness of system). Color and graphics are less important. For example, MovieCritic was rated negatively on layout and navigation. This affected ease of use and subjective usefulness ratings even though it performed well in terms of Good and Useful recommendations.

LIMITATIONS OF PRESENT STUDY / FUTURE PLANS

Conclusions drawn from this study are somewhat limited by several factors. (a) One limitation of our experiment design was that we handicapped the systems' collaborative filtering mechanisms by requiring users to simulate a first-time visit, without any browsing, clicking, or purchasing history. This deprived systems such as Amazon and MovieCritic of a major source of strength--the opportunity to learn user preferences by accumulating information from different sources over time. (b) A second limitation is the probable bias in favor of friends' recommendations: users knew which recommendations came from systems and which from their friends. In the post-test interviews, several users acknowledged that they simply had more faith in the quality of items recommended by friends. Currently, we are planning a study of music recommender systems. In that study, we intend to anonymize the source of the recommendations. Users will be asked to rate their level of interest in an item as before, but they will not find out if the item was recommended by a friend or an online system. (c) A third limitation is that we did not study a random sample of online RS. As such are results are limited to the systems we chose to study. (d) Finally, this study suffers from the same limitations as any other laboratory study: we do not know if users will behave in the same way in real life as in the lab.

REFERENCES

- David Goldberg, Daniel Nichols, Brian M. Oki, and Douglas Terry. "Using Collaborative Filtering to Weave an Information Tapestry." Communications of the ACM, December 1992. 32(12)
- Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins, "Eigentaste: A Constant-Time Collaborative Filtering Algorithm," Information Retrieval, accepted January 2001.
- P. Resnick and H.R. Varian, "Recommender systems." Communications of the ACM, 1997. 40(3) 56-58
- J. Ben Schafer, Joseph Konstan, and John Riedl. "Recommender Systems in E-Commerce." ACM Conference on Electronic Commerce 1999. http://www.cs.umn.edu/Research/GroupLens/ec-99.pdf