Proceedings of the Second AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2014)

A Crowd of Your Own: Crowdsourcing for On-Demand Personalization

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

[1]University of Illinois at Urbana-Champaign [2]Microsoft Research [3]MIT CSAIL

organis2@illinois.edu {teevan, sdumais, adam.kalai}@microsoft.com rcm@mit.edu

Abstract

Personalization is a way for computers to support people's diverse interests and needs by providing content tailored to the individual. While strides have been made in algorithmic approaches to personalization, most require access to a significant amount of data. However, even when data is limited online crowds can be used to infer an individual's personal preferences. Aided by the diversity of tastes among online crowds and their ability to understand others, we show that crowdsourcing is an effective on-demand tool for personalization. Unlike typical crowdsourcing approaches that seek a ground truth, we present and evaluate two crowdsourcing approaches designed to capture personal preferences. The first, *taste-matching*, identifies workers with similar taste to the requester and uses their taste to infer the requester's taste. The second, *taste-grokking*, asks workers to explicitly predict the requester's taste based on training examples. These techniques are evaluated on two subjective tasks, personalized image recommendation and tailored textual summaries. Taste-matching and taste-grokking both show improvement over the use of generic workers, and have different benefits and drawbacks depending on the complexity of the task and the variability of the taste space.

Introduction

Many information-seeking scenarios, such as search and recommendations, involve subjective needs where there is no clear notion of correctness. Instead, the proper items to present to a user depend on the specific user's preferences. Although crowdsourcing is commonly used for ground truth tasks, in this paper we show that it can also be used as a tool to aid subjective information-seeking by implementing two approaches for personalization that make use of paid crowdsourcing. While current approaches to personalization require large amounts of preference data from other users, the techniques we develop are valuable for personalization in small systems and over personal collections, where no data exists.

Requester	★	★★★★	★★	?
Taste-Matching:				
Worker I	★	★★★	★	★★★
Worker II	★★★	★	★★★★	★
Taste-Grokking:				
Worker III	*Sees ratings from requester*			★★★★

Table 1: Two techniques to predict ratings of a requester. In taste-matching, workers rate according to their own preferences; Worker I is most similar to the requestor and thus influences the prediction. In taste-grokking, workers see the requester's prior ratings and make predictions.

Because personal preferences are hard to capture in isolation, personalized content is often generated using prior data from other people. For example, product selection behavior can be passively observed across many users and then used to recommend particular products to individuals based on the behavior of related users. Identifying and using data from related users can aid a successful personalized experience, but the amount of data required to do so limits its applicability to large-scale systems such as search engines, heavily-used content providers, and major online retailers that have abundant interaction data. A system intended to recommend content from within a personal collection lacks the necessary data on which to base its recommendations.

This paper shows how paid crowdsourcing can fill in for a lack of prior information in on-demand personalized content. In the approach we call *taste-matching*, we show that crowd workers with similar tastes can be identified and utilized for personalization. For example, to receive a personalized rating for the rightmost salt shaker in Table 1, the requester first provides ratings for other salt shakers. After collecting ratings for these same salt shakers from crowd workers, it is possible to match the requester's ratings with Worker I and predict the requester is very likely to enjoy the unrated salt shaker. Alternatively, in an approach we call *taste-grokking*, crowd workers are asked to guess the personal preferences of a requester after seeing that requester's responses on a set of feedback questions. In the example shown in Table 1, Worker III sees the requester's earlier ratings and offers an educated guess that the requester will like the last salt shaker.

Both taste-matching and taste-grokking are human computation tasks that derive their benefits from a core feature of paid crowdsourcing – accessible crowd workers with many different tastes. Both also scale linearly to estimate many personalized suggestions from a small number of requester contributions. We evaluate their tractability in two contexts: 1) for making personalized image recommendations, and 2) for highlighting important text in a short article. Personalized recommendations allow us to explore the approaches in a well-studied domain, while text highlighting looks at a more complex use of crowdsourced personalization where the inclusion of human intelligence is particularly valuable.

There are many crowdsourcing problems for which the desired outcome relies on the subjective needs of a requester. Our approach could benefit tasks run over personal archives (e.g., identifying the best photographs from a large personal collection), datasets where many items are unique (e.g., shopping advice for handmade artwork), and dynamic datasets (e.g., finding the perfect apartment). New systems that do not yet have large-scale usage data could also use our approach to bootstrap the ability to provide recommendations. By framing the problem and evaluating multiple approaches and cases, this paper provides guidance as to how crowd markets can be used to support personalization.

Related Work

In volunteer settings, subjective contributions are common, in tasks that do not enforce a notion of a right or wrong answer. For example, many systems collect preference ratings on information objects. Sometimes these are used as a universal measure of quality (e.g., Facebook 'likes'), while other times they contribute to collaborative filtering algorithms. Other systems relish in the diversity of the crowd. The eclectic designs of crowd creation are highlighted in projects such as *Star Wars Uncut* (Pugh 2009) and *The Johnny Cash Project*. These artistic works, assembled from fragments created by thousands of contributors, intentionally punctuate the fact that humans have wildly variable tastes and styles.

In contrast, paid crowdsourcing is commonly used for human computation tasks, which "fit the general paradigm of computation" (Quinn and Bederson 2011). Many such uses make an objective assumption, seeking a consensus answer from the crowd. In systems such as the ESP Game (Ahn and Dabbish 2004), users have the explicit goal of agreeing with other people's opinions.

However, an increasing number of crowdsourcing projects are pursuing tasks that can be argued to have a subjective aspect to them, even when distinctions between users is not the focus on the system. For example, image tagging tasks (e.g., "smiling" or "beautiful") may be subjective; choosing the "best" single frame to represent a video (Bernstein et al. 2011) is partly a matter of taste; similarity ratings between objects have been shown to vary across users (Tamuz et al. 2011); and document editing (Bernstein et al. 2010) requires making stylistic decisions. Recent work has discussed problems inherent to pursuing consensus for tasks that are fundamentally subjective (Alonso et al. 2013).

We observe a class of problems where the task is time-consuming for an individual, but its subjective nature makes it difficult to delegate. Such tasks have a time cost that not all people find worthwhile. The actual value – a factor of enjoyment and wealth (Marmorstein et al 1992) – varies between individuals, and this trade-off has been noted in areas such as comparison shopping, travel planning, and job searching (ibid).

Crowdsourcing is well-positioned to address these types of problems. Some efforts have applied crowds to personalization for an individual, usually in the context of a specific application. Both EmailValet (Kokkalis et al. 2012) and Mobi (Zhang et al. 2012) allow workers to describe their preferences (for email and travel itineraries, respectively) in natural language. In our study, we use profiling tasks rather than natural language descriptions as a method for understanding the user. While paid crowdsourcing has been pursued for subjective tasks in the past, we are not aware of a generalized treatment of crowdsourcing for use in this way.

The two approaches we explore to crowdsourced personalization have precedents in the personalization literature. The taste-matching approach we study builds on collaborative filtering (e.g., Hofmann 2004) to identify crowd workers with related preferences and use them for personalization. In collaborative filtering, personalized item recommendations are typically provided by matching a target user to similar users, and inferring a predicted preference based on ratings from those similar users. However, our workers contribute data on request, thus mitigating a primary problem of collaborative filtering—sparse data, especially for unpopular or new items.

The taste-grokking approach we study seeks to convey the tastes of requestors by mean of examples. Krishnan et al. (2008) evaluated the feasibility of a similar approach in the context of film recommendations. They found that the MovieLens collaborative filtering system performed better on average than humans, but in some cases, such as for unusual or eclectic requester profiles, humans were more effective. Our research adopts this concept as a method for using paid crowds to personalize. However, unlike previous work, we assume a setting over different domains with no prior information available.

In summary, our study recognizes the effectiveness of crowdsourced services when there is a lot of data (e.g. Netflix) or when the need can be boiled down to quantitative filters, and shows that crowdsourcing can also fill the need in the more difficult cases: subjective and without data. We develop methods to support on-demand personalization, such as over personal data. The research presented in this paper differs from earlier work in that it explicitly adopts an assumption of subjectivity in the requester-worker form of crowdsourcing, where most prior work seeks to correct for it. We formalize two approaches to collecting subjective data: taste-matching and taste-grokking. In these approaches, we give researchers a new framework for thinking about personalized crowdsourcing.

Approach

We explore crowdsourcing for on-demand personalization by studying two different approaches: *taste-matching* (where crowd workers with similar tastes are identified and used for personalization) and *taste-grokking* (where crowd workers are asked to guess the personal preferences of a requester after seeing that requester's responses on a set of feedback questions). We refer to the user receiving personalized content as a *requester*. The contributors of data for the requester's needs are referred to as *workers*.

Profile Construction

Personalizing content for a given user requires a profile of that user's tastes. A requester's profile is constructed in the same manner for both taste-matching and taste-grokking. A subset of the work items to be completed in a personalized manner are selected, and performed by the requester. For example, Table 1 shows the requester profile for an image recommendation task. Three of the four pictures of salt shakers were selected for rating, and the requester gave them one, four, and two stars respectively.

The work items selected to construct a requester's profile can influence the quality of the personalization. The items selected must represent the range of items in the set to fully capture the requester's preferences. For example, if the requester in Figure 1 were only asked to rate the two silver-colored salt shakers, it would be difficult to infer that they really like figurine salt shakers. In our experiments, we select profiling task items at random, though other selection strategies can be employed. The more items that are selected the richer the resulting profile, but the more work that must be done by the requester.

Taste-matching and taste-grokking differ in how they use the profiling data in conjunction with crowd workers to provide personalized recommendations to the requester.

1. Choose profiling set $S \subset X$ of examples
2. Requester t performs work on each object in S
3. Workers $w \in W$ perform work on S
4. Worker pool $W' \subset W$ is filtered by similarity to t
5. For each subsequent task, workers $w \in W'$ perform work in their own taste on remaining data $X \setminus S$

Figure 1: The taste-matching approach

1. Choose profiling set $S \subset X$ of examples
2. Requester t performs work on each object in S
3. Work by t on S is shown to workers $w \in W$
4. Workers $w \in W$ predict t's opinions on $X \setminus S$
5. Optional 'wisdom of crowds' quality control (e.g., aggregation)

Figure 2: The taste-grokking approach

Taste-Matching

Taste-matching uses the requester's profile to identify a group of workers that are similar to the requester in how they perform a task, and uses the output from these workers on other tasks to infer the requester's output. This process is outlined in Figure 1.

Workers are initially profiled against the profiling set that was used for requesters. Similarity between the worker profiles and the requesters is calculated to determine which workers are the most appropriate personalized workers for the given requester. The similarity measure used should be appropriate to the task. For example, in the salt shaker case we use the root-mean-squared error to measure the deviation of a worker's preferences from the requester's true opinions. The most similar workers are then asked to provide additional contributions to use for personalization.

Taste-matching assumes that people with similar opinions and tastes in a domain will agree on unknown opinions. This assumption is also made in collaborative filtering. Collaborative filtering systems like Netflix use opinion data that has been previously entered on various subsets of the collection. Taste-matching can be likened to collaborative filtering, except the data is collected on-demand and explicitly. It adopts the technique of using similar people as proxies for a requestor to situations where one does not necessarily have access to preexisting ratings by other people. On-demand ratings can be used to enable new personalization scenarios. For example, a requester may have a large set of personal images that they want the crowd's help in filtering. In such a case, collaborative filtering is impossible, however taste-matching is applicable.

Taste-Grokking

In contrast to taste-matching, taste-grokking (Figure 2) does not assume that workers need to be similar to the requester in order to understand them. Instead, after being communicated the requester's tastes on some items, workers are asked to infer the requester's taste on other items. The requester's profile is shown directly to workers as an example of the requester's needs. Rather than performing work according to their own tastes, workers are asked to complete tasks based on the inferred needs of the requester; that is, they try to make sense of the requester's opinions ('grok') and contribute how they think the requester would. Results across all of the workers are aggregated to make the final prediction.

When online workers are grokking a requester, we do not attempt any form of matching. In prior work, Krishnan et al. (2008) found that better matched workers do not seem to perform better for grokking-like tasks.

Having requesters provide concrete examples of their tastes provides a simple way to communicate their needs to workers without the need to communicate detailed preferences. There are other ways personal preference might be communicated. For example, a user could describe their preferences directly, as is done with Mobi (Zhang et al. 2012). Research in search personalization, however, has found that examples of a user's need can often implicitly communicate the desired outcome better than an explicit description (Teevan et al. 2008).

Personalized Item Recommendation

We study taste-matching and taste-grokking in two contexts: 1) personalized image-based item recommendations, and 2) highlighting important text in a short article. We begin by looking at personalized item recommendations because there are many existing algorithmic approaches to this task. While the recommendation task has been well explored on popular datasets like movies, our analysis focuses on predicting a requester's opinion on images of salt shakers and local cuisine. Traditional approaches are unlikely to perform well on such datasets because there is limited existing preference data for the items they contain.

Methodology

Dataset

We collected 100 salt and pepper shaker images from Amazon's online store, and 100 popular photographs and accompanying names of restaurant meals in the cities of Boston and Seattle from Foodspotting.com. All of these images represent niche or less common spaces that attract discriminating tastes. By determining what a requester would like within the space, online crowds can help a requester identify a manageable subset of the space.

Profile Construction

Using Mechanical Turk, we asked 30 workers to rate all of the images from each image dataset on a five-point scale. For the salt shaker task, workers were shown images of a salt shaker and asked how much they liked the product. For the food task, workers were shown an image of a restaurant meal along with the name of the dish, and asked to rate how appetizing each photographed dish looked. Workers were paid $1.50 for rating a set of 100 images.

From each image set, a requester profile was constructed for each worker by randomly selecting a subset of images and their associated ratings. In the case of taste-matching, 20 images were selected. For taste-grokking, only 10 were selected, since we expected 20 images to be too many for a worker to be able to understand well. Prediction performance was evaluated by looking at how accurately taste-matching and taste-grokking predicted the worker's ratings for the remaining images.

Taste-Matching

Matching workers were identified using root-mean-squared error (RMSE), RMSE measures the deviations of the predicted preferences from the requester's true opinions and is in the same units as the ratings:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(r_{i,\text{predicted}} - r_{i,\text{true}}\right)^2}.$$

Lower error indicates a better match.

People with identical opinions can have different interpretations of what the points of a rating scale mean. To make ratings comparable across workers, ratings in taste-matching were normalized as deviations from each user's mean rating (Hofmann 2004). This was not necessary for taste-grokking, because the intention is for the workers to understand the requester's worldview (e.g., what a three or four star rating means) rather than apply their own.

Because both requesters and workers contribute their personal opinions in the ratings for taste-matching, it was possible to simulate requesters by using each worker as a pseudo-requester. Thus, we evaluate recommendations for 30 requesters.

Taste-Grokking

Prediction results in taste-grokking are requester specific, so new data needs to be collected for each requester. For this reason, each of the three tasks was run with four requesters. For each requester-task pair, two different profiling sets were evaluated, each with 30 workers performing the grokking.

In taste-grokking, the requester profiles were shown directly to workers, who then were asked to guess the requester's opinion on the 90 unseen items. For the images of salt shakers they were asked how much they expected the requester to like the salt shaker, and for cuisine they were asked how appetizing they expected the image to be to the requester. In exchange for completing the task, workers were paid a base of $1.00 plus a reward of up to $1.00 based on the accuracy of their guesses.

		Products –Salt shakers	Food - Boston	Food - Seattle
Baseline: Prediction by any worker		1.64	1.51	1.58
Taste-matching: Best 3 workers overall		0.89 (-46%)	1.02 (-32%)	1.19 (-25%)
Taste-grokking: Best 3 workers overall		0.87 (-47%)	0.78 (-48%)	0.79 (-50%)
Taste-matching	Best worker from random 5	1.43 (-13%)	1.19 (-22%)	1.26 (-20%)
	Best worker from random 10	1.35 (-18%)	1.08 (-29%)	1.08 (-31%)
Taste-grokking	Average individual	1.29 (-21%)	1.53 (+1.3%)	1.57 (-0.5%)
	Aggregated prediction (5 random workers)	1.07 (-34%)	1.38 (-9%)	1.28 (-19%)
	Aggregated prediction (5 top workers)	1.02 (-34%)	1.22 (-19%)	1.13 (-28%)

Table 2: Average RMSE of workers predicting user preferences for rating tasks. A lower RMSE value indicates better performance. Number of workers per condition: 30.

Whereas in taste-matching workers contribute their own relative truths, taste-grokking workers are all striving toward a common truth: the needs of the requester that they are interpreting. As such, it is possible to aggregate multiple predictions into a "grokking consensus". In aggregation, the mean of multiple workers' predictions is taken for a final prediction. We found that the mean was more effective for aggregation than the median.

Measurement

To evaluate the two approaches, we use RMSE to compare the ratings predicted by each approach with the requesters' true ratings. A lower RMSE represents less error and thus better predictions. As a baseline measure, we consider the predictive power of a random worker's ratings, without any taste-matching or taste-grokking. This is a measure of prediction without personalization: the baseline accounts for the non-user specific agreement (e.g., food that looks appetizing to most people).

As shown in Table 2, the baselines have RMSE values of 1.64 (salt shakers), 1.51 (Boston food), and 1.58 (Seattle food). People's tastes were more variable for the salt shaker task, with an average standard deviation of 1.32 stars, compared to 1.04 (Boston) and 1.11 (Seattle). As an alternative baseline, we examined an aggregated average of three random workers, representing more of the global consensus. This measure gives RMSE values of 1.51 (salt shakers), 1.38 (Boston food), and 1.30 (Seattle food).

Results

Taste-Matching

We begin by examining the core assumption of matching, that workers who are strongly-matched on a profiling set will be good at predicting other items as well. We found that relationship to be strong: workers with low or high RMSE on the profiling data showed similar performance on the evaluation data. For the salt shaker task, this correlation (Pearson) is 0.73; for the restaurant dish task, it is 0.67 (Seattle task) and 0.71 (Boston task).

We now turn to predicting rating performance. Table 2 shows the average performance of the three best workers (out of 30) in the collected data per condition. The three best workers (aggregated) perform 46% better than the baseline for the salt shaker task and 25-33% better for the food tasks. These numbers represent the best workers that were present in the results, but these workers may not be easy to identify a priori. Nonetheless, it suggests that the best-matched workers have a high theoretical limit on performance, higher than we will see for grokkers, if they could be reliably identified. With no prior information on workers, we need to profile workers as they arrive. For a setting more likely to be seen in practice, a requester can post a task to a crowdsourcing marketplace, wait for the first few contributors, and take the contribution of the best worker. Table 2 shows the test data performance of the best matched worker from a smaller set of five or ten random workers. Waiting for more workers before deciding on the best-match will continue to improve quality as shown in Figure 3. Although there is no formal expectation of monotonic improvement—since 'best worker' is chosen from training data and their performance shown against the test data is shown—the RMSE nonetheless decreases nearly consistently.

Taste-Grokking

Without aggregation, the average performance of any single worker's grokking prediction is shown in Table 2. In the salt shaker task this showed a 21.3% improvement over the baseline, while the food recommendation tasks showed -1.3% (Boston) and 0.5% (Seattle) improvements. In other words, without controls for worker quality, the performance on the food tasks shows that it is risky to trust a single worker's contributions.

Aggregating contributions from multiple workers into a single predicted rating set is a more effective approach, smoothing over low-quality workers or the occasional error by a good worker. Table 2 shows the performance of five-worker aggregated predictions – a realistic parameter recommended by Novotney and Callison-Burch (2009). The number of workers to aggregate is dependent on multiple factors, such as cost-quality trade-off, but Figure 4 (left) shows that the quality of averaging random workers will keep increasing, with the majority of improvements in the first 10 people. When controlling for quality by cross-validating the grokking proficiency of workers on a per-task basis, Figure 4 (right) shows that a small number of averaged workers perform best in our tasks. Both aggregation and worker quality cross-checking are common quality controls in objective tasks, and they appear similarly effective for taste-grokking for subjective tasks.

The set of items used for profiling was selected through randomization. However, random selection of training examples can potentially fail in domains where there are many dimensions that can affect a person's opinion. To measure the effect of randomization, we set up alternate taste-grokking evaluation that used an optimized rather than random set of items. K-means clustering was employed to determine taste groups, where the number of clusters k was equal to the number of profiling items n (i.e. *k=10*). An example of these clusters is seen in Figure 6. Profiling examples were sampled from these clusters, one per group. This was, in essence, a stratified random sample.

The optimized training examples (long-dash line in Figure 5) greatly improved the performance of taste-grokking. At its best, it provides an RMSE of 0.88, a 44.9% improvement over the baseline.

Summary

Personalization in the form of rating prediction is a strong fit for taste-matching and taste-grokking. Both approaches offer improvements over the baseline. The different task types benefitted from each approach differently, however: the restaurant meal recommendation task gained more from taste-matching, while the product rating task showed better recommendations from taste-grokking.

Figures 3 and 4 show how performance of these techniques changes with differently sized crowds. The value of taste-matching is in larger numbers of contributions. With a very small number of workers, the best-matched worker is not particularly strong. However, taste-matching benefits from consistent improvements from profiling additional workers. Also, once well matched workers are found, they do not have to rate the same items.

In taste-grokking, aggregation of multiple workers' predictions of a requester's rating improves quality notably. Bad workers can be further controlled for by using a held-out set to measure their reliability (Figure 4, right), but even aggregating the first random workers that submit contributions (Figure 4, left) is an effective strategy.

We also examined the effect of training examples on performance, finding improved performance with an optimized set of examples used for taste-grokking. A difficulty with this approach in practice is that the information available for determining the best examples

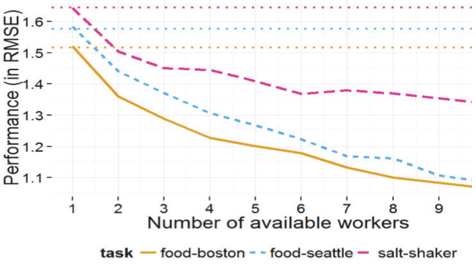


Figure 3: Performance of the best-matched workers in taste-matching from different sized pools of workers.

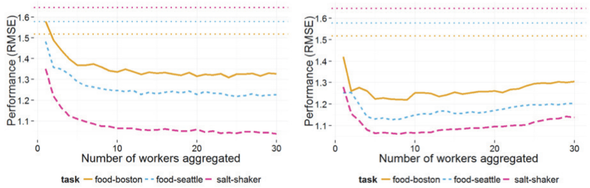


Figure 4: Performance of taste-grokking predictions aggregated from random workers (left) and the best workers (right). Shown for different sized pools of workers, dotted line denotes baseline.

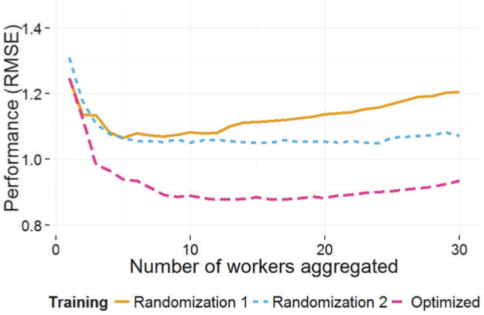


Figure 5: Effect of different taste-grokking profiling items on performance of top k aggregation. Long-dash line represents optimized profiling set

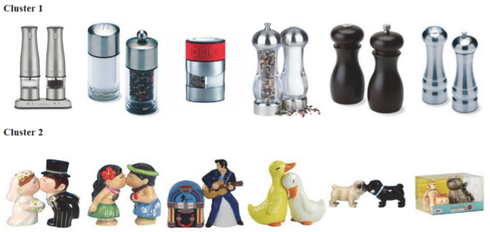


Figure 6: Example of clusters used for stratified training example sampling

may not be available. Still, these results suggest that clustering on other known variables, such as labels or facets, is worthy of future study. Most basically, a human sanity check could be applied to randomized training examples: workers or the requester could first look at the tasks that have been completed for training and assess whether they appear adequate.

Personalized Text Highlighting

We also looked at personalization for more complex tasks than simple image ratings. Taste-matching and taste-grokking were applied to a task where workers highlighted the key points of a film review. Subsequently, new highlighted texts are created for requesters, either from reviews highlighted by well-matched workers with taste-matching, or reviews that taste-grokking workers prepare for the requester.

Highlighting a text is a potentially useful task to personalize, as different types of information can appeal to different people. Many professions deal with large amounts of text that are difficult to parse. For our experiments we examined the task of highlighting key points in a film review, as shown in Figure 7. The highlights serve the purpose of summarization: the goal of the highlights being to make the text usefully skimmable for a reader.

Methodology

Dataset

The texts used for the highlighting task were six film reviews by professional critics at *The A.V. Club.* We selected film reviews because people often have strong and distinct opinions about movies. The reviews averaged 456 words each.

Profile Construction

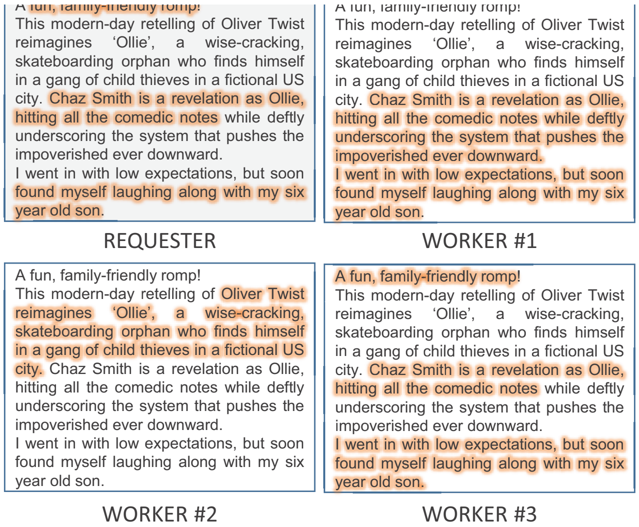
We asked 50 Mechanical Turk workers to highlight each of the six film reviews.

Workers performed the highlights using a custom interface on Amazon's Mechanical Turk. Workers were asked, "if somebody gave you a summary of this review, what would you like to know to help you decide if it is movie worth seeing?" We did not impose any constraints on what workers highlighted. Their tastes varied not only on what was important, but also how much was important. On average, they highlighted 26.6% of reviews, though a 0.95 confidence interval ranged from 3.3% to 50%.

Taste-matching and taste-grokking were used for recommending highlights to requesters based on how they highlighted one randomly selected profiling examples.

Taste-Matching

As in our image rating recommendation, both requesters and workers contribute their personal opinions in the highlights they provided for taste-matching. Thus it was possible to simulate requesters by using each worker as a pseudo-requester. We evaluate recommendations for 50 requesters.



A fun, family-friendly romp!
This modern-day retelling of Oliver Twist reimagines 'Ollie', a wise-cracking, skateboarding orphan who finds himself in a gang of child thieves in a fictional US city. Chaz Smith is a revelation as Ollie, hitting all the comedic notes while deftly underscoring the system that pushes the impoverished ever downward.
I went in with low expectations, but soon found myself laughing along with my six year old son.

REQUESTER

A fun, family-friendly romp!
This modern-day retelling of Oliver Twist reimagines 'Ollie', a wise-cracking, skateboarding orphan who finds himself in a gang of child thieves in a fictional US city. Chaz Smith is a revelation as Ollie, hitting all the comedic notes while deftly underscoring the system that pushes the impoverished ever downward.
I went in with low expectations, but soon found myself laughing along with my six year old son.

WORKER #1

A fun, family-friendly romp!
This modern-day retelling of Oliver Twist reimagines 'Ollie', a wise-cracking, skateboarding orphan who finds himself in a gang of child thieves in a fictional US city. Chaz Smith is a revelation as Ollie, hitting all the comedic notes while deftly underscoring the system that pushes the impoverished ever downward.
I went in with low expectations, but soon found myself laughing along with my six year old son.

WORKER #2

A fun, family-friendly romp!
This modern-day retelling of Oliver Twist reimagines 'Ollie', a wise-cracking, skateboarding orphan who finds himself in a gang of child thieves in a fictional US city. Chaz Smith is a revelation as Ollie, hitting all the comedic notes while deftly underscoring the system that pushes the impoverished ever downward.
I went in with low expectations, but soon found myself laughing along with my six year old son.

WORKER #3

Figure 7. An example of a highlighted movie review

Workers were matched to requesters using by looking at textual overlap on the basis of a one-review profiling set. The remaining five reviews were used wholesale as recommendations.

To identify workers that highlight the same text as requesters, while also highlighting similar amounts of texts, we use the F1 score:

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Precision is the proportion of words highlighted by the worker that are in the requestor's highlights, and *Recall* is the proportion of words highlighted by the requestor that the worker also highlighted. F1 balances precision and recall. The F1 score ranges from 0 to 1, with a higher score indicating a better match. In the example shown in Figure 7, Worker #3 has the highest F1 score for the requester, matching closely on what they highlight, while not highlighting much of what the requester is uninterested in.

Taste-Grokking

Taste-grokking workers were given a randomly selected profile example and asked to highlight what they thought the requester would find interesting based on what the requester found interesting in a training example. Workers were shown only one highlighted review, from which they inferred the style of the requester. The choice of $n = 1$ for profiling was motivated by the expected information communicated in an example.

Results for taste-grokking were evaluated for three randomly selected requesters. While the profiling set was only one review, we used two different profiling texts to communicate their tastes. For each of the six combinations, four film reviews were highlighted through taste-grokking, by 30 different workers.

Measures

The quality of a highlighting recommendation was measured using F1 to measure the text overlap between the predicted personalized highlighting and the requester's actual highlighting.

As a baseline we calculated the mean F1 score that would be achieved without matching or grokking: that is, by taking any worker's highlights as a recommendation. The baseline achieved a F1 score of 0.32.

Results

Taste-Matching

Taste-matching workers who had high F1 scores on the profiling text also performed well for the matching text. The best profiled workers showed a 20% improvement over the baseline, while widening the net to the five best profiled workers still showed 17% improvement (Table 3).

Taste-Grokking

In contrast, most taste-grokking workers did not perform well. As seen in Table 3, the average worker appeared to find it difficult to accurately grasp the nuances of a requester's needs. The average worker's highlights had an F1 score of 0.30, which was -0.7% below the baseline. However, some workers proved to be adept at the activity. The best grokking workers averaged an F1 score of 0.73, while the best worker from a field of five random ones averaged an F1 of 0.52. Determining the best workers a priori is difficult as their relative rankings did not stay consistent between a held-out set and the test set.

Summary

The review summarization task had workers highlighting key points of film texts. As expected, this longer form-task was more unwieldy for workers to perform. Taste-matching gave reliable quality improvements from well-matched workers. Taste-grokking was less predictable; while the best workers performed very well, many workers had difficulty predicting what types of information to grok from the requester's example. As a result, the average taste-grokking worker's performance was comparable to the baseline.

Discussion

When requesters want personalized results over custom datasets, does paid crowdsourcing offer a reasonable solution? Through taste-matching and taste-grokking, we found it to be an effective approach to varying degrees, dependent on the task and the dimensions of the data. These differences offer insights into how future cases of unique data can be personalized through crowdsourcing.

Factors in Choosing a Crowdsourcing Approach

Evaluating two different approaches for crowdsourcing allowed us to observe how different task considerations factor into the better approach for a practitioner hoping to use personalized crowdsourcing. Both taste-matching and taste-grokking offer improvements in tasks where there is an element of subjectivity. Grokking works better for tasks where there were fewer variables to communicate. For tasks like the text-highlighting task, which involved more nuanced judgments of a worker's style and the context of each highlighted section, taste-matching is a simpler approach.

Quality Control. In traditional tasks, the primary factor affecting worker reliability is objective quality. It is important to be robust to workers that are inattentive, sloppy, or malicious. In the case of subjective tasks, the question of "is this a good worker" is confounded with the varying subjective frames of the workers. Taste-matching and taste-grokking have different mechanisms for balancing the quality of a worker and a requester's subjective needs.

Using taste-grokking for a task gives it a ground truth: workers are pursuing the same goal. This allows the task to be treated like an objective task, cross-validating workers to measure their grokking proficiency. This situation has more accountability, discouraging dishonest or inattentive labor.

With taste-matching, the quality of a worker is negated by the fact that a sloppy worker is less likely to match up with a requester. Also, while in some normative tasks eclectic workers can be misidentified as cheaters, taste-matching makes no such assumption: such a worker might be a perfect match for a requester with unusual tastes.

Worker Satisfaction. Taste-grokking was more commonly cited in worker free-text feedback forms as an enjoyable task. This may be because the task design positioned it as a guessing challenge, where bonuses were paid for better work. However, when grokking failed by not communicating enough information in the training examples, workers also expressed distress.

Task Complexity. Whether the factors affecting subjective opinion-forming are easy to make sense of, or grok, should be factored in deciding on the appropriate approach. This issue was observed in the contrasting results for recommendation: using taste-matching on food images gave better performance than for product images, while the opposite held true for taste-grokking.

When workers were asked to explain opinions on salt shakers, they predominantly referenced visual style, a manifest taste-affecting factor. In contrast, food taste appears to have more latent, unseen factors affecting taste. We applied k-means clustering to worker tastes, finding that

		Performance (F_1)
Baseline		0.32
Taste-matching	Best-matched workers	0.39 (+20%)
	5 best-matched workers	0.38 (+17%)
Taste-grokking	Any worker	0.30 (-7%)
	Best workers (pool of 5)	0.52 (+62%)

Table 3: Performance of workers on review summarization task. Higher values indicates improvement.

food taste clusters were not as qualitatively coherent as the salt shaker clusters. In other words, a grokker might not realize the overlap between, say, beer lovers and shawarma fans.

On-Demand Collaborative Filtering

Taste-matching can be interpreted as on-demand collaborative filtering (CF) for personalized contexts. It allows one to apply intuitions from CF to areas where data does not exist; e.g. one's personal photos.

A potential benefit of matching is that it does not need to be restricted to personalization for a single person. Data collected through taste-matching is reusable for other requesters, unlike taste-grokking. In our use case – of an individual collecting personalized recommendations for their own need – the data collection needs would be comparable between matching and grokking. However, generalizing to more workers is much more tractable for matching. We also observed a higher ceiling with matching, so continuing to collect more data for more requesters keeps improving the quality of predictions, similar to mature systems like Netflix.

More generally, the application of paid crowdsourcing for bootstrapping the data of a CF system is worth exploring. Using crowd workers akin to taste-matching can address data sparsity issues related to new systems, users, or items.

Conclusion

While accommodating the diversity of people's opinions is important for subjective tasks, personalization over unique or personal collections is impeded by access to prior data. We show that in such cases, online crowd markets can address this problem. Personalized crowdsourcing affords the ability for a requesting user to get specialized work specific to their tastes, on-demand and over novel content.

Although existing crowdsourcing approaches tend to seek an objective truth, we demonstrate that making an assumption of requester-specific subjectivity is a good fit with the diversity of the crowd. We presented two task design protocols that make this assumption, *taste-matching* and *taste-grokking*, and showed that they improve over an approach that did not assume a subjective requester. Taste-matching is shown to be effective at scaling to large numbers of workers and for tasks where there are many latent variable affecting one's taste. Taste-grokking works well in areas where one's tastes are easy to communicate and was effective with a small number of workers.

Our results show the promise of crowdsourcing for personalization in domains where a lack of prior data limits the efficacy of established personalization methods.

References

Ahn, L.; and Dabbish, L. 2004. Labeling Images with a Computer Game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 319–26. Vienna, Austria: CHI' 04.

Alonso, O.; Marshall, C.C.; Najork, M. 2013. Are Some Tweets More Interesting Than Others? #HardQuestion. In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval*, 2:1–2:10. New York, NY: HCIR '13.

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 the 24th Annual ACM Symposium on User Interface Software and Technology*, 33–42. New York, NY.

Bernstein, M.S.; Little G.; Miller R.C.; Hartmann B.; Ackerman, M.S.; Karger, D.R.; Crowell, D.; and Panovich, K. 2010. Soylent: A Word Processor with a Crowd Inside. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Tech.*, New York: UIST '10.

Bernstein, M.S.; Tan, D.; Czerwinski, M.S.; and Horvitz, E. 2008. Personalization via Friendsourcing. *ACM Transactions on. Computer-Human Interactions.* 17 (2): 6:1–6:28.

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

Kokkalis, N; Köhn, T; Pfeiffer C.; Chornyi, D.; Bernstein, M.S.; and Klemmer, S.R. 2013. EmailValet: Managing Email Overload through Private, Accountable Crowdsourcing. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. New York, NY: CSCW '13.

Krishnan, V.; Narayanashetty, P.D.; Nathan, M.; Davies, R.T., and Konstan, J.A. 2008. Who Predicts Better?: Results from an Online Study Comparing Humans and an Online Recommender System. In *Proceedings of the 2008 ACM Conference on Recommender Systems*. New York, NY: RecSys '08.

Law, E.; and Ahn L. 2011. Human Computation. In *Synthesis Lectures on Artificial Intelligence and Machine Learning* 5 (3).

Marmorstein, H.; Grewal D.; and Fishe, R.P.H. 1992. The Value of Time Spent in Price-Comparison Shopping. *Journal of Consumer Research* 19.1:52-61. University of Chicago Press.

Novotney, S.; and Callison-Burch, C. 2010. Cheap, Fast and Good Enough: Automatic Speech Recognition with Non-Expert Transcription. Annual Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA: HLT '10.

Pugh, C. 2009. Star Wars Uncut: Director's Cut. www.starwarsuncut.com.

Quinn, A.J.; and Bederson, B.B. 2011. Human Computation: A Survey and Taxonomy of a Growing Field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY: CHI '11.

Tamuz, O; Liu C.; Belongie, S.; Shamir, O; and Kalai, A.T. 2011. Adaptively Learning the Crowd Kernel. In *Proceedings of the International Conference on Machine Learning*. ICML '11.

Teevan, J.; Dumais, S.T.; and Liebling, D.J. 2008. To Personalize or Not to Personalize. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 163. New York: SIGIR '08.

The Johnny Cash Project. http://www.thejohnnycashproject.com/.

Zhang, H.; Law, E.; Miller, R.; Gajos, K; Parkes, D.; and Horvitz, E. 2012. Human Computation Tasks with Global Constraints. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 217–26. New York, NY: CHI '12.