

ワーカの成長を考慮した 自己補正マイクロタスク割り当て手法

小林 正樹[†] 清水 伸幸^{††} 森嶋 厚行^{†††}

[†] 筑波大学 情報学群 知識情報・図書館学類 〒305-8550 茨城県つくば市春日 1-2

^{††} ヤフー株式会社 〒102-8282 東京都千代田区紀尾井町 1-3 東京ガーデンテラス紀尾井町 紀尾井タワー

^{†††} 筑波大学 図書館情報メディア系/知的コミュニティ基盤研究センター 〒3058550 茨城県つくば市春日 12

E-mail: [†]ts1311495@klis.tsukuba.ac.jp, ^{††}nobushim@yahoo-corp.jp, ^{†††}mori@slis.tsukuba.ac.jp

あらまし クラウドソーシングにおける研究課題のひとつとして、タスク結果の品質が上げられる。品質を左右する要因の一つに、ワーカのスキルや知識がある。本論文では、他人のタスク結果を参考に自分のタスク結果の変更を行う「自己補正」を対象として、ワーカが同一種類のタスクを繰り返す行う際に、ワーカの成長に影響を与える要素を調査する。予備的な実験の結果から次が示唆された。(1) タスクを行う際、タスクそのものや、他人の回答に関する情報を増加させると、タスクを行った後、ワーカがそのタスクを行う際の品質が向上する。(2) それらの情報が間違っていると、タスク終了後も、品質は向上しない、(3) タスクや他人の回答に関する正しい情報を増加させるのが困難な時には、品質の高い結果を返す優秀なワーカの回答を見せることが有効である。

キーワード クラウドソーシング, タスク割り当て

1. はじめに

近年、計算機のみでの解決が困難な問題に対するアプローチとして、クラウドソーシングが注目を集めている。クラウドソーシングは、ネットワークを通して不特定多数の人間（以下、ワーカと呼ぶ）に対して作業（以下、タスクと呼ぶ）を委託することで、問題解決を試みる手法である。クラウドソーシングの利点は、特定の問題を解決するためのソフトウェアを開発するよりも安価であり、早急に実施できることである。クラウドソーシングでは、問題解決のための諸作業を細分化し、複数のワーカに作業を委託するのが一般的である。特に、各タスクの処理負荷（所要時間や難易度）を抑えるなどの工夫により、簡潔に回答できるタスクはマイクロタスクと呼ばれている。本稿では、画像や映像、音声へのタグ付けや文字認識、文章構成などの作業を扱うマイクロタスク型クラウドソーシングに注目する。

クラウドソーシングにおける研究課題の1つとして、成果物の品質管理が上げられる。クラウドソーシングには、各ワーカの回答（以下、タスク結果と呼ぶ）の集合である成果物に、誤ったタスク結果が含まれるという問題がある。誤った回答が含まれる原因は2つある。まず、第一に、作業を委託する対象は不特定多数であり、各ワーカの回答能力や得意とする問題が異なることである。第二に、BOT やスパムワーカなどが実際の作業を行わずに報酬のみを受け取る目的でランダムな回答を行うことである。これまで、能力が低かったり、スパムであるようなワーカの影響を下げることによって、タスク結果の品質を向上させる手法が広く研究されている。

本稿では、これらの研究とは異なり、ワーカの能力を向上させることにより、タスク結果の品質を向上させる問題に取り組む（図1）。すなわち、一連のタスクを行った後に、タスク結果

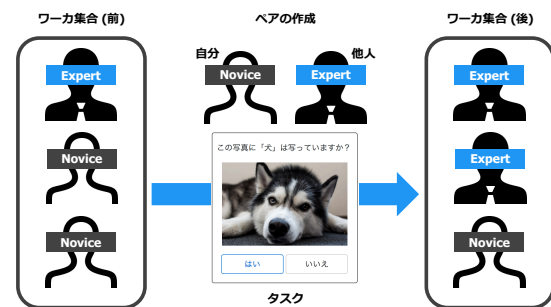


図1 タスクの自己補正を通じたワーカの成長。自己補正とは、他人の回答を見ながら自分の回答を修正する事である。本稿では、複数の自己補正タスクを行う中で、どのような要因がワーカの成長と関係があるのかを明らかにすることを目指す。

の高い品質をもたらす優秀なワーカ（本稿では Expert と呼ぶ）の割合を増やすことを目標とし、具体的には、タスクの自己補正 [1] を複数回行う際に、どのような要素がワーカの能力をより良く向上させるのに影響があるのかを調査する。自己補正は、問題に回答するタスク（以下、第1タスクと呼ぶ）と、第1タスクでの自身の回答を確認するタスク（以下、第2タスクと呼ぶ）の2段階で構成される（図2）。第1タスクでは、一般的なクラウドソーシングと同様に問題に回答する作業を行う。第2タスクでは、提示される回答を見て、自身の回答を見直すという作業を行う。自身の回答を見直すことで、回答の誤りに気づくことが出来るのが自己補正の利点である。この気づきによってタスク結果の品質が向上する事はすでに論文 [1] で示されて

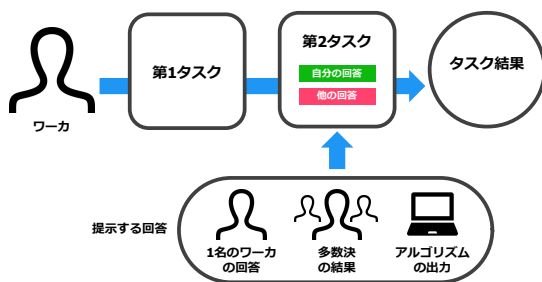


図2 自己補正の概要

いるが、本論文では、単に個々のタスク結果の品質を問題にするのではなく、同一種類の複数のタスクで自己補正を行う際に、どのような要因がワーカーの成長に影響するのかを調査する。

今回、ワーカーの成長に影響を与える可能性があるとして検証した要素は次の3つである。

- (1) **タスクの情報**：タスクの説明に関して、より詳しい情報を提示した場合の影響を調べる。
- (2) **他者の回答の情報**：他者の回答に関して、より詳しい情報を提示した場合の影響を調べる。
- (3) **他者の回答の品質**：より優秀な（品質の高いタスク結果をもたらす）人の回答を提示した場合の影響を調べる。

2. 関連研究

これまでクラウドソーシングにおける成果物の品質管理に関する研究が広く行われている。ここではクラウドソーシングにおける成果物の品質管理の問題について、4つの観点から関連研究を整理する。

2.1 ワーカーの能力向上

新規のワーカーに対して訓練を行うことに注目した芦川らの研究がある[2]。芦川らは、新規のワーカーに対して訓練タスクをこなした後に、本番タスクを割り当てることで、本番タスクでの回答精度を向上させる手法を提案した。論文[3]では、ワーカーの学習のために必要な最小限のタスクを選択する手法を提案している。

本研究は、訓練や学習のためのタスクを用意するのではなく、タスク作業と自己補正を繰り返す中で、どのような要因がワーカーの成長につながるかを明らかにしようとするものである。

2.2 他者の情報を提示

自己補正を提案したShahらの研究は、2段階のタスクによりワーカー自身がタスク結果の品質を改善する枠組みと、そのための報酬設計を提案をするものである[1]。他の研究としては、複雑なタスクを解くワーカーの回答品質を向上させるための、タスク設計手法を提案するDoroudiらの研究がある[4]。この研究では、優秀なワーカーの解答手順を示すことで問題を解くワーカーの回答品質が向上するが、現実的な方法として他者の解答手順を示す手法が有効であるという結論を示した。

これらの研究では、他者の情報を提示することによるタスク結果の品質向上については議論しているが、それらがどのよう

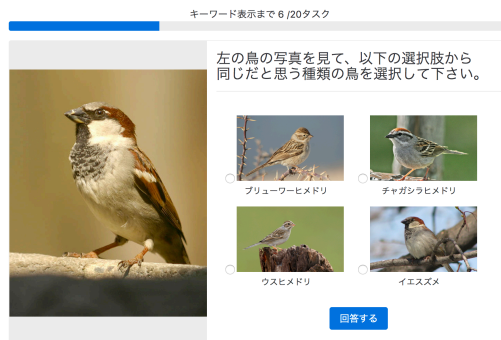


図3 自己補正の第1タスク



図4 自己補正の第2タスク：質問文の下部に、自分の回答と他者の回答が表示されている

にワーカーの成長につながるかについては議論していない。

2.3 ワーカーの評価

過去のタスク結果を用いてワーカーの評価をする手法として、回答が既知のタスクを用いる方法がある。この方法では、回答が既知のタスクにおける誤り率を、ワーカーの能力とする。清水らは、回答既知タスクの出題方法の検討をした[5]。

一方で、回答が既知のタスクを必要とせずに、ワーカーの能力を推定する手法が提案されている。その1つが、EM法を用いてワーカーの能力と最終的なタスク結果を交互に推定する方法である[6][7][8]。Joglekarらは、多数決における回答の一致に基づいて、ワーカーの誤り率を推定する手法を提案した[9][10]。Jiangらは、能力が様でないワーカー集合を2つのグループに分割する手法を提案した[11]。Jiangらは、ワーカーの能力に基づいて分割することでタスク結果の品質が向上することを示した。

大谷らは、階層分類タスクの集約手法を提案した[12]。大谷らの評価実験では、提案手法が正確なラベルを出力するだけでなく、ワーカーの能力を推定することにも有効であることを示した。

自己補正で提示する回答を与えるワーカーの能力を一つの要因として調査する際、これらの手法が利用可能と考えられる。

3. 自己補正と調査する要素

3.1 自己補正

自己補正は、(1) 問題に回答する第1タスクと、(2) 第1タス



図 5 選択肢の画像を提示しない (左), 提示する (右)



図 6 他者回答の理由を提示しない例 (左), 提示する例 (右)

クでの自身の回答を確認する第 2 タスクで構成される。本稿における自己補正の各タスクの役割を以下で説明する。

【自己補正の第 1 タスク】

通常のクラウドソーシングで用いられる、質問と選択肢が与えられるタスクである (図 3)。以降では単に第 1 タスクと呼ぶ。

【自己補正の第 2 タスク】

自己補正の第 1 タスクの内容に加えて、第 1 タスクでの自身の回答と、他者の回答が表示されるタスクである (図 4)。以降では単に第 2 タスクと呼ぶ。

3.2 調査する要素

今回、ワーカーの成長に影響を与える可能性があるとして検証した要素は 3 つである。詳細は後述するが、実験で利用したタスクでは、鳥の写真を提示し、写っている鳥が 4 つの選択肢の中のいずれかをワーカーに尋ね、次の 3 つの要素の影響を調査した。

(1) タスクに関する情報。タスクの説明に関してより詳しい情報を提示した場合の影響を調べた。具体的には、単に候補の選択肢を 4 つ示す場合と、その候補のそれぞれについて写真を示した場合の比較を行った (図 5)。

(2) 他のワーカーの回答に関する情報。他者の回答に関してより詳しい情報を提示した場合の影響を調べた。具体的には、第 2 タスクにおいて、単に他のワーカーが行った回答を提示する場合と、そのワーカーがなぜその回答を選んだかの理由をそのワーカー自身が入力したものを提示する場合について、比較を行った (図 6)。

(3) 他者の回答の品質より優秀な (品質の高いタスク結果をもたらす) ワーカーの回答を第 2 タスクで提示した場合の影響を調べた。具体的には、タスクの正答率が 90% 以上のワーカーの回答だけを見せた場合と、ワーカーの正答率を気にせずに提示した場合について、比較した。

4. タスク割り当て手法

3.2 節で説明した 3 つの要素のうち、(3) において優秀なワーカーを第 2 タスクに割り当てる方法に関しては自明でないため、本節で説明する。

4.1 問題設定

タスク集合を $T = \{t_1, t_2, \dots, t_q\}$ 、ワーカー集合を $W = \{w_1, w_2, \dots, w_h\}$ とする。ワーカー集合の要素 $w_j \in W$ が優秀なワーカーであるかを真理値で返す関数を $isExpert$ とする。 $isExpert$ 関数が真を返すようなワーカー (以下、*Expert* ワーカーと呼ぶ) の集合を E 、偽を返すようなワーカー (以下、*Novice* ワーカーと呼ぶ) を N とする。さらに、ワーカー $w_j \in W$ とワーカー $w_k \in W$ のペアにタスク $t_i \in T$ をタスク割り当てることを $assignment(t_i, w_j, w_k)$ と表す。このタスク割り当て $assignment$ の集合を $Assignment$ とする。

本問題の目的は、タスク集合 T とワーカー集合 W が与えられた時に、タスク割り当て集合 $Assignment = \{(t_i, e, n) \mid t_i \in T, e \in E, n \in N\}$ を出力することである。

ワーカー w_j は第 1 タスクのみを処理し、ワーカー w_k は第 1 タスクと第 2 タスクを処理する。その際、ワーカー w_k が処理する第 2 タスクにおいて、ワーカー w_j の第 1 タスクの結果を提示する。

4.2 初期設定

ここでは、初期化の手続きについて説明する。初期化の手続きを Algorithm 1 に示す。

Expert ワーカーの集合 E に対してタスク $t_i \in T$ を割り当てる。その際、*Expert* ワーカーの集合 E は空集合でないものとする。*Submit* は与えられたタスク t_i を、与えられたワーカー集合またはワーカーが回答できる状態にする手続きである。

4.3 能力に基づくワーカーの分類

ここでは、能力に基づいてワーカーを *Novice* ワーカーと *Expert* ワーカーに分類する手法を説明する。ワーカーの能力に基づいて、与えられたワーカー W 集合の要素 $w \in W$ が *Expert* ワーカーであるかを真理値で返すブール値関数 $isExpert$ を式 1 に示す。

$$isExpert : w \rightarrow Boolean \quad (1)$$

Expert ワーカーの集合 E は、 $isExpert$ 関数が *true* を出力するワーカー集合 W の要素の集合である。 $isExpert$ 関数を用いて、*Expert* ワーカーの集合 E を式 2 のように表す。

$$E = \{w \in W \mid isExpert(w)\} \quad (2)$$

同様に、*Novice* ワーカーの集合 N は、 $isExpert$ 関数が *false* を出力するワーカー集合 W の要素の集合である。 $isExpert$ 関数を用いて、*Novice* ワーカーの集合 N を式 3 のように表す。

Algorithm 1 初期化の手続き

```

1:  $Assignment \leftarrow \{(t_i, \emptyset, \emptyset) \mid t_i \in T\}$ 
2: for  $(t_i, \emptyset, \emptyset) \in Assignment$  do
3:    $Submit(t_i, E)$ 
4: end for

```

$$N = \{w \in W \mid \neg isExpert(w)\} \quad (3)$$

$isExpert$ 関数が $true$ を返す条件には、任意のワーカ能力推定手法を用いることが出来る。

4.3.1 回答既知タスクを用いる方法

ここでは、ゴールドスタンダード質問の正答率を用いてワーカの能力を推定する方法を説明する。ゴールドスタンダード質問は問題に対する正答が明らかなタスクである。ゴールドスタンダード質問に対する回答を正答と照らし合わせてワーカの正答率を求めることで、ワーカの能力を推定する。

ワーカが回答したゴールドスタンダード質問のタスク結果の集合 G と、 G の正答の集合 A が与えられたとき、 G の正答率に基づいて、ワーカの能力を式 4 で推定する。

$$p = \frac{|\{g_i \in G \mid g_i = a_i, a_i \in A\}|}{|G|} \quad (4)$$

$isExpert$ 関数の出力を、式 4 が返すワーカの能力をしきい値 M を用いて式 5 のように場合分けをする。

$$isExpert = \begin{cases} true & (M < p \leq 1) \\ false & (0 \leq p < M) \end{cases} \quad (5)$$

ここでは、ゴールドスタンダード質問の正答率を用いてワーカの能力を推定する方法を説明した。

4.4 ワーカがタスクに回答したときの動作

ここでは、ワーカがタスクに回答したときの動作について説明する。ワーカ w_j がタスク t_i したときに実行する手続きを Algorithm 2 に示す。

1 行目から 5 行目では、ワーカ w_j が処理したタスク t_i に関する $assignment$ の更新をする。6 行目から 11 行目では、まずワーカ w_j が $Expert$ ワーカであるかを $isExpert$ 関数で判定する。次に、 $Expert$ ワーカである場合には、ワーカ w_j に対して $Expert$ ワーカ集合 E に割り当てたタスクが回答できる状態にする処理を行う。

5. 評価実験

本節では、3.2 節で説明した 3 つの要素が、ワーカの成長にどのように影響するかを調査する。まず、5.1 節で評価実験に用いるデータセットについて説明する。次に、5.2 節で実験の流れを説明し、5.3 節でワーカの成長を評価する方法について説明する。最後に 5.4 節で実験結果について説明する。

5.1 データセット

評価実験では、データセットを用いて、回答が既知のタスクを作成する。データセットには、200 種類の鳥類の画像が含まれている CaltechUCSD Birds 200 [13] を用いた。次に定義する 3 つのルールに基づいて鳥の種類を選出することで、異な

Algorithm 2 Perform

Input: w_j, t_i

Output: $Assignment \subseteq T \times W \times W$

```

1: if  $(t_i, \emptyset, \emptyset) \in Assignment$  then
2:    $Assignment \leftarrow Assignment - \{(t_i, \emptyset, \emptyset)\} \cup \{(t_i, w_j, \emptyset)\}$ 
3: else
4:    $Assignment \leftarrow Assignment - \{(t_i, w_k, \emptyset)\} \cup \{(t_i, w_k, w_j)\}$ 
5: end if
6: if  $(isExpert(w_k))$  then
7:    $E \leftarrow E \cup \{w_k\}$ 
8:   for  $(t_i, \emptyset, \emptyset) \in Assignment$  do
9:      $Submit(t_i, \{w_k\})$ 
10:  end for
11: end if

```

るデータを用いた複数のタスクセット (Hard1,2,3 Normal1,2,3 Hard1,2,3) を作成する。

【Hard 1, 2, 3】

Sibley 分類体系において目階級が等しい種類を選び、容姿が類似するものを目視で選択する。

【Normal 1, 2, 3】

Sibley 分類体系において目階級が等しい種類を選ぶ。

【Easy 1, 2, 3】

Sibley 分類体系において目階級で異なる種類を選ぶ。

5.2 手順

評価実験で用いるタスクでは、選択式の 4 択問題を扱う。タスクは、質問文と選択肢が与えられ、選択肢から 1 項目を選び回答とするものである。ワーカがタスクに連続で取り組む際、タスクの質問文のみが変化し、選択肢の内容は変化しないものとする。

評価実験では、ワーカは次の順番でタスクに取り組む。

- (1) 評価タスク (前) に 5 回取り組む
- (2) 自己補正タスクに 10 回取り組む
- (3) 評価タスク (後) に 5 回取り組む

評価タスクはワーカの成長度合いを測定するためのタスクである。ワーカが実験のためのタスクに取り組む直前 (以降、評価タスク (前) と呼ぶ) と直後 (以降、評価タスク (後) と呼ぶ) に行う。評価タスク (前) と評価タスク (後) では同じ質問を扱う。評価タスクの正解率を算出することで、ワーカの成長度合いを測定する。

3 章で説明したワーカの成長に影響を与えられと考えられるタスクの要素について表 1 で整理する。本稿では、タスクの情報として選択肢の画像を、他者回答の情報として回答の理由を用いた。自己補正タスクで提示する他者の回答やコメントには、実験に用いるタスクを評価実験の対象ワーカ集合とは異なるワーカ集合に委託し、そのワーカ集合から得たタスク結果を用いる。

上記の方法により作成したタスクは、Yahoo!クラウドソーシング^(注1)を通じてワーカに委託する。

(注1) : <http://crowdsourcing.yahoo.co.jp/>

表 1 タスクのパターン

パターン	タスクの情報	他者回答の情報	他者回答の品質
A	画像あり	コメントなし	条件なし, 正答率 90%以上
B	画像あり	コメントあり	条件なし, 正答率 90%以上
C	画像なし	コメントなし	条件なし, 正答率 90%以上
D	画像なし	コメントあり	条件なし, 正答率 90%以上

表 2 パターン A の実験結果

タスクの難易度	他者回答	評価 (前)	評価 (後)	成長度合い
Hard 1	0.511	0.728	0.781	0.052
	0.970	0.731	0.785	0.053
Hard 2	0.514	0.779	0.802	0.023
	0.964	0.808	0.816	0.008
Normal 1	0.628	0.829	0.804	0.025
	0.921	0.874	0.800	0.074
Normal 2	0.682	0.776	0.786	-0.010
	0.920	0.860	0.880	-0.020
Normal 3	0.701	0.797	0.784	0.012
	0.900	0.840	0.787	0.053
Easy 1	0.941	1	0.989	0.011
	0.941	1	0.989	0.011
Easy 2	0.943	0.988	0.996	-0.008
	0.943	0.988	0.996	-0.008
Easy 3	0.970	0.975	0.995	-0.020
	0.986	0.975	0.995	-0.020

5.3 評価

ワーカの成長を測定する方法を説明する。評価実験では、回答が既知の問題を扱うため、タスクの正答率を求めることが可能である。評価タスク (前) と評価タスク (後) の正答率から、ワーカの成長度合いを式 6 で求める。

$$\text{成長度合い} = \text{評価タスク後の正答率} - \text{評価タスク前の正答率} \quad (6)$$

5.4 結果

5.4.1 パターン A

パターン A のタスクでの実験結果を表 2 に示す。1 列目はタスクの難易度を、2 列目は自己補正の第 2 タスクで提示した他者回答の平均を、3 列目は評価タスク (前) の正答率の平均を、4 列目は評価タスク (後) の正答率の平均を、5 列目は成長度合いの平均を表している。各行は難易度ごとの実験結果を表しており、各難易度の 1 行目は全体での集計結果を、2 行目は他者の回答の正答率が 0.9 以上の結果のみを集計したものである。

難易度 Easy の結果と、難易度 Normal, 難易度 Hard を比較すると、後者の方がワーカの成長を確認できていることが分かる。理由として、簡単なタスクでは評価タスク (前) での正答率が高いため、結果としてワーカの成長度合いが小さくなることが考えられる。以降の実験では難易度 Hard のみを扱う。

5.4.2 パターン B

パターン B のタスクでの実験結果を表 3 に示す。表の見方は 5.4.1 節で説明した表 2 についての説明と同様である。

難易度 Hard 1 では Expert ワーカの回答を提示することで

表 3 パターン B の実験結果

タスクの難易度	他者回答	評価 (前)	評価 (後)	成長度合い
Hard 1	0.775	0.770	0.814	0.043
	0.903	0.693	0.843	0.15
Hard 2	0.754	0.853	0.856	0.003
	0.921	0.914	0.871	-0.043

表 4 パターン C の実験結果

タスクの難易度	他者回答	評価 (前)	評価 (後)	成長度合い
Hard 1	0.471	0.673	0.740	0.066
	0.964	0.655	0.818	0.164
Hard 2	0.488	0.621	0.666	0.045
	0.971	0.529	0.614	0.086
Hard 3	0.479	0.686	0.686	0
	0.970	0.520	0.6	0.08

表 5 パターン D の実験結果

タスクの難易度	他者回答	評価 (前)	評価 (後)	成長度合い
Hard 1	0.685	0.697	0.699	0.003
	0.9	0.621	0.588	-0.033

成長度合いの改善が確認できた。一方で、難易度 Hard 2 では、Expert の回答を提示しても成長度合いの改善が確認できない。この結果から、他者のコメントの品質がワーカの成長に大きく影響していると考えられる。

5.4.3 パターン C

パターン C のタスクでの実験結果を表 4 に示す。表の見方は 5.4.1 節で説明した表 2 についての説明と同様である。

パターン C はすべての実験においてワーカの成長度合いが向上したことが分かる。

5.4.4 パターン D

パターン D のタスクでの実験結果を表 5 に示す。表の見方は 5.4.1 節で説明した表 2 についての説明と同様である。

6. 考察

6.1 タスクで表示する参考情報の量

タスクで提示する参考情報を増やした場合、多くの例においてワーカの成長を確認できた。成長が確認できなかった例においても、大幅に正答率が減少している例は確認できなかった。

6.2 フィードバックで提示する情報の量

タスクに取り組むワーカへのフィードバックとして、他ワーカが回答した際の判断理由を提示することで、ワーカの成長が見られることを、評価実験のパターン B で確認した。ただし、評価実験のパターン B でのタスク難易度 Hard 1 の結果はワーカの成長が見られるのに対し、タスク難易度 Hard 2 の結果からはワーカの成長が見られない。このことから、提示する判断理由の品質がワーカの成長に大きな影響を与えていると考えられる。今回の実験で収集したコメントを観察した所、有用であると考えられるコメントに「頭の毛の模様が似ているから」「くちばしの先端が尖っている」などがあり、有用でないと考えられるコメントとして「なんとなく」「雰囲気で」などが存在した。そこで、タスクに取り組むワーカへのフィードバックとし

て、他ワーカーが回答した際の判断理由を提示する場合、提示する判断理由を選別することが求められる。

6.3 他者回答の品質

信頼できるワーカーの回答を自己補正の第2タスクで提示することで、ワーカーが回答精度が改善されることをパターンCの結果から確認した。選択肢の画像を示すパターンAよりも、選択肢の画像を示さないパターンCにおけるワーカーの成長度合いが高くなる理由として、選択肢で示す画像がすべての質問において適切とは限らないことが挙げられる。今回の評価実験では鳥の種類を分類する問題を用いたが、同一の種類の鳥であっても、(1)雌雄による特徴の違い、(2)成長の度合いによる特徴の違い、などの要因により選択肢で示す1枚の画像でその種類の特徴すべてを網羅するのは困難である。

6.4 タスクの難易度

パターンAのタスクについて、複数の難易度による実験の結果(表2)から、難易度を低く設定したタスクセットではEasyやNormalの実験ではワーカーの成長が確認出来なかった。大半のワーカーが高い正答率となるようなタスクでは、評価タスク(前)と評価タスク(後)の正答率の差が小さくなるためである。

7. まとめと今後の課題

本稿では、クラウドソーシングにおいてワーカーの能力を向上させることで、タスク結果の品質を向上させる問題に取り組んだ。一連のタスクを行った後に、高い品質のタスク結果をもたらすExpertワーカーの割合を増やすことを目標とし、自己補正タスクを複数回行い際に、どのような要素がワーカーの能力を向上させるのに影響するかを調査した。提案手法は、自己補正の第2タスクでExpertワーカーの回答を提示するためのタスク割り当てを行うものである。評価実験では、本研究で注目したタスクに関する3つの要因について次のような知見を得た。

【タスクに関する情報】

タスクに関する情報は多いほうが、ワーカーがより成長することが示唆された。ただし、ワーカーが判断に用いる材料が増えると、リクエストの意図する判断基準とは異なる情報に基づいて判断すると考えられ、強調したい情報がある場合は、それを目立たせる、それ以外の情報を排除する等の工夫が必要である。

【他ワーカーの回答に関する情報】

他ワーカーの回答に関する情報として、回答の理由を提示する場合としない場合を比較したところ、提示する回答理由の品質によつては提示しない場合と同等の成長にとどまることが示唆された。よつて、回答の理由を提示する場合、信頼できる回答の理由を選別するとの処理を行うことで、よりワーカーの成長に貢献できると考えられる。

【他者の回答の品質】

優秀なワーカーの回答を提示することは、ワーカーの成長の良い影響があることが評価実験で示唆された。評価実験では、優秀なワーカーの回答を提示したことで、ワーカーの成長に悪影響を与える例は確認できず、他の要因と組み合わせることでよりワーカーを成長させることが出来ると考えられる。

今後の課題として、(1)別のデータセットや条件による評価

実験、(2)回答が既知のタスクが存在しない場合に優秀なワーカーの回答を提示する手法の検討がある。

謝 辞

本研究の一部は、JST CREST および JSPS 科研費(#25240012)による。

文 献

- [1] Nihar Shah and Denny Zhou. No oops, you won't do it again: Mechanisms for self-correction in crowdsourcing. pp. 1–10, June 2016.
- [2] Masayuki Ashikawa, Takahiro Kawamura, and Akihiko Ohsuga. Proposal of grade training method in private crowdsourcing system. In *Third AAAI Conference on Human Computation and Crowdsourcing*, 2015.
- [3] Pinar Donmez, Jaime G Carbonell, and Jeff Schneider. Efficiently learning the accuracy of labeling sources for selective sampling. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 259–268. ACM, 2009.
- [4] Shayan Doroudi, Ece Kamar, Emma Brunskill, and Eric Horvitz. Toward a learning science for complex crowdsourcing tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 2623–2634. ACM, 2016.
- [5] 清水伸幸, 山下達雄, 塚本浩司, 颯々野学. クラウドソーシングにおける成果物の品質維持のためのダミー問題出題手法の検討. 言語処理学会第20回年次大会, 2014.
- [6] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.
- [7] Jacob Whitehill, Ting fan Wu, Jacob Bergsma, Javier R. Movellan, and Paul L. Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pp. 2035–2043. Curran Associates, Inc., 2009.
- [8] Peter Welinder, Steve Branson, Pietro Perona, and Serge J. Belongie. The multidimensional wisdom of crowds. In J. Lafferty, C. Williams, J. Shawe-taylor, R.s. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pp. 2424–2432. 2010.
- [9] Manas Joglekar, Hector Garcia-Molina, and Aditya Parameswaran. Evaluating the crowd with confidence. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pp. 686–694, New York, NY, USA, 2013. ACM.
- [10] Manas Joglekar, Hector Garcia-Molina, and Aditya G. Parameswaran. Comprehensive and reliable crowd assessment algorithms. *CoRR*, Vol. abs/1411.3377, , 2014.
- [11] H. Jiang and S. Matsubara. Improving crowdsourcing efficiency based on division strategy. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, Vol. 2, pp. 425–429, Dec 2012.
- [12] Naoki Otani, Yukino Baba, and Hisashi Kashima. Quality control of crowdsourced classification using hierarchical class structures. *Expert Systems with Applications*, Vol. 58, pp. 155 – 163, 2016.
- [13] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.