Remembering the Stars? Effect of Time on Preference Retrieval from Memory

Dirk Bollen  
Human-Technology Interaction  
Eindhoven University of Technology  
The Netherlands  
d.g.f.m.bollen@tue.nl

Mark Graus  
Human-Technology Interaction  
Eindhoven University of Technology  
The Netherlands  
m.p.graus@tue.nl

Martijn C. Willemsen  
Human-Technology Interaction  
Eindhoven University of Technology  
The Netherlands  
m.c.willemsen@tue.nl

ABSTRACT
Many recommendation systems rely on explicit ratings provided by their users. Often these ratings are provided long after consuming the item, relying heavily on people’s representation of the quality of the item in memory. This paper investigates a psychological process, the “positivity effect”, that influences the retrieval of quality judgments from our memory by which pleasant items are being processed and recalled from memory more effectively than unpleasant items. In an offline study on the MovieLens data we used the time between release date and rating date as a proxy for the time between consumption and rating. Ratings for movies tend to increase over time, consistent with the positivity effect. A subsequent online user study used a direct measure of time between rating and consumption, by asking users to rate movies (recently aired on television) and to explicitly report how long ago they watched these movies. In contrast to the offline study we find that ratings tend to decline over time showing reduced accuracy in ratings for items experienced long ago. We discuss the impact these rating dynamics might have on recommender algorithms, especially in cases where a new user has to submit his preferences to a system.

Categories and Subject Descriptors
H.1.2 [Models and principles]: User/Machine Systems-software psychology; H.4.2 [Information Systems Applications]: Types of Systems-decision support; H.5.2 [Information Interfaces and Presentation]: User Interfaces-evaluation/methodology, interaction styles, user centered design

General Terms

Keywords
memory biases, temporal effects, explicit feedback

1. INTRODUCTION
Watching a movie, at home or in a movie theater, provides an experience with all sensory information available and evoking an affective response. After seeing a movie all this information, vividly available, influences our judgments and preference processes. This information will quickly be transferred and encoded into one’s memory. The concrete information as well as the emotional and affective responses related to the experience of watching a movie tend to fade over time as the result of a natural forgetting process. This would imply that one’s ratings for the same item might differ when given directly after the experience of the item, compared to a few months or years later. For example, the positivity effect that was established in memory research [2,3] shows more effective encoding and recall of positive information, which would suggest that we would rate items more positively if the time between consumption and rating increases.

In the present paper, we investigate the effect of time on preference retrieval in two ways. Using existing offline data from MovieLens1 we took the difference between release date and rating data as a (less than ideal) proxy for the time between consuming and rating an item. Our data shows that ratings for movies increase over time, showing a positivity effect. In a second online study, we used a more accurate measurement of the time between consumption and rating by asking users after providing ratings to report directly how long ago they had experienced the item. We also asked for explanations for some ratings. This study revealed a negativity effect of time on ratings, but a positivity effect in terms of what items users like to explain.

2. RELATED WORK
2.1 Positivity effects in memory research
In psychological literature the process of memories becoming more positive over time is known as the positivity effect. This effect is part of the Pollyanna principle that describes that pleasant experiences are being processed more efficiently and accurately in memory than less pleasant experiences [2]. This causes retrieval of information from our memories to be biased towards positive aspects. The result of this is that people forget the negative aspects of an experience, while remembering the positive aspects. The positivity effect is a well-established and robust effect in psychology. Walker et al. [3] provide an overview of several studies in what they call the fading affect bias: the process underlying the stronger decrease of the emotional intensity for unpleasant experiences in autobiographical memory. Kwon et al.[1] show that positivity effects hold over different cultures.

2.2 Temporal Effects in Recommender Systems
A number of studies acknowledge the effect of time on user preference. Wang et al. [4] addressed these temporal dynamics, that they call preference evolution, using an Ant Collaborative Filtering algorithm in which users and movies are marked with a

---

1 Available at: www.grouplens.org
pheromone trail, analogous to how ants communicate. When a user rates an item positively, pheromones are transferred from the user to the item and vice versa. As more ratings are processed, the pheromone patterns for all movies and users become more complex and predictions can be made based on similarity in pheromone patterns between users and movies. Preference evolution is taken into account by letting the pheromone decay over time and processing the ratings in the same order as they were submitted to the system, which results in later ratings contributing more to the final model. While this algorithm takes temporal dynamics into account by putting more weight on recent ratings it provides no understanding in how preferences change as time between consumption and rating increase.

Koren [5] attributed the change in rating behavior to what he calls concept drift. By means of an offline study of the Netflix contest dataset he investigated what common patterns can be found in how ratings evolve over time. An important finding was that as the time between movie release and rating increases, the rating tends to become higher.

3. OFFLINE STUDY: MOVIELENS DATA

Koren’s [5] results show that older movies are rated higher on average. However, to provide a good answer to our question, of how the time between viewing and rating influences ratings, the difference between release date and time of rating (as used by Koren [5]) might not be the most accurate proxy.

To more exhaustively investigate memory effects, we study the difference between popular and unpopular movies. The rationale is that popular movies appear more often on television and are discussed more often, this results in the actual experience remaining more freshly in memory and thus being less prone to memory distortions. For less popular movies, the time between release date and rating date is a better proxy for the time that has passed between experience and rating. In other words, any memory distortions should be more prominent in less popular movies, for which we rely on memory more.

For our analysis, we calculate as independent variable $\Delta t$, the time difference between the year of release (which is 2000 since we only considered movies released in that year) and the moment of rating by the MovieLens users. For further investigation we divided the movies into popular and unpopular movies, by performing a median split on the number of ratings for each movie.

Figure 1 depicts the results of a multilevel regression with a random intercept for movies and two fixed effects, $\Delta t$ and popularity. The results show that an increase of time difference ($\Delta t$) has a positive significant effect on the mean rating ($t=9.12$). This effect is qualified by a significant interaction with movie popularity ($t=-3.06$). Ratings for less popular movies increase about 0.04 stars per year (dotted line), whereas ratings for more popular movies only increase about 0.02 stars per year (solid line).

These results show, like [5], that ratings for movies tend to increase over time. It also shows that this effect is stronger for less popular movies for which we expect that people have to rely on their memory more.

4. USER STUDY

Our offline MovieLens study provided some initial support for the existence of the positivity effect in which unpopular ratings tend to increase more over time, since users need to rely more on memory than for popular ones. However, the measure $\Delta t$ used in our offline study is at best a (imprecise) proxy for the time factor we are really interested in: the time in between experiencing (or consuming) and rating.

To the best of our knowledge there are no studies that have taken this time as a measure into account. The goal of our study was to investigate how the time passing between the moment of consumption and the moment of rating influences a user’s ratings by asking the users directly for an estimate how long ago they have consumed the item.

4.1 Design

For the experimental setup IMDb was crawled to find information consisting of a movie cover and a synopsis for 156 movies that were aired on public Dutch television in the month prior to the study. These movies were selected in order to maximize the probability that the data would include ratings with a low $\Delta t$ as it is likely that our participants had at least seen a few of these movies on TV in the month prior. Participants in the study were provided lists of 10 randomly selected movies at a time and asked to rate the movies they knew. For every movie they rated they were asked to indicate how long ago they watched it via a dropdown menu (six categories: last week, last month, last 6 months, last year, last three years, longer ago). After having provided at least 20 ratings, participants would see a button allowing them to finish the task. They could also continue rating as long as they wanted (one person rated a total of 71 movies out of 156). After the rating task, participants were asked to provide motivations for two ratings they gave. One dropdown menu displayed the titles and ratings of movies that the participant indicated having seen within the last month, another displayed the same information for movies seen a year ago or longer. Below each dropdown menu a text input area allowed the user to enter her explanation.
4.1.1 Participants

In the course of 2 days, 100 participants finished the task (40 male, 60 female, mean age: 29.7 years, SD: 7.9 years). On average they rated 20.41 movies (SD: 9.0 movies).

4.1.2 Results

The first step of the analysis consists of visually investigating the rating distributions. In total 2041 ratings were submitted by the participants. These ratings are visually represented in the mosaic plot in Figure 2, with the width of the columns representing the proportion of ratings in a given timeslot. Only 30% of the provided ratings were of movies seen in the last year.

![Figure 2: Distribution of Ratings across timeslots](image_url)

Figure 2 shows that as time progresses the proportion of positive ratings decreases. In the first four timeslots four and five star ratings make up for more than 50% of all ratings, while in the last time slot they only make up for 25%. This seems to run against the hypothesis of a positivity effect.

![Figure 3: Progression of Average Rating (dotted line for initially high rated items, dashed for initially low rated items)](image_url)

In order to test for the hypothesized positivity effect a multilevel regression was performed, using the rating value as dependent variable. Random intercepts were included for participant and movie, similar to how biases are handled in biased matrix factorization algorithms. This compensates for differences in ratings by individual users and individual movies that are rated higher or lower on average. As a fixed effect we used Time (a linear effect from 0 to 5 for each timeslot) and its quadratic effect Time Squared. In addition the movies were divided in initially high and low rated movies to provide more insight into where the pattern observed in Figure 2 originates from. Initially high rated movies (High rated = 1) were movies that in the first three time slots received a high rating (>3 stars).

In Figure 3 we plotted the average rating score as a function of time for each group. The dotted line represents the initially high rated movies, the solid line the initially low rated movies. We see a steady decline in mean rating for high rated movies, and an initial increase followed by a decrease for the mean rating of low rated movies. The multilevel regression, shown in Table 1, confirms the pattern of Figure 3. As far as the influence of time is concerned, the model shows that initially the ratings rise (the linear time coefficient is positive). However, as time progresses the Time² coefficient weighs heavier on the ratings, and they decrease. Additionally a significant interaction of Time × High Rated shows that for high rated movies the rating decreases faster.

Overall the results seem to suggest a negativity effect, rather than a positivity effect, which is also more pronounced for initially highly rated items.²

<table>
<thead>
<tr>
<th>Table 1: Linear multilevel model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficient</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>Time</td>
</tr>
<tr>
<td>High Rated</td>
</tr>
<tr>
<td>Time Squared</td>
</tr>
<tr>
<td>Time × High Rated</td>
</tr>
</tbody>
</table>

An additional analysis was performed on the explanations people provided for their ratings. After submitting their ratings participants were asked to explain why they gave the rating for a movie they saw in the last month and one they saw longer than one year ago. They could select which rating to explain by selecting it from two drop-down menus, one for each category. Two conclusions can be drawn from this data, displayed in Figure 4. Firstly, given how the ratings are distributed, the choice for explaining movies appears to have a bias toward positive movies. For movies seen more than a year ago four and five star ratings make up for about 30% of ratings, however when asked to explain a rating, participants chose to explain a four or five star rating for almost 75% of the time. For the recent movies (last month), the proportions of explanations and ratings are more similar. For these ratings however participants are more inclined to rate either good or bad movies, as can be seen by the decrease in three star rating explanations. Secondly, the length of the explanation was investigated. Although no significant difference could be found for older versus more recent ratings, people used significantly more characters to explain a positively rated item than a negatively rated item (on average 11 characters more per rating value).

Both these findings concerning the explanation people provided are related to the positivity effect from memory research discussed earlier. As positive experiences are encoded and recalled better, it might have been easier for our participants to describe the positively rated items they had seen longer ago.

² We also ran an analysis similar to our offline study, using the difference between the time of rating (2012) and the year of release of the movies in this set. This analysis also shows a significant negativity effect (a decrease of 0.02 stars per year).
This paper investigated the time dynamics underlying ratings in recommender systems. An offline analysis shows a positivity effect in which movies are rated higher as time between release date and rating date increases, especially for movies that are less popular for which we expect people to rely more on their memory. This increase in average rating over time can be mainly attributed to the increase present in low ratings for non-popular movies, which is higher than the increase in high ratings or in popular movies.

However, the real variable of interest was the time difference between the moment of consumption and the moment of rating. In a second experimental study participants were asked to explicitly indicate how long ago they watched a movie. The measure $\Delta t$ in this study is thus the time difference between moment of consumption and moment of rating as opposed to moment of release and moment of rating. However, in this study a negativity effect instead of a positivity effect was found. Results showed a decline in rating score as movies were rated longer after being watched. Additionally this effect was more pronounced for movies that received high ratings shortly after being watched.

A possible explanation for these contrasting results can be found in the different ways of interaction between actual recommender systems and our experimental setup. In the presented study participants were asked to rate all movies they recognized and indicate how long ago they watched the movie. Recommender systems typically ask new users to rate randomly sampled movies for training, which is similar to our experimental design. However, once enough training ratings have been provided, a recommender system presents movies the user is likely to appreciate. Older movies have more data and are thus more often accurately recommended, which leads to accurate recommendations and thus more high ratings. Because users mainly encounter movies they are likely to appreciate during the interaction, the amount of negative ratings is also to be expected lower than what we found in our online study.

An additional explanation can be found in the result of our user study that over time, ratings tend to regress towards the mean. The best discrimination between good and bad movies can be found when ratings are given no longer than a year after watching a movie.

These dynamics are important to take into account when designing recommendation algorithms. A situation where these dynamics might have a particularly negative influence on recommender accuracy is when a new user enters a system. Often, new users are asked to rate a number of more or less randomly selected movies so that the recommender system has enough information to calculate predictions that can later, through natural interaction, be fine-tuned. In a natural interaction it might be more common that a user rates a movie after having just seen it. With the initial ratings however, the user may rate movies that she has seen long ago, which can lead to suboptimal recommendations. In fact, depending on how data is gathered, current recommender systems may predict how users will like a movie after a certain amount of time, instead of how she will like it right now.

In order to take these effects into account in future recommender systems, two paths can be taken. First of all, users of a recommender system could be asked how long ago they watched a movie when they rate it, similar to how it was done in this study. While this additional data is likely to improve recommendation quality, a negative consequence is that users will have to provide more input. Another way to go about this is investigating the patterns related to the passing of time in more detail, with the goal of understanding and compensating for this influence on rating values.

The authors propose future studies with a more natural interaction with a recommender system, to further investigate the temporal dynamics of preferences.

5. CONCLUSION

These dynamics are important to take into account when designing recommendation algorithms. A situation where these dynamics might have a particularly negative influence on recommender accuracy is when a new user enters a system. Often, new users are asked to rate a number of more or less randomly selected movies so that the recommender system has enough information to calculate predictions that can later, through natural interaction, be fine-tuned. In a natural interaction it might be more common that a user rates a movie after having just seen it. With the initial ratings however, the user may rate movies that she has seen long ago, which can lead to suboptimal recommendations. In fact, depending on how data is gathered, current recommender systems may predict how users will like a movie after a certain amount of time, instead of how she will like it right now.

In order to take these effects into account in future recommender systems, two paths can be taken. First of all, users of a recommender system could be asked how long ago they watched a movie when they rate it, similar to how it was done in this study. While this additional data is likely to improve recommendation quality, a negative consequence is that users will have to provide more input. Another way to go about this is investigating the patterns related to the passing of time in more detail, with the goal of understanding and compensating for this influence on rating values.

The authors propose future studies with a more natural interaction with a recommender system, to further investigate the temporal dynamics of preferences.

6. REFERENCES


