Task Recommendation Experiments

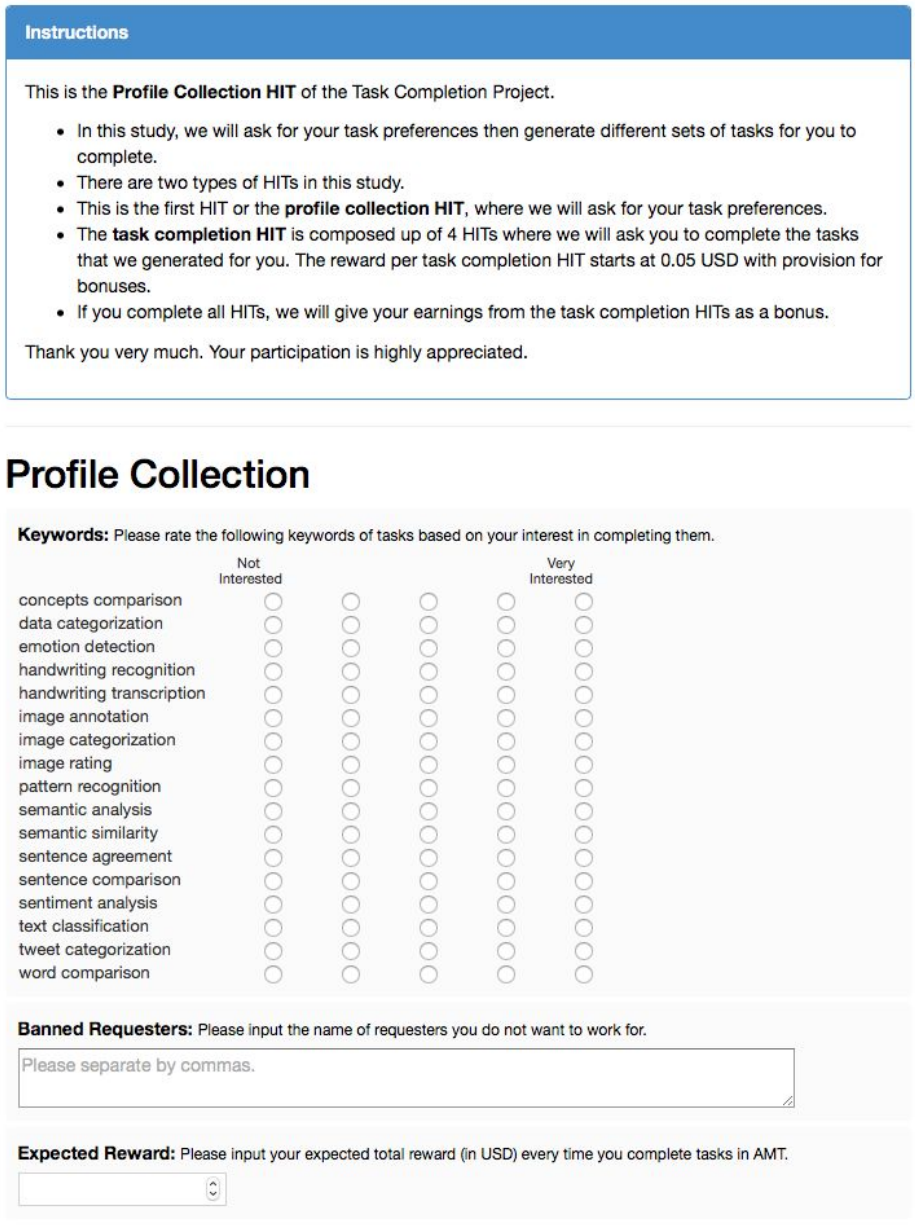
In the real data experiments, we use OUR algorithm to generate task sets for crowdsourcing workers, based on different scenarios then ask workers to complete individual tasks and rate the task sets based on how well they liked them. We then evaluate overall completion time, task throughput, quality with respect to a ground truth, total reward, and ratings of the task sets.

Task Dataset

The dataset consists of 20,000 tasks from Figure Eight`'s open data library. Each task belongs to one of the 10 different task types such as tweet classification, image transcription, sentiment analysis, and entity resolution. A task type is assigned a set of keywords that best describe the skills required by the task, creation date, expected completion time or duration, and a reward ranging between $\$0.01$ and $\$0.10$, proportional to its expected completion time. The tasks are {\em micro-tasks}, tasks which take less than a minute to complete.

Experiments Flow

1. Collect 200 user profiles per context; 50 profiles per context.



Instructions

This is the **Profile Collection HIT** of the Task Completion Project.

- In this study, we will ask for your task preferences then generate different sets of tasks for you to complete.
- There are two types of HITs in this study.
- This is the first HIT or the **profile collection HIT**, where we will ask for your task preferences.
- The **task completion HIT** is composed up of 4 HITs where we will ask you to complete the tasks that we generated for you. The reward per task completion HIT starts at 0.05 USD with provision for bonuses.
- If you complete all HITs, we will give your earnings from the task completion HITs as a bonus.

Thank you very much. Your participation is highly appreciated.

Profile Collection

Keywords: Please rate the following keywords of tasks based on your interest in completing them.

	Not Interested				Very Interested
concepts comparison	○	○	○	○	○
data categorization	○	○	○	○	○
emotion detection	○	○	○	○	○
handwriting recognition	○	○	○	○	○
handwriting transcription	○	○	○	○	○
image annotation	○	○	○	○	○
image categorization	○	○	○	○	○
image rating	○	○	○	○	○
pattern recognition	○	○	○	○	○
semantic analysis	○	○	○	○	○
semantic similarity	○	○	○	○	○
sentence agreement	○	○	○	○	○
sentence comparison	○	○	○	○	○
sentiment analysis	○	○	○	○	○
text classification	○	○	○	○	○
tweet categorization	○	○	○	○	○
word comparison	○	○	○	○	○

Banned Requesters: Please input the name of requesters you do not want to work for.

Please separate by commas.

Expected Reward: Please input your expected total reward (in USD) every time you complete tasks in AMT.

2. Generate 5 task sessions for the scenarios below.
   Note: For long sessions, there are 5 sets per session, 10 songs per set; For short sessions, 3 sets per session, 3 songs per set.

   Case 1 - min intra(skill), max inter(reward)

In the first case, we minimize skill diversity within a window thus each window has tasks that require similar skills, and maximize reward diversity across windows thus each window's average reward is varied.

Case 2 - max intra(skill), min inter (reward)
In the second case, we maximize skill diversity within a window thus each window has tasks that require different skills, and minimize reward diversity across windows thus all the windows' average rewards are similar.

Case 3 - max intra(duration), max inter (skill)
In the third case, we maximize duration diversity within a window. This means that the tasks in a window have different expected completion times, which can be an indicator of a task's difficulty. We also maximize skill diversity across windows thus the skills required for every window are different.

Case 4 - min intra(creation date), min inter(skill)
Lastly, we minimize the diversity of task creation dates within a window thus each window has tasks published around the same time. Additionally, we minimize skill diversity across windows thus similar skills are required to complete tasks for every window.

Case 5 – no diversity

3. Task Completion
   We then upload the task sets into a task browser application where workers can view and complete the task sets generated for them. Once uploaded, we invite the workers through AMT to do a task completion HIT, where a worker is directed to view and complete tasks in the task browser application. Figure X shows a task set as seen by workers, where each tab represents a window. Workers have the option to choose only the tasks they want to complete. When a worker decides to finish the task set, a unique code is given to the worker, which must be inputted in AMT for the worker to receive a reward. Additionally, we ask workers to rate the task set from 1 to 5 based on how well they liked the task sets.

1 2 3

Finish Task

1) Please read the following new year's resolution tweet.

#NewYearsResolution To save my money and not blowing it on food

What do you think is the category of the tweet?

- Health & Fitness
- Humor
- Personal Growth
- Philanthropic
- Recreation & Leisure
- Family/Friends/Relationships
- Career
- Finance
- Education/Training
- Time Management/Organization

Submit

Title: 2015 New Year's resolutions
Description: *Classify tweets about new year resolutions.*
Keywords: sentiment analysis, emotion detection, resolutions
Requester: Aiko
Reward: 0.05 USD

2) How similar are the following terms?

Term 1: *three families*
Term 2: *three families*

Completely different | Very different | Slightly different | Neutral | Quite similar | Very similar | Exactly the same

Submit

Title: Similarity judgment of word combinations
Description: Evaluate the similarity of two sets of words.
Keywords: word comparison, semantic similarity
Requester: Makoto
Reward: 0.03 USD

Next