Matching and Grokking: Approaches to Personalized Crowdsourcing

Peter Organisciak[1], Jaime Teevan[2], Susan Dumais[2], Robert C. Miller[3], Adam Tauman Kalai[4]

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

[2]Microsoft Research
Redmond, WA
{teevan, sdumais}@microsoft.com

[3]MIT CSAIL
Cambridge, MA
rcm@mit.edu

[4]Microsoft Research New England
Cambridge, MA
adam.kalai@microsoft.com

Abstract

Personalization aims to tailor content to a person's individual tastes. As a result, the tasks that benefit from personalization are inherently subjective. Many of the most robust approaches to personalization rely on large sets of other people's preferences. However, existing preference data is not always available. In these cases, we propose leveraging online crowds to provide on-demand personalization. We introduce and evaluate two methods for personalized crowdsourcing: *taste-matching* for finding crowd workers who are similar to the requester, and *taste-grokking*, where crowd workers explicitly predict the requester's tastes. Both approaches show improvement over a non-personalized baseline, with taste-grokking performing well in simpler tasks and taste-matching performing well with larger crowds and tasks with latent decision-making variables.

1 Introduction

Paid online crowds can augment user-facing systems in difficult-to-automate settings. However, while this form of crowdsourcing tends to support tasks that have a ground truth, many scenarios involve subjective judgments that are particular to the needs of a user.

This paper applies crowdsourcing to personalization in subjective, user-specific contexts. While popular approaches to personalization rely on large amounts of preference data from other users, the application of *personalized crowdsourcing* is a valuable alternative in cases where less data is available. This opens the door to personalization by providers, who do not have large user bases, or for novel or highly specific content, and even over personal collections.

The main contribution of this paper is a set of two approaches for personalized crowdsourcing over on-demand data: *taste-matching* and *taste-grokking*. These are two possible ways to consider personalized crowdsourcing, and their successes provide insights for future methods. Taste-matching recruits crowd workers based on their similarity to the requester, their contributions more effective than that of a generally recruited worker. In contrast, taste-grokking does not screen workers, instead relying on workers' ability to make sense of, or grok, a requester's needs based on a succinctly communicated taste-profile.

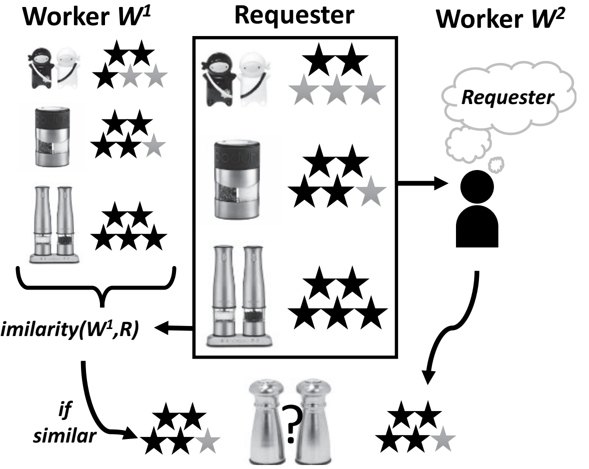


Figure 1: Examples of taste-matching, left, and taste-grokking, right.

Figure 1 demonstrates a basic opinion-rating example of taste-matching. After a requester provides their opinion on a profiling task (center), a taste-matching worker completes the same task (left). A similarity method determines how good of a 'match' the worker is to the requester, and if they are similar, they complete more ratings. These ratings predict what the requester would like. Taste-grokking (right) also uses the requester's opinions on the profiling example, showing them to the taste-grokking workers. Workers use this information about the requester to make an educated guess of the requester's opinion on future items.

Both taste-matching and taste-grokking offer improvements in subjective tasks, but each has strengths in different contexts. In this paper, we evaluate personalization in a simple and well-studied item recommendation task for products and food, and in a more complex task of highlighting important text in a short article. We find that taste-grokking is effective at the less complex task and appears to be more satisfying for workers, while taste-matching is more feasible in spaces with many latent decision-making factors and more cost-effective for long-term engagements.

Our approach can benefit a variety of problems that rely on the subjective needs of a user, such as tasks over personal data (e.g., choosing the best photographs from a large personal archive), tasks where items are unique (e.g., shopping for handmade artwork), and tasks where the items change often (e.g., finding the perfect apartment). Another use is for new systems that do not yet have rich preference data from users. This paper introduces the concept of personalized crowdsourcing, and demonstrates ways to pursue it by evaluating two approaches in multiple contexts.

2 Related Work

The common application of paid crowdsourcing is toward problems that make an objective assumption, with the expectation of a 'correct' answer. This is because many paid uses of crowds are in the paradigm of human computation, completing work in a manner akin to computation [Quinn and Bederson, 2011]. However, numerous projects have pursued crowdsourcing tasks that are either subjective or influenced by different worker biases and requester needs. For example, tagging images, judging the best frame of a video [Bernstein et al., 2011], rating similarity between images [Tamuz et al., 2011], and editing documents [Bernstein et al., 2010] all exhibit variances between how workers interpret them. Recent work has discussed the challenges of deriving consensus from paid crowd imbued by underlying subjectivity [Alonso et al., 2013].

Tasks that do not enforce a notion of a correct answer are common in volunteer crowdsourcing settings, such as opinion ratings. Occasionally, projects make use of the variability of crowds to artistic effect, such as in a project to animate a Johnny Cash music video one frame at a time [The Johnny Cash Project].

We focus on the applicability of personalized crowdsourcing for problems where a task is time-consuming for an individual, but difficult to delegate because the proper completion of the task is specific to the target person. The time cost of properly completing a task temps the benefits of doing so. Marmorstein et al. [1992] note that the point at which the time-cost trade-off results in inefficient decision-making varies by individual, but that it has been observed in areas such as comparison-shopping, travel planning, and job-hunting.

There are recent focused efforts to address this class of problem through paid crowdsourcing. EmailValet [Kokkalis et al., 2012] allows people to find an email assistant from the crowd, communicating their preferences by describing them. Mobi takes a similar approach for travel-planning [Zhang et al., 2012]. Our study builds on these approaches, while considering the problem space of subjective tasks more generally. While this study's evaluated methods involve statistical matching and communication-by-example, EmailValet and Mobi's approach of communicating preference by natural language description is another possible method for personalization.

Taste-matching and taste-grokking have precedents in personalization research. Krishnan et al. [2008] evaluate an approach to personalization that is similar to taste-grokking, communicating taste-by-example for human recommenders in the context of film recommendation. Though they found that the mature MovieLens collaborative filtering system performed better in general, the human recommendations were more effective for unusual or eclectic requester profiles. Our research pursues a similar approach, though focusing on more novel settings than film where there is no access to prior data.

```
1. Choose profiling set $S \subset X$ of examples
2. Requester $t$ performs work on each object in $S$
If taste-matching
    A.3. Workers $w \in W$ perform work on $S$
    A.4. Worker pool subset $W' \subset W$ is selected by
    similarity to $t$
    A.5. For each subsequent task, workers $w \in W'$
    perform work on remaining data $X \setminus S$
If taste-grokking:
    B.3. Work by $t$ on $S$ is shown to workers $w \in W$
    B.4. Workers $w \in W$ predict $t$'s opinions on $X \setminus$
    $S$
    B.5. Optional 'wisdom of crowds' quality
    control (e.g., aggregation)
```

Figure 2: The taste-matching and taste-grokking approaches

Taste-matching is similar to collaborative filtering [e.g., Resnick et al, 1994; Hofmann, 2004] in that it relies on the opinions and behaviors of similar people to a target user. Since workers in taste-matching contribute data on request, a central concern in collaborative filtering is sidestepped—sparse data for new or unseen items.

In summary, our study builds on past work in crowdsourcing and personalization in order to consider on-demand crowdsourcing for subjective tasks. While many prior uses of personalization (e.g. recommendations on Netflix, Last.fm) use implicit and explicit crowd-contributed data, our study applies many of the same intuitions to more difficult cases of sparse spaces and on-demand individual needs. We introduce and explore two approaches to collecting subjective data, taste-matching and taste-grokking, and through them provide future researchers a framework for thinking about personalized crowdsourcing.

3 Approach

We study two approaches for on-demand personalization through crowdsourcing: *taste-matching*, where crowd workers are screened based on their similarity to the requesting user, and *taste-grokking*, where crowd workers try to guess the preferences of a requesting users based on a set of profiling questions. We refer to the person receiving personalized content as a *requester*. Though personalized crowdsourcing is not inherently dependent on paid contributions, this paper focuses on paid settings, so we refer to contributors as *workers*.

3.1 Profile Construction

Both protocols first profile the requester's preferences or tastes, who completes a subset of subjective work items. For example, if the personalization task calls for recommendations of an online product for the requester, the profiling step may have the requester rate whether they like or dislike similar items. For example, Figure 1 shows the requester profile for examples from a product recommendation task.

The selection and quantity of items to use for profiling can influence the quality of the personalization, as will be discussed later. It is important to capture the breadth of the requester's subjective profile, especially in taste-grokking where workers are trying to understand a requester. For most experiments, we select profiling items at random, though item selection that is more purposive is possible.

The profiling data is used differently by taste-matching and taste-grokking. In matching it is used statistically, to match people on similarity, while in grokking it is used explicitly by crowd workers.

3.2 Taste-Matching

Taste-matching uses the requester's profile to identify workers who are similar to the requester for the given task.

When recruiting workers, they are first given the profiling set to complete. A metric appropriate to the task measures similarity the similarity of workers to requesters. The deviation from the requester's ratings measures similarity for this study's item recommendation task, while this study's text highlighting task uses highlighting overlap.

After a similarity to the requester is determined, the most-similar workers provide additional contributions for personalization (e.g., more ratings, more text highlights). This study evaluates this approach of screening workers, but a more complex use can accept all contributions and weigh them according to similarity.

Taste-matching is similar to collaborative filtering (CF), which uses similar users' opinions to make recommendations. Both approaches assume that people who agree on a subset of a domain will have similar opinions and tastes elsewhere in that domain. Where taste-matching differs is that data is explicitly rather than implicitly collected, and done so on-demand. This can open up new personalization settings: for example, if a requester has a large set of personal vacation photos to cull down, taste-matching can be used where CF would not have been possible.

3.3 Taste-Grokking

Taste-grokking makes a different assumption than taste-matching, pursuing the notion that workers explicitly shown a profile of the requester can sufficiently infer the requester's needs. Rather than performing subjective tasks in their own style, workers 'grok' – or make sense of – what the requester needs based on the profiling set. Then they perform future work against that mental model. Since they are working against a common notion of truth, the requester's preferences are easily aggregated across multiple workers' grokked work. We do not attempt a combination of methods, i.e. applying taste-matching to grokking workers, but similar work in the area of film recommendation found that better matched human recommenders do not seem to perform better on this type of inferential task [Krishnan et al., 2008].

This study communicates the requester's needs by example – e.g., asking them to rate images that grokking workers will use to predict future ratings. Communicating an example has been found to be more effective than explicit description of needs in the area of personalized search [Teevan et al., 2008].

3.4 Evaluation

We study personalized crowdsourcing across two domains: personalized item recommendation for food or products based on images (discussed in Section 4), and text summarization by personalized highlighting (Section 5).

4 Personalized Item Recommendation

Recommendation by inferring a target person's rating of an item is a common personalization task. While it often done in rich domains such as film, we focus on two sample domains with limited existing preference data: predicting a requester's opinion of salt and pepper shakers, and of local cuisine. Both of these domains represent less common but notably subjective spaces.

The product recommendation dataset consisted of 100 salt and pepper shaker images from Amazon's online store, while the cuisine recommendation dataset consisted of images and names of 100 popular restaurant meals in the cities of Boston and Seattle, from Foodspotting.com.

4.1 Methodology

Data Collection

Data was collected using Amazon's Mechanical Turk crowd marketplace, with workers rating all the images on a five-point scale. For the product images, workers rated on how much they like the salt and pepper shakers. For the food images, workers saw an image of a meal alongside the name of the dish, and rated how appetizing it appeared. These same questions were carried over to taste-grokking, but asked in the context of the targeted requester rather than the workers themselves.

For both taste-matching and taste-grokking, requesters are first profiled on their opinions of the shakers or cuisine, respectively. The profiling set was randomly selected from the full set of requester ratings: 20 images for taste-matching, or 10 images for taste-grokking. The smaller grokking set is due to the expectation that trying to comprehend 20 images would be unnecessarily complex, to the detriment of performance and worker satisfaction. The remaining ten ratings from the requester's profile are retained for cross-validation, to identify grokking workers that are performing particularly well. In all cases, we evaluated the personalized prediction quality against a requester's held-out opinions of 80 items.

Taste-Matching

The similarity of workers to requesters is measured by the deviations of the workers' ratings from those of the requesters, using root-mean-squared error (RMSE). RMSE is in the same units as the ratings, and lower error indicates a better match. Rather than matching by absolute rating, people's ratings are normalized as deviations from their mean rating to account for differing attitudes of what the choices on the rating scale mean [Hofmann, 2004].

Since contributions are requester-independent and matching is post-hoc, we simulated requesters from worker contributions for evaluation. The payment for workers was $1.50 for rating a set of 100 images.

Taste-Grokking

In taste-grokking, the 10 item profiling set of requester's ratings is shown directly to workers. Workers are asked to guess what the requester would judge for the next 90 items; i.e. how much they would like the salt shaker or how appetizing they would find the photo of cuisine.

We evaluated both individual grokking recommendations and aggregations of multiple predictions. Independent taste-grokking workers have the same goal, to interpret the requester, making it more sensible to use a consensus prediction.

Rating predictions were evaluated twice for four different requesters (for the one product task and two cuisine tasks), each time using a different profiling set. *30* workers performed each set of predictions and were paid between $1.00-$2.00, depending on the accuracy of their predictions.

Measurement

RMSE was used to measure how closely the rating predictions from taste-matching or taste-grokking mirrored the true ratings provided by requesters. A lower RMSE represents less deviation from the true ratings.

To evaluate the two approaches, we use RMSE to compare the ratings predicted by each approach with the requesters' true ratings. The baseline measure is the RMSE of a non-personalized prediction; that is, the quality of a prediction from any given contributor.

	Products	Food #1	Food #2
Baseline	1.64	1.51	1.58
Best Matched of 5	1.43 (-13%)	1.19 (-22%)	1.26 (-20%)
Best Matched of 10	1.35 (-18%)	1.08 (-29%)	1.08 (-31%)

Table 1: Average RMSE of taste-matching predictions

	Products	Food #1	Food #2
Baseline	1.64	1.51	1.58
Individual grokkers	1.29 (-21%)	1.53 (+1.3%)	1.57 (-0.5%)
5 random grokkers	1.07 (-34%)	1.38 (-9%)	1.28 (-19%)
5 top grokkers	1.02 (-34%)	1.22 (-19%)	1.08 (-28%)

Table 2: Average RMSE of taste-grokking predictions

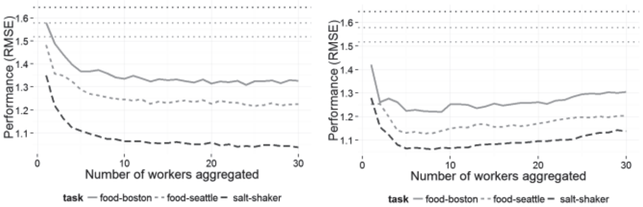


Figure 3: Performance of taste-grokking predictions aggregated from N random workers (left) and the N best workers (right). Shown for different sized pools of workers.

4.2 Results

Taste-Matching

In a realistic taste-matching setting, a requester can post a task, wait for *n* contributors to be matched, and then take the contribution of the best-matched worker. Table 1 shows the performance of taste-matching considered in this way, when five or ten workers are available. For all datasets, taste-matching resulted in improvements in predicting the preferences of a requester, with stronger results on the food dataset. Increasing the pool of possible workers consistently improves quality.

Taste-Grokking

Taste-grokking was evaluated both with aggregation and for individual workers (Table 2). Without aggregation, the salt shaker recommendation task showed a 21.3% improvement over the baseline, while performance for the food datasets did not show improvement.

Aggregating multiple predictions into a single rating is more effective, smoothing over individual workers' errors. In Table 2 we show aggregations of five workers – a number of redundant workers recommended by Novotney and Callison-Burch [2010] for encoding tasks – for both five randomly chosen workers or the five best workers as chosen by a small held-out set. To consider the cost-quality trade-off of adding additional contributions, Figure 3 shows the performance of N=1-30 workers. Improvements are seen over all datasets, but again taste-grokking is more effective for the salt and pepper shaker products. Aggregating multiple contributions and cross-validating work against a ground truth are common quality control techniques in crowdsourcing objective tasks, and they appear similarly effective in taste-grokking.

The above results used a randomly selected profiling set. We also evaluated the performance of grokking over an optimized training set, using K-means clustering to determine taste-groups (where *k* is equal to 10, the profile set size), then sampling one item from each group for the profiling set. The optimized training examples greatly improved the performance of taste-grokking (Figure 4).

4.3 Summary

On-demand crowd-based rating prediction through both taste-grokking and taste-matching provides improvements over an non-personalized approach. The two evaluated domains improved differently from each approach: recommending cuisine is more effective when matching similar workers to a requester, while recommending products is better done by workers grokking a requester and explicitly guessing that requester's ratings.

Taste-matching is more effective with larger numbers of contributions, while taste-grokking was more amenable to aggregation. Since there is a notion of truth when grokking workers are trying to interpret a requester, taste-grokking also allows for quality control: in our case, we used a small held-out set to see who was good at grokking.

In addition to evaluating the overall performance of taste-matching and taste-grokking, a number of auxiliary results were observed. This includes a comparison of the effect of profiling set selection in taste-grokking, where a stratified sampling method that selects from diverse clusters in the data was found to be more effective than completely randomized sets. We also found that the best taste-grokkers performed better than the best matched-workers, although how well the best grokkers can be identified a priori remains to be seen.

5 Personalized Text Highlighting

In order to consider personalized crowdsourcing in a more complex domain, the next task evaluates taste-matching and taste-grokking when applied to a summarization-based task. Specifically, personalized crowdsourcing is used to highlight key points in a film review (as mocked up in Figure 5). The motivation for text-highlighting as a personalization task is to enable tailored summaries, dependent on what types of information a person is searching for. It is common to deal with large numbers of texts – perhaps a paralegal researching court decisions or a scholar reading papers – and being able to summarize a text for specific to the reader's needs is potentially useful. Furthermore, while deciding what text is interesting or uninteresting is subjective, it depends on user context in addition to opinions, providing a somewhat different take on user-specific tasks than the item recommendation task.

5.1 Methodology

The texts used for the highlighting task were six film reviews by professional critics at *The A.V. Club*, averaging 456 words.

For each of six reviews, 50 paid crowd workers highlighted film reviews for passages deemed useful in deciding to see the film. For taste-matching, requesters were simulated from other worker submissions.

Requesters highlighted only one review for the profiling set, both to minimize their effort and because more than one review was expected to be too difficult to interpret for grokking. Workers highlighted up to six reviews, choosing what and how much they highlighted. Without restriction, there was a great deal of variance in the quantity of highlights.

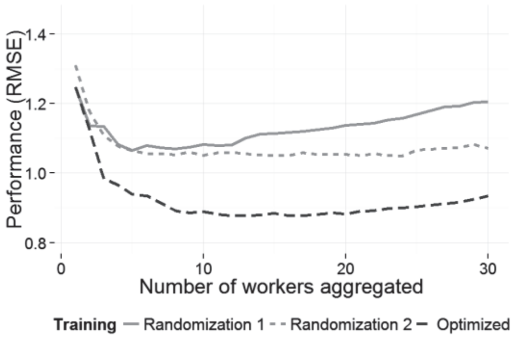
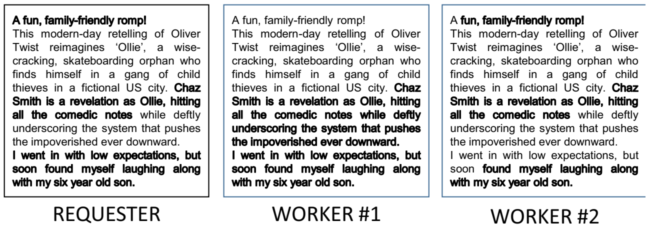


Figure 4: Effect of different taste-grokking profiling items on performance of top k aggregation. Long-dash line represents optimized profiling set.



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

Figure 5: An example of a highlighted movie review

The F1 score measures matching similarity and text highlighting quality for this task. F1 is the harmonic mean between precision (what proportion of the worker's highlighting overlaps with the requester's highlights) and recall (how much of the requester's highlights did the worker cover). F1 ranges from 0 to 1, with a higher score indicating a better match. In the example shown in Figure 5, Worker #2 has the higher F1 score, matching closely what the requester highlighted, as well as avoiding passages that the requester did not highlight.

For taste-matching, highlights were collected from 50 crowd workers. They highlighted information they would find useful in choosing to see a film. Workers were matched to requesters based on textual overlap on the one-review profiling set, measured by F1 score. One other review was held as an alternate profiling set, while the highlights on the remaining four films were used for evaluation.

For taste-grokking, data was collected for three different requesters, and with two different profiling sets. For each of these six conditions, 30 crowd workers highlighted reviews for four films. Workers were shown a single highlighted review by the requester, and asked to highlight what they thought the requester would find interesting in the other reviews.

5.2 Results

The baseline was the mean quality of a non-personalized highlighted text, as determined through F1. It showed an F1 score of *0.32*.

In taste-matching, workers that were matched by a high F1 measure for the profiling text highlights improved on the baseline. The best-matched workers had a mean F1 of *0.39*, a *20%* improvement, while the 5 best-matched workers averaged an F1 of *0.38* across all conditions, an improvement of *17%.* These improvements suggest that the highlighting task does indeed have a subjective component, and people who highlight similarly on a known text can be expected to do so in the future.

In contrast, taste-grokking was not nearly as robust. Understanding a requester's needs from their highlights on one film review proved difficult to generalize to other reviews, and the average taste-grokking suggestions had an F1 of *0.30*, which was somewhat worse than the baseline. Interestingly, some workers proved to be adept at the activity: the best grokking workers in the sets of thirty averaged *F1=0.73 (+128%)*, while the best worker from a random set of five averaged *F1=0.52 (+62%)*. This shows that grokking workers theoretically can generate much stronger highlights than matching, though in practice we did not find a method for identifying 'super-grokkers' a priori and grokkers in general performed poorly.

5.3 Summary

Application of personalized crowdsourcing to text highlighting for summarization was a more difficult task than image-based item recommendation. Taste-matching provided reliably quality improvements. Taste-grokking was less predictable: the best workers performed very well, but the average worker had difficultly anticipating the highlighting style of a requester, with results comparable to the baseline.

6 Discussion

We found personalized crowdsourcing to be a feasible approach to on-demand personalization, with two approaches that were effective to varying degrees. Though taste-grokking and taste-matching are only two possible approaches to personalized crowdsourcing, they show strength over different tasks and task types. These strengths offer insight into personalized crowdsourcing and help in considering future research.

Both taste-matching and taste-grokking present improvements over traditionally completed crowd-contributed data collection in cases where there is an element of subjectivity. Taste-matching performs better for complex spaces, with many potential decision-making factors, and less explicit manifestations of those factors. Grokking, in contrast, works particularly well for predicting the manifest and visually imbued salt and pepper shaker task.

Taste-grokking was notably stronger than taste-matching in terms of worker satisfaction. In voluntary feedback, numerous taste-grokking workers expressed that they enjoyed the novelty of the task. It is possible that the slightly competitive form of payment contributed partially to this attitude. Conversely, grokking failures, as observed at least once with a poor profiling set, appeared to be notably distressing.

Another area where taste-grokking excelled was in reframing subjective tasks around a notion of truth: the requester's tastes. This makes it easier to measure objective quality: how good is a worker at grokking for this requester? In contrast, a typical subjective task confounds worker quality with the variability of human opinions and interpretations. While quality metrics are more difficult with taste-matching, they are also less relevant since an inattentive or sloppy worker will simply not match a requester. However, a strategic cheater could conceivable give realistic contributions on profiling but poor future contributions.

Where taste-grokking performed poorly was when opinion-forming factors were more difficult to grok. When asking workers to justify their ratings on salt and pepper shakers, they primarily referenced visual, easily seen factors: this may account for some of taste-grokking's strength on that task. In contrast, some correlations in cuisine rating were less obvious, like an overlap between beer and shawarma lovers. The 'grokability' of a task should be an important consideration in deciding how to pursue personalized crowdsourcing, and is worthy of further study.

Taste-matching also benefits from the fact that it does not need to be specifically collected for each requester. Contributions are reusable between people, and one can imagine the technique bootstrapping a more mature system when there is a lack of existing data. This approach may be seen as a type of on-demand collaborative filtering. Applying paid crowds to the larger collaborative filtering systems is worth exploring further. Taste-matching can augment a larger system when new users join or when new items are added to its collection.

7 Conclusion

There are many subjective settings in human-computer interaction where people are best served by personalization, but a lack of prior information makes personalization difficult. We show that paid crowdsourcing is useful for on-demand personalization in such cases. With personalized crowdsourcing, a requester can feasibly collect data for personal or otherwise private datasets, novel domains, or new systems.

Much on-demand crowdsourcing focuses on seeking consensus or an objective truth. However, we find that the diverse and qualitative nature of the crowd makes them well-suited for subjective task completion. This is demonstrated through two protocols for designing personalized crowdsourcing tasks: *taste-matching* and *taste-grokking*. Both protocols usually improve over an un-personalized baseline, but they each show different strengths. Taste-matching is effective for more complex tasks, particularly with many latent decision-making factors affecting one's task, and is useful for scaling to large numbers of requesters and workers. Taste-grokking works well in easier to communicate domains, is more effective with small numbers of workers, and appears to be a more engaging approach to explicit data collection.

Crowdsourcing shows promise for on-demand personalization. Our results show this with two approaches over two task types and various domains, and suggest that further research into personalized crowdsourcing is worthwhile.

References

[Ahn and Dabbish, 2004] Luis von Ahn and Laura Dabbish. Labeling Images with a Computer Game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 319–26, Vienna, Austria, 2004.

[Alonso et al., 2013] Omar Alonso, Catherine C. Marshall, and Mark Najork. 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, 2013.

[Bernstein et al., 2011] Michael S. Bernstein, Joel Brandt, Robert C. Miller, and David R. Karger. Crowds in Two Seconds: Enabling Realtime Crowd-Powered Interfaces. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, pages 33–42, 2011.

[Bernstein et al., 2010] Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2010. Soylent: A Word Processor with a Crowd Inside. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*, pages 313–22, 2010.

[Bernstein et al., 2008] Michael S. Bernstein, Desney Tan, Greg Smith, Mary Czerwinski, and Eric Horvitz. Personalization via Friendsourcing. *ACM Transactions on Computer-Human Interaction* 17 (2), pages 6:1–6:28, 2008.

[Hofmann, 2004] Thomas Hofmann. Latent Semantic Models for Collaborative Filtering. *ACM Trans. Inf. Syst.* 22 (1): pages 89–115, 2004.

[Kokkalis et al., 2013] Nicolas Kokkalis, Thomas Köhn, Carl Pfeiffer, Dima Chornyi, Michael S. Bernstein, and Scott R. Klemmer. EmailValet: Managing Email Overload through Private, Accountable Crowdsourcing. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, pages 1291–1300, 2013.

[Krishnan et al., 2008] Vinod Krishnan, Pradeep Kumar Narayanashetty, Mukesh Nathan, Richard T. Davies, and Joseph A. Konstan. 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*, pages 211–18, 2008.

[Marmorstein et al., 1992] Howard Marmorstein, Dhruv Grewal, and Raymond P.H. Fishe. The Value of Time Spent in Price-Comparison Shopping. *Journal of Consumer Research* 19(1):52-61, 1992. University of Chicago Press.

[Novotney and Callison-Burch, 2010] Scott Novotney and Chris Callison-Burch. Cheap, Fast and Good Enough: Automatic Speech Recognition with Non-Expert Transcription. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 207–15, Stroudsburg, PA, USA, 2010.

[Quinn and Bederson, 2011] Alexander J. Quinn and Benjamin B. Bederson. Human Computation: A Survey and Taxonomy of a Growing Field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* pages 1403-1402. New York, NY, 2011.

[Resnick et al., 1994] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. "GroupLens." In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, pages 175–86, 1994. ACM Press.

[Tamuz et al., 2011] Omer Tamuz, Ce Liu, Serge Belongie, Ohad Shamir, and Adam Tauman Kalai. Adaptively Learning the Crowd Kernel. In *Proceedings of the International Conference on Machine Learning*, 2011.

[Teevan 2008] Jaime Teevan, Susan T. Dumais, and Daniel J. Liebling. To Personalize or Not to Personalize. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 163-170, 2008.

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

[Zhang et al., 2012] Haoqi Zhang, Edith Law, Rob Miller, Krzystof Gajos, David Parkes, and Eric Horvitz. Human Computation Tasks with Global Constraints. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 217–226. New York, NY, 2012.