A Topic Model For Movie Choices and Ratings

Timothy N. Rubin (trubin@uci.edu)
Mark Steyvers (mark.steyvers@uci.edu)
Department of Cognitive Sciences, University of California, Irvine
Irvine, CA, 92697-5100

Abstract

User input to recommendation systems such as Netflix provide an excellent opportunity to study human choice and preferences. We present a probabilistic model that captures two processes that underlie human input to recommendation systems; the process by which individuals choose items to rate, and the process by which they select a rating for those items. Using movie rating data collected by Netflix, we demonstrate that this model can generate accurate predictions about missing movie ratings. Furthermore, we show that the implicit information that users reveal through their choice processes can be used to improve prediction accuracy even in the total absence of explicit ratings.

Keywords: Choice, Decision Making, Recommendation Systems, Topic Models, LDA, Machine learning

Introduction

Recommendation systems are becoming increasingly important in industry and academia. While the field of recommender systems is heavily researched in the area of machine learning and data-mining (see Adomavicius & Tuzhilin, 2005 for an overview), it has been largely ignored by the cognitive science community. This is somewhat surprising, because an accurate model of human preferences requires understanding the basic psychological processes underlying choice and judgment. In addition, the goal of any recommendation engine is ultimately to provide a good prediction of what a particular individual will like. This requires an understanding of individual differences as they relate to preference judgments and choice behavior.

Consider the process by which you produce a movie rating. Typically, you first choose a movie to watch, then watch the movie and form an opinion of it, and finally translate this opinion into a discrete rating. This full sequence of events is important in determining what ratings are actually observed by a commercial recommendation system such as Netflix. And at each point in this process, choice plays a key role. We choose movies to watch based both on our preferences and on the situation—what mood we are in, what type of movie we feel like that night, and who we are with. And our opinion of a movie can be significantly influenced by the conditions in which we saw it (for example, you might love horror movies, but have a bad opinion of *The Shining* because it gave your child nightmares for a month). Even the process of picking a discrete rating based on an internal representation of preference involves choice.

In addition to determining which ratings are observed, choices reveal information about peoples' preferences; without knowing someone's actual movie ratings, we can get a sense of their movie tastes from which movies they see. Hofmann (2004) described the two complementary sources of information about user preference as *implicit* data (which movies users watch or otherwise show interest in), and *explicit* data (the ratings users assign to movies). The notion of implicit vs. explicit data presents an interesting question—how much, exactly, can we learn about an individual's preferences through their choices alone? Suppose that all we know about a user is that they have watched *Full Metal Jacket*, *The Godfather*, and *Goodfellas*. How accurately can we predict ratings that this user will give to other movies based solely on this information? And more to the point, how well can we make recommendations for them? Now suppose that we are told that they gave ratings of 3, 5 and 4 to these movies respectively (on a scale of 1-5). How much additional knowledge do we now have about this user? How much better can we make predictions (and recommendations) for this user?

In this paper, we attempt to answer these questions by developing a model of human ratings that describes the process by which individuals choose movies and then produce a rating for them. We develop a probabilistic framework for understanding individual differences in preference, and specify a generative model that describes how users choose movies to watch and choose ratings for these movies. After demonstrating that this model can produce interpretable dimensions of movie preferences, we compare how well this model can make predictions given different amounts of both implicit and explicit user data. We apply this model to a subset of the Netflix dataset that was released as part of a competition for researchers to develop the next generation of recommender systems (Bennett and Lanning, 2007).

The Current State of Recommendation Systems

The majority of recommendation systems currently use collaborative-filtering based techniques such as a k-Nearest-Neighbors algorithm (kNN) (Schafer et. al, 2007). Collaborative-filtering approaches typically generate recommendations for a user by finding items that have been given high ratings by similar users (where "similarity" is measured using a metric such as the Pearson correlation coefficient between the users observed ratings). While this often produces accurate predictions, the psychological underpinnings of this model are unclear; these approaches do not model latent psychological features, nor do they account for individual differences in choices. Furthermore, while collaborative filtering produces clusters which can illuminate groups of similar items, they do not produce dimensions that are readily interpretable; although

knowledge that two movies have positive covariance can be useful for predictions, it does not tell us why these two movies are similar.

Another common technique for analyzing user-ratings is singular-value decomposition (SVD), in which a matrix of ratings for a set of users is decomposed into spaces where users as well as movies are modeled as points in a high-dimensional space (Sarwar et al., 2000). This technique captures the notion that individuals can be characterized by a set of latent features. However, it is difficult to extend the SVD representation to allow for variations in the ways users and items are represented; because there are no separable dimensions, users and items cannot be similar in some respects but dissimilar in others. Furthermore, this technique does not capture the processes by which items are chosen or ratings are generated.

Modeling User Choice When considering rating data that are volunteered by a user, there are two separate processes that have significant impact on which items are rated. The first process involves movie choice—why does a particular user choose to watch a particular set of movies but not others? The second process guides rating choice—given that a user has watched some movie, what determines whether they will actually provide a rating for it, and if they do provide a rating, how do they choose a rating that reflects their opinion of the movie?

Marlin et al. (2007) showed that users are more likely to rate items for which they have a strong opinion (particularly when the opinion is favorable). These authors go on to demonstrate the significance of missing-data models for producing unbiased predictions for user-ratings. This is an important result, but for the purposes of this paper we do not account for this missing-data mechanism. Rather, we focus on the largely ignored questions of how users choose movies to watch, and choose ratings to represent their opinions of the movie.

The Ratings Topic Model

This paper presents the Ratings Topic Model, a probabilistic model of movie ratings (Figure 1). The model attempts to capture two related processes: the process of choosing a movie to watch, and the process of choosing a rating for the movie. Our model combines features of Latent Dirichlet Allocation (LDA) and the ordered-logit model to explain both processes. LDA is an established probabilistic framework for extracting latent dimensions from data, particularly in the field of corpus analysis (Blei et al., 2003). The ordered-logit model is an econometric model for Likert rating scales (Train, 2003), and is related to the polytomous Rasch model studied in psychometrics (Andrich, 1978). Our model is related to a model proposed by Hofmann (2004). However, Hofmann (2004) focuses on a formulation of this model in which user choice processes are not explicitly considered and do not influence users' ratings. Furthermore, his model lacks a generative process by which users convert their preferences into discrete ratings.

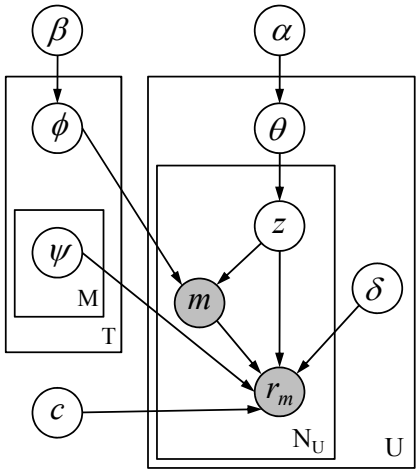

Figure 1: Graphical Model for the Ratings Topic Model

The Ratings Topic Model addresses some of the weaknesses inherent to both collaborative filtering and SVD-based approaches to modeling ratings. In addition to describing the role that choice processes play in determining what data is observed, LDA produces a set of separable latent dimensions of human preference. Without modeling separable dimensions, it is difficult to explain the underlying reasons why sets of items are rated similarly. This is particularly true with something as complex as human preferences, since items can be liked or disliked for different reasons by different users. Additionally, items or people can be highly similar with respect to one feature (e.g. a particular genre), while being dissimilar with respect to a different feature. For example, which of these would you consider more similar to the television series *The Sopranos*: *Casino*, or *Sex and The City*? It is likely that people would disagree on this answer, because although the genre of *The Sopranos* is closer to that of *Casino*, *Sex and the City* is similar to *The Sopranos* in that it they are both critically acclaimed television series produced by H.B.O.

Our probabilistic approach employs LDA to model user movie choices and preference, and an ordered-logit model to capture the process by which preferences are converted into an observed rating. We assume that users can be modeled as mixtures of topics, and that each topic represents a probability distribution over movies and preferences. In this process, once a user has selected a topic, some movies are more likely than others to be watched, and some movies are more likely than others to be enjoyed. Intuitively, we can think of a topic as any feature that might guide what people choose to watch or how they rate it (e.g. genre, release date). Once a movie has been selected, the user's rating for the movie is a function of the topic used to choose it.

The Ratings Topic Model is a generative model in that it defines a process to generate the distribution of preferences and choice probabilities for each topic, and the process by which users produce a set of ratings on the basis of these topics. For all topics $z = 1 \ldots T$, we pick a multinomial probability distribution over movies ϕ, which determines the

probability $p(m_i|z_i = j)$ of choosing each movie, $m = 1 \ldots M$, given a topic, z. For each topic, movies are independently assigned a preference parameter $\psi_{m,t}$ which determines how much a user will enjoy the movie given the topic used to choose it.

For each user, we first sample a multinomial mixture of topics (θ) from a Dirichlet prior α. This mixture determines the probability $p(z_i|u)$ that the user's choice and rating will come from topic z. Each time we produce a rating for a user, we first select a topic according $p(z_i|u)$, and then select a movie from that topic according to $p(m_i|z_i = j)$. The probability that the user will choose movie i is given by:

$$p(m_i) = \sum_{j=1}^{T} p(m_i|z_i = j)P(z_i = j)$$

Once a movie has been selected, a numerical rating for that movie is generated according to the probabilities specified by the ordered-logit component of the model.

The ordered-logit model treats ratings as a function of utility (U), which we define as the sum of the preference parameter and a bias parameter: $U_{u,m} = \psi_{t,m} + \delta_u + \varepsilon$. The bias parameter δ_u is specific to each user and determines the general tendency of a user to give favorable ratings. The probability of observing rating r_i is defined as probability that U falls between the rating thresholds $c_i$and c_{i+1}. Noise is modeled using a logistic function, such that:

$$P(r = r_{u,m}|\, \psi_{t,m}, \delta_u, c) = P(c_i < U_{u,m} < c_{i+1}) = \frac{1}{1 + e^{\psi_{t,m} + \delta_u - c_{i+1}}} - \frac{1}{1 + e^{\psi_{t,m} + \delta_u - c_i}}$$

The rating-thresholds c determine which values of U correspond to each of the possible observed ratings (1...5) and are set globally – all users are assumed to have to same set of rating thresholds (but different biases). Figure 2 Illustrates how relative rating probabilities change as a function of U.

Model parameters were learned through Markov-Chain Monte Carlo methods, using a hybrid of Gibbs sampling and Metropolis-Hastings steps. Details of inference procedure are provided in supplementary material.[1]

Dataset The Ratings Topic Model was evaluated on a subset of the Netflix dataset. This dataset is comprised of over 100 million anonymized user ratings on movies and television shows collected between 1998 and 2005. For model evaluation we selected a relatively dense subset of 500 movies and 10,000 users, containing approximately 950,000 ratings (about 20% of elements were thus filled, in contrast to 1% for the full Netflix dataset). The model was run using T = 1, 10, 20, 25 and 50 Topics.

Topic Examples

For every topic, a number of informative features can be visualized: (1) a ranking based on $p(m_i|z)$ that shows the movies most likely to be chosen given that a user has selected the topic, (2) a ranking based on $\psi_{m,t}$, showing the movies which have the highest and lowest expected ratings given the topic, and (3) a ranking based on $p(r, m_i|z)$ illustrating the movies with the highest joint probability of being chosen and being assigned rating of either a 1 or 5. Figure 3 illustrates these features using three topics taken taken from a single Gibbs sample using T=25.

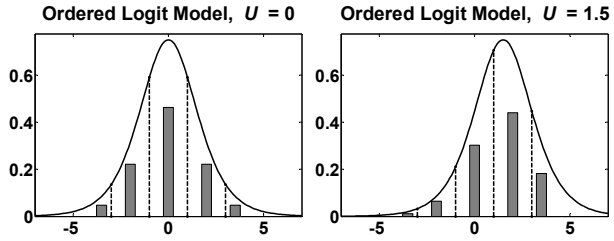

Figure 2: Left panel illustrates the logistic distribution for U=0, with rating thresholds depicted by dashed vertical lines. The shaded bars show probabilities of each rating. The right panel illustrates how rating probabilities change when U shifts from 0 to 1.5. When U=0 the most likely rating is a "3"; when U=1.5, the most likely rating is a "4".

Probability of Movies Given a Topic A Topic's probability distributions over movies models the processes guiding movie choice. Since movie choice is an overt process, it is not surprising that this feature typically discriminates topics in an intuitive manner. The movies that are most likely to be chosen given some topic usually have obvious thematic similarities. Looking the examples given in Figure 3 we can see that the movies most likely to be chosen under each topic are from similar genres. For example, the movies most likely to be chosen under Topic 4 are all horror films, with an emphasis on "classic horror" films. The movies most likely to be chosen under Topic 20 are fairly recent romantic comedies, while those in Topic 23 are mostly recent crime dramas.

Expected Movie Ratings Given a Topic While the choice dimension of a topic is highly interpretable, it does not always reflect user preferences; just because people are likely to watch a movie doesn't mean that they are likely to enjoy it. To interpret the topic along the dimenion of preference, we can look at which movies have the highest and lowest expected ratings given some topic (this is a function of parameter $\psi_{m,t}$).

For example, consider a person that often chooses movies according the distribution in Topic 20 (i.e., he is very likely to watch romantic comedies), and suppose that he is browsing for this type of movie one night. The model predicts that he is likely to enjoy movies with high values of $\psi_{m,t=20}$. Thus, even though he is more likely to choose *10 Things I Hate About You* than season 5 of *Sex and the City*, the model predicts that he will be more likely to enjoy *Sex and the City*. On the other hand, a person that is in the mood for a crime drama and therefore chooses a movie from Topic 23 is expected to strongly dislike *Sex and the City*.

[1] http://www.socsci.uci.edu/~trubin/

This does not mean that everyone who watches a lot of *The Sopranos* is expected to dislike *Sex and the City*; it just means that if they choose *Sex and the City* from this topic they they are unlikely to enjoy it. In fact, a user choosing from Topic 5 (not shown) has a high probability of giving both *The Sopranos* and *Sex and the City* high ratings.

Joint Probabilities of Ratings and Choices While the predicted rating for a movie under a topic can be highly informative, it often does not tell the whole story. Since these predictions are conditional on users actually choosing the movie under the topic, the probability of observing high or low ratings from movies that have the highest and lowest expected values may be relatively small. On the other hand, if we consider the movies that are most likely to be both chosen and given a high or low rating, we are often able to find the most liked or disliked movies that are topic-relevant. For example, *The English Patient* has a very low expected rating under Topic 4, but if also has a low probability of being chosen. However, looking now at the movies that are most likely to be given a rating of 1 under the topic, we see mostly topically relevant movies that tend to be disliked, such as *The Ring Two* and *The Grudge*. Since these movies are often chosen and subsequently disliked, we label these movies as those that are "most likely to dissapoint" users. Conversely, movies that have a probability of being chosen and then liked we label as "most likely to please".

Sometimes the fact that there is a very high preference for a movie can overcome the fact that it isn't among the most likely movies to be chosen, as with *Sex and The City* in Topic 23; although none of the seasons of this show are among the top 15 most likely to be chosen, they compose all seven of the top "most likely to please" spots, because they have such a high expected rating.

In some cases, the connection between the preference and the choice dimension is not totally intuitive. For example, both *Labyrinth* and *The Neverending Story* are expected to be well-liked by users picking from Topic 4, even though they don't have a high probability of being chosen. These sorts of "unlikely favorites" are particularly interesting when we consider the domain of recommendation systems. While it might generally be a smart approach to recommend movies with high probabilities of being chosen and also being liked, these recommendations may not always be particularly useful since the user is likely to choose them anyway. The most interesting and useful recommendations might be those movies that are unlikely choices but that nevertheless are likely to be enjoyed.

Predicting User Ratings and Choices

A standard approach for model assessment is to see how well a model can predict unobserved data. For this purpose, we removed five ratings from each user in our Netflix subset. These items were used as a test set, while all remaining ratings were used to train the model using $T = 1$, 10, 20, 25 and 50 topics. Several performance measures

Choice Dimension $p(m \mid t)$		Preference Dimension $E(r \mid m, t)$		Joint Probability $p(r, m \mid t)$
		Topic 4		
p	***Most Likely Choices***	*E(r)*	***Highest Rated***	***Most Likely To Please***
.031	Poltergeist	4.4	Labyrinth	The Exorcist
.030	Carrie	4.2	The Exorcist	Poltergeist
.029	A Nightmare on Elm Street	4.2	The NeverEnding Story	Misery
.027	Halloween	4.2	Aliens	Halloween
.025	Misery	4.1	Alien	A Nightmare on Elm Street
.024	Scream	4.0	Primal Fear	Carrie
.023	Saw	4.0	Superman: The Movie	The Lost Boys
.022	The Exorcist	4.0	Misery	Scream
.022	The Grudge	4.0	Poltergeist	Saw
.021	The Lost Boys	4.0	South Park: Bigger, Long…	Alien
.021	Friday the 13th	4.0	Lean on Me	Bram Stoker's Dracula
.020	Final Destination 2	4.0	The Life of David Gale	Aliens
.020	Stir of Echoes	3.9	Bram Stoker's Dracula	Stir of Echoes
.020	Sleepy Hollow	3.9	Thelma & Louise	Frailty
.019	Frailty	3.9	Halloween	Dawn of the Dead
.017	From Hell	3.9	The Lost Boys	Labyrinth
.017	I Know What You Did La…	3.9	Sleepers	Fatal Attraction
.016	The Haunting	3.9	Hostage	The NeverEnding Story
.016	Rosemary's Baby			
.016	Hide and Seek	*E(r)*	***Lowest Rated***	***Most Likely To Dissapoint***
.016	Bram Stoker's Dracula	2.3	Where the Heart Is	Dreamcatcher
.016	Dreamcatcher	2.3	Dr. Dolittle 2	The Ring Two
.015	Stigmata	2.2	Sneakers	White Noise
.015	Resident Evil	2.2	Team America: World Pol…	The Haunting
.014	The Ring Two	2.1	The English Patient	Catwoman
.014	The Gift	2.1	Black Sheep	The Grudge
.014	Fatal Attraction	2.0	Catwoman	Hide and Seek
.013	Alien	1.9	8 Mile	Scary Movie 2
		Topic 20		
p	***Most Likely Choices***	*E(r)*	***Highest Rated***	***Most Likely To Please***
.019	Ever After: A Cinderella St…	4.9	Sex & the City: Season 6-1	Sex & the City: Season 3
.018	10 Things I Hate About You	4.9	Sex & the City: Season 4	Sex & the City: Season 2
.015	Kate & Leopold	4.9	Sex & the City: Season 3	Sex & the City: Season 6-1
.015	Save the Last Dance	4.8	Sex & the City: Season 6-2	Sex & the City: Season 1
.015	Pretty in Pink	4.8	Sex & the City: Season 1	Sex & the City: Season 4
.014	Clueless	4.8	Sex & the City: Season 2	Sex & the City: Season 5
.013	She's All That	4.8	Sex & the City: Season 5	Sex & the City: Season 6-2
.013	The Prince and Me	4.7	Friends: Season 1	Friends: Season 2
.013	Say Anything	4.7	Friends: Season 2	Friends: Season 1
.013	Practical Magic	4.4	Sleeping Beauty	Say Anything
.012	America's Sweethearts	4.4	The Parent Trap	10 Things I Hate About You
.012	Bridget Jones: The Edge…	4.4	Singin' in the Rain	Clueless
.012	Win a Date with Tad Ham…	4.2	Sense and Sensibility	Pretty in Pink
.012	Cruel Intentions	4.2	Life as a House	Ever After: A Cinderella St…
.011	What a Girl Wants	4.2	Primal Fear	Sliding Doors
.011	Chasing Amy	4.2	The Phantom of the Opera	Breakfast at Tiffany's
.011	My Girl	4.2	Beauty and the Beast	The Parent Trap
.011	Sex & the City: Season 2	4.1	Say Anything	Little Women
.011	Down With Love			
.011	40 Days and 40 Nights	*E(r)*	***Lowest Rated***	***Most Likely To Dissapoint***
.011	Bring It On	2.3	Waiting for Guffman	Little Black Book
.011	Sliding Doors	2.3	Saving Silverman	Kate & Leopold
.011	Return to Me	2.2	Team America: World Police	Alfie
.011	Where the Heart Is	2.2	The Naked Gun	Intolerable Cruelty
.011	Sex & the City: Season 3	2.1	Eyes Wide Shut	Eyes Wide Shut
.010	Uptown Girls	2.1	Half Baked	I Heart Huckabees
.010	Sex & the City: Season 1	2.1	The Cell	America's Sweethearts
.010	Hope Floats	1.9	Little Nicky	Win a Date with Tad Ham…
		Topic 23		
p	***Most Likely Choices***	*E(r)*	***Highest Rated***	***Most Likely To Please***
.032	The Sopranos: Season 1	4.8	24: Season 1	The Sopranos: Season 1
.032	The Sopranos: Season 2	4.8	Band of Brothers	The Sopranos: Season 2
.031	The Sopranos: Season 3	4.8	The Sopranos: Season 1	The Sopranos: Season 3
.030	The Sopranos: Season 4	4.7	The Sopranos: Season 2	The Sopranos: Season 4
.024	Heat	4.7	The Sopranos: Season 3	Casino
.023	Casino	4.7	The Sopranos: Season 4	Heat
.020	Donnie Brasco	4.5	Casino	Band of Brothers
.017	Rounders	4.3	Glory	24: Season 1
.014	Swingers	4.3	Swingers	Swingers
.014	The Untouchables	4.3	Hoosiers	Rounders
.014	Sleepers	4.3	The Last of the Mohicans	Donnie Brasco
.014	The Score	4.3	Friday	Glory
.013	Primal Fear	4.3	Heat	Lock, Stock and Two Smo…
.012	Lock, Stock and Two Smo…	4.2	Apocalypse Now Redux	The Untouchables
.012	The Godfather, Part III	4.2	City of God	Primal Fear
.012	True Romance	4.2	Lock, Stock and Two Smo…	Apocalypse Now Redux
.012	The Professional	4.1	The Good, the Bad and the …	True Romance
.012	The Insider	4.1	Primal Fear	Sleepers
.011	Boyz N the Hood			
.011	Glory	*E(r)*	***Lowest Rated***	***Most Likely To Dissapoint***
.011	The Game	2.0	I Heart Huckabees	White Chicks
.011	Spy Game	2.0	The Transporter	The Transporter
.011	Apocalypse Now Redux	2.0	Beauty and the Beast	Sex & the City: Season 3
.010	Band of Brothers	2.0	Sex & the City: Season 4	Sex & the City: Season 6-1
.010	25th Hour	1.9	Sex & the City: Season 6-2	Alexander: Director's Cut
.010	The Hurricane	1.9	Sex & the City: Season 6-1	The Cell
.010	24: Season 1	1.9	Sex & the City: Season 3	Eyes Wide Shut
.010	Raging Bull	1.4	White Chicks	Sex & the City: Season 4

Figure 3: Topic features from a single Gibbs Sample, $T = 25$

were computed to evaluate how accurately the model could predict test data using different numbers of topics. Performance was compared across different values for *T*, and against several baseline predictors.

To evaluate the accuracy of rating predictions we computed both the percent of correct predictions (using a single *maximum a posteriori* prediction for each rating), and the perplexity of the posterior predictive distribution. Perplexity is a standard measure of performance in the field of information-retrieval, and is computed as $e^{\left[-\log\frac{1}{n}\sum p(r_i)\right]}$. Perfect performance (i.e. assigning all probability to the true rating) yields a perplexity of one, while a completely uninformative prediction (assigning uniform probability to all ratings) yields a maximum perplexity of 5. The three baseline predictions for ratings we used were (1) the full marginal distribution of ratings across all users and movies in the training set, (2) the marginal distribution of ratings for the movie being rated, and (3) the optimal blend of the movie's marginal distribution with the user's marginal distribution of ratings. As shown in Figure 4, the Ratings Topic Model outperformed the baseline predictions when the number of topics was greater than one. The model made the most accurate predictions when 25 topics were used.

In addition to making predictions about ratings, the Ratings Topic Model makes predictions about user choices; for each user and movie, it assigns a $p(m_i|u_j)$, where $\sum_i p(m_i|u_j) = 1$. Predictions are made after training items have been removed, such that the prediction goal is to assign as much probability to the five test items as possible. The accuracy of these predictions was measured using perplexity. In this case, an uninformative prediction (which assigns uniform probability to all movies) yields a perplexity equal to the number of movies remaining after training items are removed. For the purposes of comparison, perplexity was also computed for the following two baseline predictions: (1) assigning uniform probability to all movies being chosen, and (2) assigning each movie its marginal probability of being chosen across all training data. Results are shown in Figure 4. The Ratings Topic Model outperformed the baseline predictions when $T > 1$, and achieved best performance with T=25.

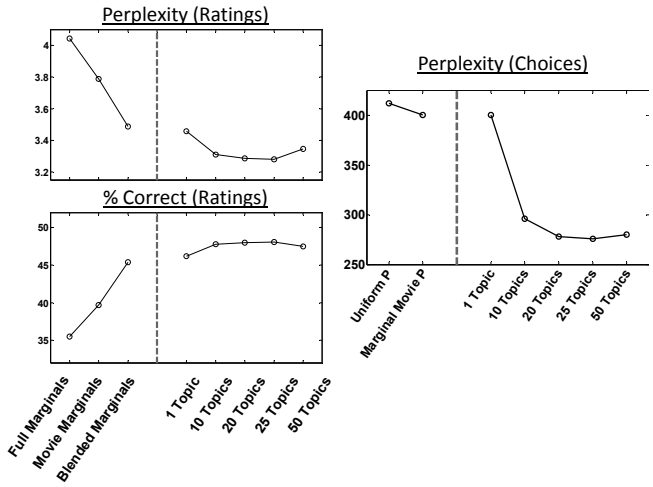

Figure 4: Accuracy of model and baseline measures for rating predictions (left) and choice predictions (right)

Implicit vs. Explicit Data

The results described in the previous section demonstrate that the Ratings Topic Model makes reasonably accurate predictions about both user choices and user ratings. For these purposes, the model uses both implicit and explicit preference data (user choices and ratings, respectively). However, it is still unclear whether the choice data itself can be used to improve rating predictions (and accordingly, whether it can improve user recommendations). In other words, is implicit data useful only for the purpose of understanding user choices, or does it capture information about user preferences, which are only explicitly observed through the ratings themselves?

To address this question, we systematically varied the amount of explicit information (i.e., the number of movie ratings) and implicit information (i.e., the number of movie choices) that was observed for each user and measured how this affects prediction accuracy for missing ratings. For this simulation, we removed a subset of 1,000 test-users from our 10,000 user subset. Complete data for the 9,000 remaining users was used to train the model on 25 topics. Topic parameters $\psi_{t,m}$ and ϕ were then fixed, so that it was only necessary to fit parameters θ and δ for each test-user.

For model evaluation, all but 50 ratings for each test-user were removed, such that we had a 1,000 user x 500 movie matrix, with 50 ratings observed in each row. This matrix was then randomly split into a training set and validation set containing 40 and 10 ratings per user respectively. The model was trained under 45 different conditions in which the number of observed ratings and choices was manipulated. (Note that since it is impossible to observe a rating without a choice, the number of choices observed here refers to the number of choices that were observed *in addition* to the observed ratings). For each condition, posterior estimates of parameters were averaged over *N* chains to generate predictions for validation data. Measures of performance under each condition were obtained using five-fold cross validation, such that all ratings in the test-set were used once in the validation set.

Measuring Performance User bias accounts for a large amount of variance in Netflix user ratings. Since bias can only be observed from users' explicit ratings, prediction accuracy does not provide a good measure to determine how much we can learn about preferences from implicit vs. explicit data. Furthermore, while it is important to account for bias when trying to accurately predict missing ratings, it is unimportant when we are interested in understanding user preferences or when making recommendations. More relevant for these purposes is the ability to predict the *relative* enjoyment of different movies. Therefore, we evaluated model performance by measuring how well it could predict which movies were rated higher than others.

For each user, all pairs of unequal ratings in the validation set provide a single comparison about relative movie preference; for each of these comparisons, we computed the posterior predicted probability that user u will give movie j a higher rating than movie k:

$$p(r_u > r_k) = \sum_{v=1}^{5} p(r_j > r_k \,|r_j = v)\, p(r_j = v)$$

We computed two measures of the accuracy of this prediction across all paired-comparisons for all users. First, we computed the perplexity of the estimate (where the baseline value of perplexity for this prediction is 2, which is obtained by assigning a .5 probability that movie j will be rated higher than movie k). In addition, we generated a binary prediction using the *maximum a posteriori* estimate of which rating would be higher, and computed the percent of these predictions that were correct. Baseline for this binary measure is 50%, since it is the expected result if we were to make random guesses. The condition with zero ratings and choices presented in the table below provides a second baseline for these measures; without any ratings or choices, predictions for all users are generated using the prior values for parameters ϕ and δ.

Figure 5 shows the perplexity and percent correct for all paired-comparisons, averaged across the five validation sets using five-fold cross validation. Looking within each row from right to left, we can see that given a fixed number of training ratings, the model is able to improve its predictions using additional knowledge about user choices. For example, for a user with 5 ratings, knowledge about 20 additional choices improves performance about as much as 10 additional ratings. Even without any ratings, knowledge about choice can significantly improve performance; the model achieves similar performance when trained with 40 choices as it does when trained with 15 ratings.

Perplexity

	Choices								
Ratings	0	5	10	15	20	25	30	35	40
0	1.823	1.806	1.793	1.787	1.782	1.777	1.775	1.774	1.771
5	1.799	1.786	1.780	1.775	1.771	1.770	1.769	1.768	
10	1.781	1.777	1.772	1.771	1.767	1.765	1.764		
15	1.772	1.766	1.768	1.762	1.762	1.760			
20	1.764	1.764	1.762	1.761	1.758				
25	1.759	1.759	1.756	1.756					
30	1.754	1.754	1.755						
35	1.750	1.753							
40	1.749								

Pct. Correct

	Choices								
Ratings	0	5	10	15	20	25	30	35	40
0	68.0	68.7	69.2	69.5	69.7	70.0	70.0	70.0	70.2
5	68.8	69.5	69.8	69.8	70.3	70.2	70.3	70.3	
10	69.7	69.8	70.1	70.1	70.4	70.4	70.4		
15	70.1	70.4	70.2	70.5	70.4	70.5			
20	70.5	70.3	70.5	70.6	70.7				
25	70.6	70.5	70.7	70.7					
30	70.9	70.9	70.6						
35	71.0	70.8							
40	71.0								

Figure 5: Prediction perplexity and percent correct for paired-comparisons, when model is trained with different amounts of choice and ratings data

Conclusion

The Ratings Topic Model provides a general framework for understanding the processes that underlie individual's rating behaviors in recommendation systems. The model can make accurate predictions about both unobserved ratings and choices, while generating interpretable dimensions that guide these processes. Furthermore, we have shown that the model can use implicit choice data in to improve predictions about a user's explicit ratings, even in the complete absence of ratings data. In addition to this being of psychological interest, it is a useful feature for real-world recommendation systems since such systems have access to a large amount of implicit preference data.

References

Adomavicius, G., Tuzhilin, A. (2005). Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering, 17*(6), 734-749.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-73.

Bennet, J., Lanning, S., *The Netflix Prize*. *KDD Cup*, 2007.

Blei, D. M., Ng, A.Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research,* 3, 993–1022.

Griffiths, T. L., Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101,* 5228-5235

Hofmann, T. (2004). Latent Semantic Models for Collaborative Filtering. *ACM Transactions on Information Systems*. 22(1): 89-115.

Marlin, B.M., Zemel, R. S., Roweis, S., Slaney, M. (2007). Collaborative filtering and the missing at random assumption, *UAI 2007: Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*.

Sarwar, B. M., Karypis, G., Konstan, J. A., Reidl, J. (2000). Application of dimensionality reduction in recommender system–A case study. *Proceedings of the ACM WebKDD 2000 Web Mining for E-Commerce Workshop.*

Schafer, J. B., Frankowski, D., Herlocker, J., Sen, S. (2007) Collaborative Filtering Recommender Systems. *The Adaptive Web: Methods and Strategies of Web Personalization*, Springer.

Train, K. (2003). *Discrete Choice Methods with Simulation*. Cambridge University Press.