Personalized Human Computation

Peter Organisciak[1], Jaime Teevan[2], Susan Dumais[2], Robert C. Miller[3], Adam Tauman Kalai[2]

[1]University of Illinois at Urbana-Champaign
organis2@illinois.edu

[2]Microsoft Research
{teevan, sdumais, adam.kalai}@microsoft.com

[3]Massachusetts Institute of Technology
rcm@mit.edu

Abstract

Significant effort in machine learning and information retrieval has been devoted to identifying personalized content such as recommendations and search results. Personalized human computation has the potential to go beyond existing techniques like collaborative filtering to provide personalized results on demand, over personal data, and for complex tasks. This work-in-progress compares two approaches to personalized human computation. In both, users annotate a small set of training examples which are then used by the crowd to annotate unseen items. In the first approach, which we call *taste-matching*, crowd members are asked to annotate the same set of training examples, and the ratings of similar users on other items are then used to infer personalized ratings. In the second approach, *taste-grokking*, the crowd is presented with the training examples and asked to use them predict the ratings of the target user on other items.

Introduction

We are studying how to complete non-normative tasks in crowdsourcing environments. Most research thus far in human computation has focused on how to generate consensus among disparate workers. Our goal is to understand how to collect results from online crowds when the standard of quality is based upon the individual tastes of a particular user rather than an objective truth. We present an initial comparison of two protocols for collecting personalized crowdsourcing results, *taste-matching* and *taste-grokking*.

Motivation and Related Work

Personalized search and recommender systems employ collaborative filtering algorithms and other techniques to generate personalized results based on prior data from humans and other sources. For example, movie selection behavior may be passively observed across many users and then used to recommend particular movies to individuals based on the behavior of related users. Our work builds on related literature in active collaborative filtering (e.g., Mahltz and Ehrlich, 1995) to explore how the crowd can be used to generate personalized results.

Rather than relying on users to passively provide annotations for personalization, we propose using crowd workers to actively collect these annotations on demand. A paid crowd can be employed at a moment's notice (Bernstein *et al*, 2011) to address the cold-start problems on previously unannotated sets of objects. This means that annotations can be collected over new types of data sets, such as personal photo collections. Additionally, because human intelligence is involved in the process, personalization can be embedded in complex creative tasks.

Many existing crowdsourcing tasks address personalization to the extent that they try to address an individual's specific needs. For example, Mobi (Zhang *et al*, 2012) provides crowdsourced itinerary planning in response to a short textual description of a trip. Likewise, selecting the "best" frame to represent a video (Bernstein *et al*, 2011) has an element of taste.

In existing systems, for the crowd to meet a user's need the user must state their desired outcome explicitly, which can be challenging to do well. Research in personalization has found that examples of a user's need can often implicitly communicate the desired outcome better than an explicit description (Teevan et al. 2010). For example, when considering the photos a person likes, someone who highly rates photos that are slightly blurry implicitly conveys that focus is not a crucial feature. It is unlikely, however, that the user would think to actively describe their photo preferences by saying, "I don't mind photos that are a little blurry."

Randomly choose training set $S \subset X$ of examples
Target user t rates each object in S
Workers $w \in W$ provide their own feedback on S
Workers are filtered based on taste similarity to target t
For each subsequent task, workers $w \in W$ provide feedback of their own tastes on remaining data $X \setminus S$

Figure 1: Taste-matching Protocol

Randomly choose training set $S \subset X$ of examples
Target user t rates each object in S
Examples in S and their ratings presented to the crowd
Workers $w \in W$ predict t's ratings on $X \setminus S$
Results are aggregated to project ratings

Figure 2: Taste-grokking protocol

Approach

We explore two different ways to personalize human computation. In the first, which we call *taste-matching*, workers who are subjectively similar to the requester are identified and asked to provide annotations. This approach functions like collaborative filtering: people with similar opinions in a domain can be expected to align on unknown or future opinions. In the second, which we call *taste-grokking*, workers are provided with examples of the requester's taste and asked to infer how the requester might annotate other items. Workers are not required to be similar to predict the subjective tastes of a requester, as a human worker with very different tastes may still be able to infer the requester's needs if sufficiently well communicated.

For simplicity, we assume that the goal is to provide personalized projected ratings on a given rating scale of a large set X of objects. However, although the protocols are described for rating, similar ideas may be used in more complex human computation protocols.

The taste-matching protocol (Figure 1) profiles requesters by asking them to annotate a number of training examples. Workers are profiled using the same examples. Similarity between the worker profiles and the requesters' is then calculated, allowing us to determine which workers are the most appropriate personalized workers for the requester. The number of training examples is dependent on the task.

The taste-grokking protocol (Figure 2) converts the task into one with a presumed ground truth, allowing us to use existing reliability metrics. For example, a held-out set of training examples may be used to evaluate worker quality.

Preliminary Experiments

We explored taste-matching and taste-grokking using Mechanical Turk to annotate 100 images of salt & pepper shakers from Amazon.com. Thirty workers rated the 100 images on a scale of 1-5 stars. Then, by using a subset of the ratings from one worker as a "requester", we evaluated the performance of our two different approaches in predicting that requesters other ratings.

Baseline	1.59
Taste-matching	1.10
Taste-grokking	1.07

Table 1: RMSE of target user predictions

To develop a baseline, we selected ten random examples for training, and used the remaining 90 for testing. We then calculated the average root-mean-squared error (RMSE) of predictions over the test examples for each worker in the pool as a predictor of each other worker. For taste-matching, the top quartile of users on the training set were used as the well-matched worker. Their ratings were used to predict on the test examples for each user in the pool. For taste-grokking, ten workers were chosen, shown the target's ratings on the ten examples and their predictions were averaged on each of the 90 test examples.

The results from these experiments are shown in Table 1. Both personalized approaches improved in quality over a baseline where neither was applied. For these results, taste-grokking with predictions from an aggregation of multiple workers appears to work slightly better with less variance, and we hope to see if this holds true across different tasks.

While workers reported enjoying both tasks, preliminary feedback suggests that the taste-grokking was generally more enjoyable. Early work suggests that there is an effect based upon which examples are used to teach the taste.

Next Steps

By examining taste-matching and taste-grokking across multiple domains, we aim to see if one approach outperforms the other in general, or if different approaches are optimal for different domains. We also plan to apply taste-matching and taste-grokking to more complex and creative tasks. Finally, we are exploring various parameters that affect the protocols in different contexts, such as the optimizing the choice of training examples, choosing sample sizes, and balancing cost and quality improvements.

References

Bernstein, M. S.; Brandt, J.; Miller, R. C.; and Karger, D.R. 2011. Crowds in two seconds: enabling realtime crowd-powered interfaces. In *Proceedings of UIST 2011*, 33-42.

Maltz, D.; and Ehrlich, K. 1995. Pointing the way: active collaborative filtering. In *Proceedings of CHI 1995*, 202-209.

Teevan, J.; Dumais, S.T.; and Horvitz, E. 2010. Potential for personalization. *TOCHI*, 17(1).

Zhang, H.; Law, E; Miller, R.C; Gajos, K.; Parkes, D.; and Horvitz, E. 2012. Human computation tasks with global constraints. In *Proceedings of CHI 2012*, 217-226.