Asking the Right Questions: Active Learning for Collaborative Filtering

Thomas Hofmann
Department of Computer Science
Brown University, Providence RI
th@cs.brown.edu

For many practical applications in e-commerce and in information retrieval, it is important to predict user preferences and to automatically generate recommendations. While such systems often perform reasonably well in practice, once a long enough user profile is collected, the recommendation accuracy is typically far less satisfying during the initial steps, where very little user feedback is available. This phenomenon is also known as the "early rater problem". Many deployed recommendation systems thus require an initial phase, where users vote on a number of items before the system can then be utilized to perform recommendations.

We propose to use an active learning approach to address this problem. The idea is to generate a strategy that aims at soliciting informative user ratings that provide the maximum expected accuracy for future predictions and recommendations. Formally, we will denote by u a generic user from some population of users, by y an item, e.g. a movie, and by v a vote. We will focus on the situation where the vote variables are discrete, for example, v may refer to the number of stars on a five star rating scale.

In previous work, we have proposed a statistical model for collaborative filtering, which is extended here to the active learning scenario. The model is of the following functional form,

$$P(v,y|u) = \sum_z P(v|y,z)P(y|z)P(z|u)\,, \tag{1}$$

where z is one out of a pre-specified number of states of a latent variable. Each z is supposed to model an unobserved interest pattern that represents some characteristic rating behavior shared by a subgroup of users.

We investigate an active learning setting with an explicit exploratory phase. The user is asked to rate items, one at a time and based on the outcomes new items are selected sequentially until the user decides to terminate the interview phase. In a more complicated setting one would want to combine recommendation of useful items with querying in order to continuously improve the prediction. Yet the exploration-exploitation dilemma makes this case harder to handle.

There are two subproblems that need to be addressed. The first problem deals with the question of whether or not a user is able to provide a rating for the selected item. This is very important, since in many situations it is unreasonable to assume that every user is able to answer every request for a rating. For example, not every user has seen (or even heard of) every movie. Conveniently enough, the model already provides a relative estimate for the probability that given we consider a rating of a user u, it will be on y, irrespective of the actual vote v. The latter is simply given by $P(y|u) = \sum_v P(v, y|u) = \sum_z P(y|z)P(z|u)$. Since the model always reflects our currently best estimate, this probability is the best we have available in the context of active learning.

The second problem is linked to the uncertainty of our current parameter estimate. One would like to select an item such that the user's vote would improve the parameter estimates and ultimately lead to more accurate predictions. In order to get a tractable procedure, we ignore the effect of new votes on the estimation accuracy for the parameters $P(v|z, y)$ and $P(y|z)$. This can be justified by the fact that in the limit of an infinite number of users, individual votes become negligible. Hence the only set of parameters we will consider are the user-specific mixing proportions $P(z|u)$ that indicate to what extend user votes are explained by the specific factor represented by z. We pursue a Bayesian pre-posterior approach and propose to apply the decision-theoretic framework based on the concept of the *Expected Value of Information* of an experiment. Preliminary results on the EachMovie data set confirm the validity of our approach.