User Response to Two Algorithms as a Test of Collaborative Filtering

Adam W. Shearer
Lewis and Clark College
Portland, OR 97219, USA
Tel: 1-503-335-8088
In cooperation with The University of Minnesota GroupLens Research Team
E-mail: shearer@lclark.edu

ABSTRACT
The purpose of this experiment was to determine whether recommendations based on collaborative filtering (CF) are perceived as superior to recommendations based on user population averages. The test vehicle was a movie recommender. 29 subjects were divided into 2 groups, each group using one of these systems. The recommneder systems suggested movies which subjects later viewed. Each subject filled out pre and post-questionnaires about their experience. Subjects using the CF algorithm rated more movies. Subjects placed slightly more confidence in the recommendations of the population averages algorithm. Both algorithms were over-confident compared to subjects ratings. Subjects found both recommender systems to be an effective source of finding entertainment. User responses did not reveal a noticeable difference between the two algorithms.

Keywords: Collaborative Filtering, Recommender System

INTRODUCTION

In today s society we have a challenge. We are faced with a vast amount of information and we must choose what is relevant to our needs. People are always looking for ways to save time. One way to do this is through the use of collaborative filtering (CF). CF helps address information overload by using the opinions of users in a community to make personal recommendations for any number of desirable items including consumer goods, scientific journals, news, etc. to each user [1].

The purpose of this experiment was to determine whether recommendations based on collaborative filtering are perceived as superior to recommendations based on user population averages. When using information retrieval systems, do users experience more confidence and a higher level of success when using a collaborative filtering algorithm that will produces personalized suggestions rather than when they are using an algorithm that produces suggestions based on average ratings of an entire user population? In asking this question we are learning whether or not these two mathematical processes will produce recommendations that differ in quality. The test vehicle for this experiment was a web site called MovieLens, a free service provided by GroupLens Research at The University of Minnesota [3].

Automated collaborative filtering systems, like the one used in MovieLens, work by collecting human judgments (known as ratings) for items in a given domain and matching together people who share the same information needs or the same tastes [2]. In other words, someone using a CF system will give ratings one at a time to items they have had experience with. Then the filter puts these users into community groups of other like-minded individuals who have given the same, or similar, ratings to the same items. After the system has grouped users through the use of an algorithm it makes suggestions to these users based on the likes and dislikes of the other users in their community. Users of a CF system share their analytical judgments and opinions regarding each item that they consume so that other users of the system can better decide which items to consume [2].

METHOD

Subjects

29 computer literate persons from The University of Minnesota community and the Portland area were compensated with ten dollars worth of movie gift cards for participating in this experiment. 14 were in the control group and 15 were in the experimental group. The experiment took on average 30 minutes to complete.

Equipment

The MovieLens web site was used on both Macintosh and PC computers. All subjects used Netscape Navigator. All subjects used standard mouse input devices. The experimental web sites were arranged specifically for this study and were nearly identical to the public MovieLens web site [3] with a few slight differences. Features that did

not pertain to the experiment were removed and the sites disregarded all personal information.

Procedure

Subjects were assigned to either the experimental group, collaborative filter condition (CFC), or the control group, population average condition (PAC). This was a single blind experiment. For the first phase of the experiment subjects were presented with lists of movies and rated the ones they had seen using a five star rating scale. Subjects were instructed to rate as many movies as possible and were told that as they rated more movies the predictions given by MovieLens would become more accurate. Each subject rated at least 23 movies. When finished they chose a genre and time period from which they would view a movie. A top-ten-list was created and subjects were asked to rent the highest rated movie possible. At the end of the session subjects completed a short pre-questionnaire concerning their experience with MovieLens. Subjects were given a $5 movie gift card to rent their selection. The movies were viewed in the privacy of their homes. Immediately after viewing their movie subjects completed another short post-questionnaire similar in format to the pre-questionnaire. Subjects also rated the movie on a 5 star scale.

RESULTS

In the first phase of the experiment, subjects were allowed to choose how many movies to rate. CFC subjects rated on average 78.3 movies. PAC subjects rated on average 59.8 movies. An unpaired t-test found a non-significant trend for CFC subjects to rate more movies than PAC subjects $t(22)=1.49, p=.15$.

The next data analyzed involved subject responses to the pre-questionnaire. Unpaired t-tests showed no significant difference between the CFC and the PAC when asked to rate the following; positive experience, ease of use, and interest in using MovieLens in the future. A marginally significant result was found when subjects were asked to rate their confidence in the predictions given. PAC subjects were found to have more confidence $t(26)=1.85, p=.08$.

When looking at subject responses to the post-questionnaire unpaired t-tests showed no significant difference between the CFC and the PAC in the following areas; positive experience, satisfaction after viewing, novelty of suggestion, and interest in using MovieLens in the future. While there was no significant difference between CFC and PAC when evaluating MovieLens as an effective source of finding entertainment, MovieLens was viewed favorably by both groups $t(17)=3.55, p=.002$.

Comparing MovieLens predicted ratings with subjects ratings of movies reveals the strongest finding. A two-way ANOVA shows that predicted ratings (4.64) were higher than subject ratings (3.25) for both algorithms $F(1,16)=12.5, p=.003$. There was a marginally significant interaction between the predicted and subject ratings depending on the algorithm used $F(1,16)=2.97, p=.1$. The disparity between the predicted and subject rating was greater for the CFC (4.82 vs. 2.96) than it was for the PAC (4.36 vs. 3.71).

DISCUSSION

Analysis of the pre-questionnaire tells us that the differences between the CFC and PAC web sites were undetectable to subjects. Subjects from both groups found the experience to be positive, the site easy to use, and suggested that they may use the MovieLens system again in the future. Analysis of the pre-questionnaire also reveals that the PAC subjects tended to have more confidence in the suggestions made by MovieLens than the CFC subjects. One can infer that this is because the PAC subjects were more likely to have prior knowledge of their suggestions. This finding does not support the use of CF. If a consumer does not have confidence in a recommendation they will not pursue it.

Analysis of the post-questionnaire tells us that after viewing their suggested movies subjects still reported no difference in positive experience or the likelihood of repeated use. They also reported no difference in satisfaction with MovieLens, the novelty of suggestions given by MovieLens, or MovieLens being an effective source of finding entertainment. Subjects who completed this experiment did not find either algorithm to be superior.

MovieLens predictions were significantly higher than subject ratings, particularly in the CFC. This overconfidence of the collaborative filtering algorithm may be due to the fact that it predicts extreme highs and lows. The population average algorithm is more likely to make neutral predictions based on their very large sample sizes.

From a marketing perspective, observing the number of items rated in each user group may be the most salient finding. More ratings translate to greater exposure to the web site. The collaborative filtering algorithm tends to encourage users to stay on the site longer.

REFERENCES

1. Badrul M. Sarwar, Joseph A. Konstan, Al Borchers, Jon Herlocker, Brad Miller and John Riedl. Using filtering agents to improve prediction quality in the GroupLens research collaborative filtering system, in *Proceedings of the ACM 1998 conference Computer supported cooperative work* (1998), 345
2. Jonathan L. Herlocker, Joseph A. Konstan, Al Borchers and John Riedl. An algorithmic framework for performing collaborative filtering, in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (1999), 230 — 237
3. MovieLens web site. Available at www.movielens.umn.edu