

クラウドソーシングにおける 長期労働が及ぼす作業能力の変化

松田 義貴¹ 鈴木 優¹ 中村 哲¹

概要：本研究では、クラウドソーシングにおけるワーカーの作業能力の変化について分析を行う。既存の研究では、ワーカーの作業能力は変化しないものとして扱われてきた。しかし、心理学の分野では長時間の労働は作業品質の低下をもたらすと言われているが、クラウドソーシングにおいても同様のことが言えるのかは分かっていない。そこで、我々はワーカーの作業時間に着目し、長時間の労働が及ぼす作業能力の変化を分析する。本研究により、ワーカーが長時間連続で作業を続けた場合でも、正答率は低下しないことが分かった。作業開始直後の正答率が悪いワーカーに限っては正答率が向上する場合があった。この正答率が向上したワーカーは再び正答率が低下することはなく、安定した正答率で作業を続けた。また、ワーカーが長時間連続で作業を続けた場合でも、タスクの処理時間は低下しないことが分かった。一方で、長期間にわたって作業を行うとタスク処理時間は向上するという傾向が見られた。今回得られた知見は今後、ワーカーの能力の推定やタスク設計に活かすことが期待できる。

Change in worker quality exerted by long-term work in crowdsourcing

Yoshitaka Matsuda¹ Yu Suzuki¹ Satoshi Nakamura¹

1. はじめに

クラウドソーシングとはインターネットを介して不特定多数の人々にタスクを依頼し、作業報酬として金銭的価値を支払う仕組みである。このクラウドソーシングの活用事例として機械学習のためのラベル付きデータの収集がある [1]。以後、話を簡単にするため、マイクロタスク型のラベル付け作業を話題にして議論を進める。例えば、ワーカーはある製品に関するツイートを読み、投稿者がその製品に対してポジティブな意見を持っていればポジティブのラベルを付与するような評判分析のためのタスクである。

このクラウドソーシングにおける問題点は、作業結果の品質が担保されないことである。ワーカーと呼ばれる作業員の中には、タスクの難易度によっては正しいラベルをつけることができないワーカーや金銭目的など悪意を持ったスパムワーカーが存在する。このようなワーカーの存在によって、必

ずしも正しいラベルを付けることができるとは限らない。

そこで、できる限り正しいラベル付けをする手法としてワーカーの冗長化が挙げられる。ワーカーの冗長化は、同一のタスクに対して複数人のワーカーを割り当て、それぞれのワーカーが付けたラベルを統合し、最終的なラベルを導く手法である。最も単純な統合手法としては、多数決によってラベルを決定する手法が挙げられる [2]。しかし、マイクロタスク型のクラウドソーシングでは、1人のワーカーが複数のタスクを行うことが一般的である。この特徴によってワーカーの作業能力を推定することが可能になり、ワーカーの能力を考慮した統合手法が提案されている。Dawid と Skene らは EM アルゴリズムによってワーカーの能力と正しいラベルを交互に推定する手法を提案した [3]。Whitehill や Welinder もワーカーの能力を考慮したラベルの統合手法を提案している [4][5]。これらの研究では、ワーカーの能力は常に一定であると仮定している。

しかし、心理学の分野では、長時間の労働を行うことは疲労や集中力の低下から作業品質の低下をもたらすと言わ

¹ 奈良先端科学技術大学院大学 情報科学研究科
Graduate School of Information Science, Nara Institute of
Science and Technology

れている [6]. クラウドソーシングにおいて, ワーカが長時間労働を行うことでワーカの作業能力が低下することは分かっている. また, ワーカがタスクに慣れることでワーカの作業能力が向上することも分かっている. ワーカの作業能力が変化するならば, その変化を読み取り, ラベルの統合手法に反映させることによって, 既存の研究よりも高精度な統合ができることは明らかである. そのためにはまず, ワーカの作業能力が変化するかどうか分析することが必要である.

そこで, Hata らはワーカの作業能力の変化に関する分析を行った [7]. Hata らはワーカは安定した正答率でタスクに取り組むと述べている. 一方で, タスクの処理時間は短縮されるという分析結果を示した. つまり, クラウドソーシングにおいては正答率としての作業能力は変化しないが, タスク処理時間としての作業能力は向上することが分かった. この研究の前提条件では, ワーカは自由に作業をやめることができるようにタスクが設計されている. しかし, Hata らは作業を早く終えたワーカの作業も, 長時間連続で作業を続けたワーカの作業も, 同一の尺に正規化し, 分析を行っている. また, ワーカは連続して作業を行っているものとしているが, 現実的には, ワーカは休憩を挟むことや複数の日にわたって作業を行っている. 疲労や集中力の低下などは長時間連続で作業を行うことによって生じるため, 疲労や集中力の低下など人間の特徴を正確に捉えることができていない可能性があると考えた.

そこで本研究では, ワーカの作業時間に着目し, 長時間の労働が及ぼすワーカの作業能力の変化について分析を行う. 本研究により, ワーカが長時間連続で作業を続けた場合でも, 正答率は低下しないことが分かった. この結果は心理学の知見に反しているが, 強制的な労働ではないというマイクロタスク型のクラウドソーシングの特徴が現れたと言える. つまり, ワーカは疲労が蓄積したり集中力が低下したりする前に, 作業から離反することが分かった. 一方で, 正答率が向上するワーカが存在を確認できた. そのワーカとは作業開始直後の正答率が悪いワーカである. また, 正答率が向上したワーカは再び正答率が低下することではなく, 安定した正答率で作業を続けた. 処理時間に関しても同様で, マイクロタスク型のクラウドソーシングの特徴からタスクの処理時間も低下しないことが分かった. また, 長期間にわたって作業を行うとタスク処理時間は向上する傾向が見られた.

2. クラウドソーシングのタスク

我々はワーカが長期で労働することによるワーカの作業能力の変化を分析する. 本章ではまず, この分析のために行ったクラウドソーシングのタスクについて説明する. また, 我々が行ったタスクによって収集した作業結果の数についても述べる.

我々が行ったタスクはマイクロタスク型である. 我々はあるスマートフォンに関連したキーワードを含むツイートを取得し, ワーカにそのツイートへのラベル付けを課した. ワーカは一つ一つのツイートに対して以下に示す五つのラベルのいずれかを付与する.

- **Positive**: ツイート投稿者がスマートフォンに対して何か具体的に良い部分があると思っている.
- **Negative**: ツイート投稿者がスマートフォンに対して何か具体的に悪い部分があると思っている.
- **Positive & Negative**: ツイート投稿者がスマートフォンに対して何か具体的に良い部分と悪い部分の両方があると思っている.
- **Neutral**: ツイート投稿者がスマートフォンに対して何か感じているが良い悪いの判断ができない. もしくは良い悪いと感じているが具体的な部分がない.
- **NA**: ツイート投稿者がスマートフォンに対して何も感じていない. 関係がない. 意味がわからない. 広告など.

例えば, 「スマートフォンは画面が綺麗だな」というツイートに対して, ワーカは Positive を選択しなければならない. 「スマートフォンに変えようかな」というツイートは, スマートフォンについて良いと考えていることが予想できるが, 具体的に何について良いと考えているか分からないので, ワーカは Neutral を選択する必要がある.

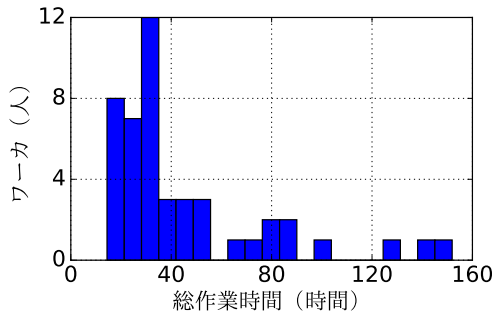
我々は, 合計 250,164 ツイートを Twitter API を用いて収集した. さらに, 1 ツイートずつタスクに従事できるプラットフォームを構築し, 一つのツイートのラベル付けに最低 5 人のワーカを割り当てた. 合計 1,009 人のワーカがこのタスクに従事し, のべ 1,250,923 の作業結果を得た.

3. 分析

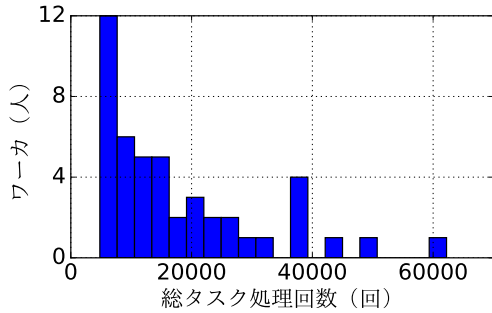
我々は長期労働におけるワーカの作業能力の変化について分析を行う. 本章では, その分析手法とその結果について述べる. まず, 長期労働の定義を行う. 次に, 作業能力の一つとして考える正答率に関する分析手法とその結果について説明する. その後, もう一つの作業能力として考えるタスクの処理時間に関する分析手法とその結果について説明する.

3.1 長期労働の定義

長期ワーカ労働と長期一連作業労働の 2 種類の長期労働を定義する. 長期ワーカ労働とは長期労働者による労働である. 長期労働者とは総作業時間と総タスク処理回数が共に上位 5% に含まれるワーカのことを指す. 長期一連作業労働とは長期一連作業内の労働である. 長期一連作業とは一連作業内の作業時間とタスク処理回数が共に上位 5% に



(a) 作業時間の分布



(b) タスク処理回数の分布

図 1: 長期ワーカー労働の作業時間とタスク処理回数の分布

含まれる一連作業のことを指す。

ここで、一連作業を定義する必要がある。堀江の研究では、60 分間の VDT (Visual Display Terminals) 作業に対して 10 分間の休憩を取ることが、作業者の心身諸反応および作業効率の観点から効果的であることを示している [8]。本研究で行ったクラウドソーシングのタスクも VDT 作業であり、10 分以上の休憩がワーカーのリフレッシュに繋がると考えられる。そこで本研究では、タスク処理間隔が 10 分以内の連続したタスク処理を一連作業とする。

長期ワーカー労働に該当するワーカーは 46 ワーカー存在し、全ワーカーの 4.6% であった。この 46 ワーカーの作業時間とタスク処理回数の分布はそれぞれ図 1 の通りである。また、長期一連作業労働に該当する作業は 255 作業存在し、一連作業の 3.1% であった。この 255 作業の作業時間とタスク処理回数の分布はそれぞれ図 2 の通りである。

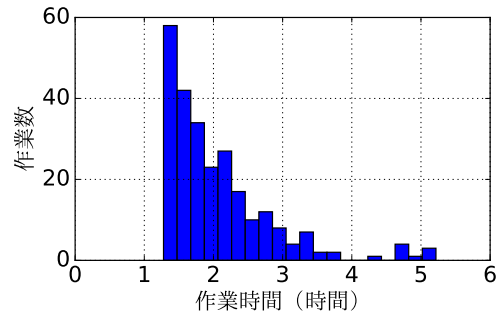
3.2 正答率の変化

本節では、ワーカーの長期労働における正答率の変化に関する分析を行う。まず、分析手法について説明する。その後、長期ワーカー労働と長期一連作業労働それぞれの分析結果について説明する。

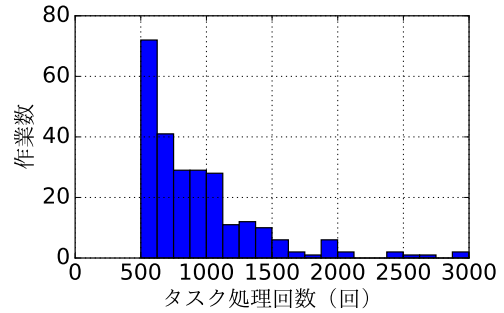
3.2.1 分析手法

任意の期間におけるタスクの正答率 P を式 (1) のように定義する。

$$P = \frac{N_{correct}}{N} \quad (1)$$



(a) 作業時間の分布



(b) タスク処理回数の分布

図 2: 長期一連作業労働の作業時間とタスク処理回数の分布

ここで、 N は任意の期間内のタスク処理回数、 $N_{correct}$ は同じ期間内に正しい選択肢を選んだタスク処理回数である。本研究で行ったタスクには、正解の選択肢があらかじめ用意されているわけではないので、正解の選択肢を決定する必要がある。そこで、2 章で説明した通り、同一タスクに対して最低 5 人のワーカーが取り組んでおり、5 人の多数決によって正解の選択肢を決定する。

長期労働期間中における最初の 150 タスクの正答率を P_{first} 、最後 10% のタスクの正答率 P_{last} とする。我々はこの P_{first} と P_{last} の差分 P_{tranc} を正答率の変化と考え、以下の式 (2) で定義する。

$$P_{tranc} = P_{first} - P_{last} \quad (2)$$

例えば、最初 150 タスクの正答率 P_{first} が 80% で、最後 10% のタスクの正答率 P_{last} が 90% の場合、差分 P_{tranc} は -10% となる。このように $P_{tranc} < 0$ のとき、正答率が向上していることを表す。逆に、 $P_{tranc} > 0$ のときは正答率が低下、 $P_{tranc} = 0$ のときは正答率が変化していないと考えることができる。

次に、最初 150 タスクの正答率を用いる理由を述べる。最初のタスク数を F 、差分の絶対値を $|P_{tranc}|$ とし、 F と $|P_{tranc}|$ の平均の関係を表したグラフを図 3 に示す。約 150 タスクを境に差分の絶対値が変化しなくなっていることが分かる。よって、150 タスク処理した時点で、ワーカーの正答率を判断することは妥当であると言える。

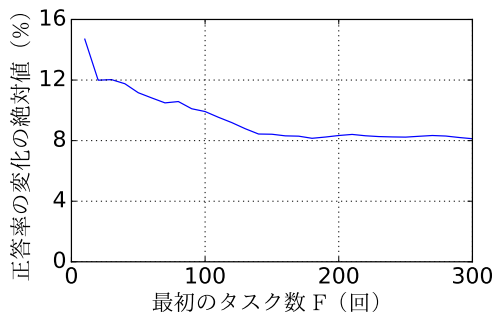


図 3: 最初のタスク数 F と差 P_{tranc} の絶対値の関係

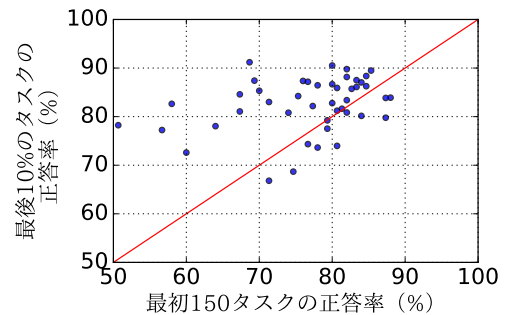
3.2.2 分析結果

まず、長期ワーカー労働が及ぼす正答率の変化についての分析を行う。長期ワーカー労働を行った 46 ワーカーそれぞれの最初 150 タスクの正答率と最後 10%のタスクの正答率を図 4a に示す。一つ一つの点はワーカーを表し、対角線を結ぶラインは $P_{tranc} = 0$ を表している。対角線を結ぶラインより上側のワーカーは $P_{tranc} > 0$ となり、正答率が上昇していると判断することができる。反対に、対角線を結ぶラインより下側のワーカーは $P_{tranc} < 0$ となり、正答率が下降していると判断できる。最初 150 タスクの正答率が低いワーカーの多くは、最後 10%の正答率が大幅に向上していることが見て取れる。一方で、最初 150 タスクの正答率が高いワーカーの正答率には大きな変化が見られない。

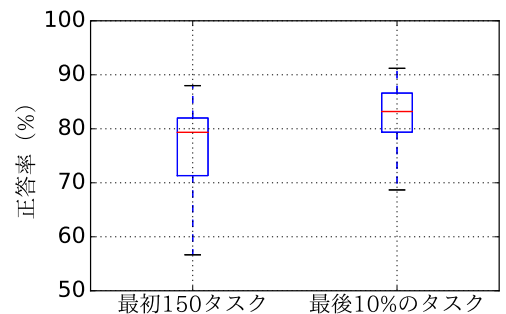
さらに、最初 150 タスクの正答率と最後 10%のタスクの正答率の分布を図 4b に示す。左側の箱ひげ図が最初 150 タスクの正答率の分布、右側の箱ひげ図が最後 10%のタスクの正答率の分布である。正答率の中央値が高くなっており、正答率の低いワーカーの底上げも確認できる。

また、最初 150 タスクの正答率と最後 10%のタスクの正答率の差分 P_{tranc} の分布を図 4c に示す。約 7 割のワーカーは $\pm 10\%$ 以下の変化しか見られず、絶対値の平均は 8.4% の変化であった。

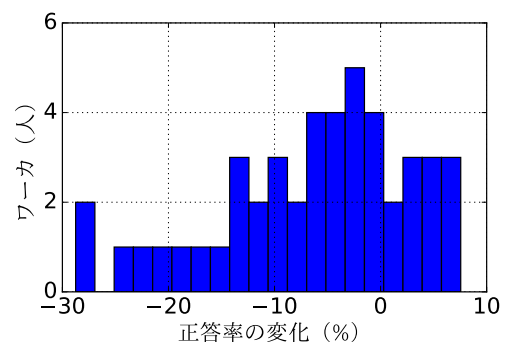
特徴的なワーカーの正答率の変化を図 5 に示す。横軸がタスク処理回数、縦軸が直近 200 タスクの正答率である。図 5a のワーカーは正答率を大きく向上させたワーカーの 1 人である。作業開始直後は 70%程度の正答率であったが、タスク処理回数が増えるにつれて、正答率が向上している。正答率が 80%に到着して以降は、90%との間で小刻みな変化を繰り返している。正答率が向上するワーカーの多くはこのワーカーのように、作業開始直後から一気に正答率が向上し、ある程度の正答率になってからは正答率が大きく変化することがなかった。図 5b は正答率の変化が小さかったワーカーの 1 人である。短期的には正答率が上下しているが、長期的な傾向は見られなかった。ほとんどのワーカーの正答率は図 5b のワーカーのように、短期的な変化だけで長期的な変化がなかった。また、長期的に正答率が大幅に低下する



(a) 各ワーカーの最初 150 タスクと最後 10%のタスクの正答率



(b) 各ワーカーの最初 150 タスクと最後 10%のタスクの正答率の分布



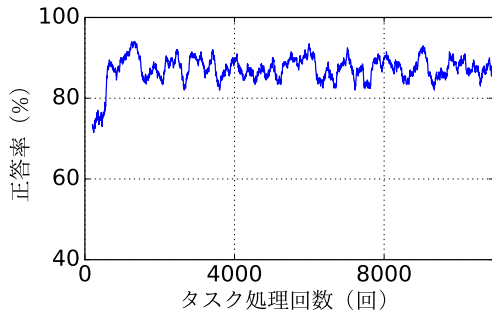
(c) 各ワーカーの最初 150 タスクと最後 10%のタスクの正答率の差の分布

図 4: 長期ワーカー労働の正答率の変化

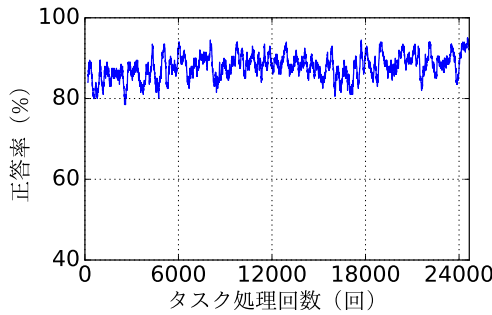
ワーカーは存在しなかった。

次に、長期一連作業労働が及ぼす正答率の変化についての分析を行う。長期一連作業労働である 255 作業それぞれの最初 150 タスクの正答率と最後 10%のタスクの正答率を図 6a に示す。それぞれの点は各作業を表す。最初 150 タスクの正答率が低い作業では、最後 10%のタスクの正答率が向上しているが、最初 150 タスクの正答率が高い作業では、向上や低下する作業もあるが全体的には変化していない作業が多かった。

さらに、最初 150 タスクの正答率と最後 10%のタスクの正答率の分布を図 6b に示す。左側の箱ひげ図が最初 150 タスクの正答率の分布、右側の箱ひげ図が最後 10%のタスクの正答率の分布である。同じような分布の形をしてお



(a) 正答率が向上するワーカー例



(b) 正答率が変化しないワーカー例

図 5: 長期ワーカー労働における正答率の変化のワーカー例

り、全体的な正答率の変化は見られなかった。

図 6c に示すように、最初 150 タスクの正答率と最後 10%のタスクの正答率の差分 P_{tranc} は、半数以上の作業で $\pm 5\%$ 以下であり、絶対値の平均も 5.8%とほとんどの作業で、大きな正答率の変化は見られなかった。

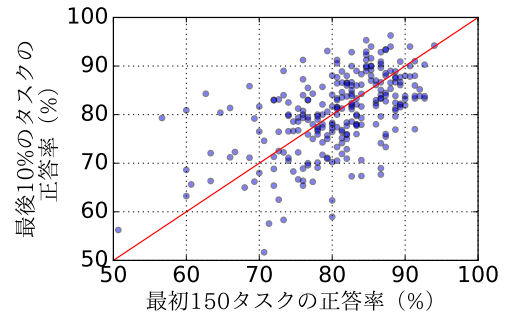
特徴的な長期一連作業労働中の正答率の変化を図 7 に示す。横軸がタスク処理回数、縦軸が直近 50 タスクの正答率である。図 7a の長期一連作業労働は正答率が向上した例であり、作業開始から徐々に正答率が向上している。長期ワーカー労働における結果から、このワーカーはその後、向上後の正答率付近での作業が見込まれる。図 7b の長期一連作業労働では、800 タスクをすぎた頃から正答率が急激に低下している。ほとんどの作業では、図 7c のように、長期的な正答率の変化はなく、安定した作業を行っていた。

3.3 タスク処理時間の変化

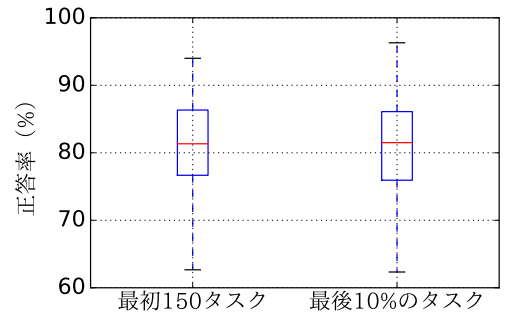
本節では、ワーカーの長期労働におけるタスク処理時間の変化に関する分析を行う。まず、分析手法について説明する。その後、長期ワーカー労働と長期一連作業労働それぞれの分析結果について説明する。

3.3.1 分析手法

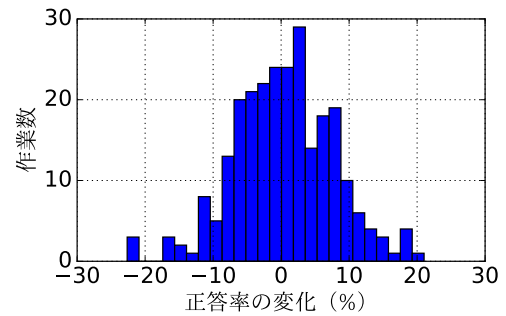
まず、ワーカーが行ったのべ 1,250,923 タスクの 1 タスクあたりのタスク処理時間の分布を図 8 に示す。第一四分位数、中央値、第三四分位数はそれぞれ、4 秒、6 秒、9 秒であった。四分位範囲の 1.5 倍の 16 秒を越える処理時間



(a) 各作業の最初 150 タスクと最後 10%のタスクの正答率



(b) 各作業の最初 150 タスクと最後 10%のタスクの正答率の分布



(c) 各作業の最初 150 タスクと最後 10%のタスクの正答率の差の分布

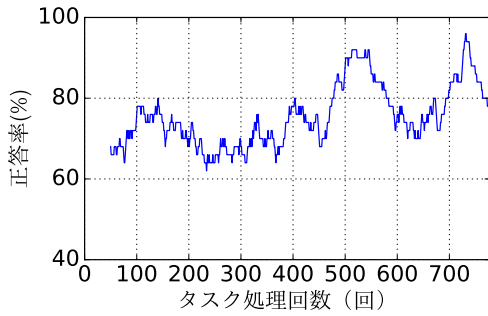
図 6: 長期一連作業労働における正答率の変化

については外れ値として扱う。なぜならば、1 タスクの処理時間が 16 秒以上必要であったタスクは、タスクの難易度が他のタスクに比べ高く、多くの処理時間が必要であった、もしくは、作業画面を開いたまま小休止を挟んだことによって正味の処理時間ではなかったことが考えられるためである。

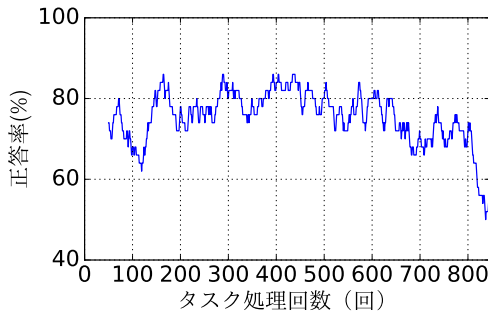
任意の期間における 1 タスクあたりの平均処理時間 T を式 (3) のように定義する。

$$T = \frac{1}{|W|} \sum_{t \in W} T_t \quad (3)$$

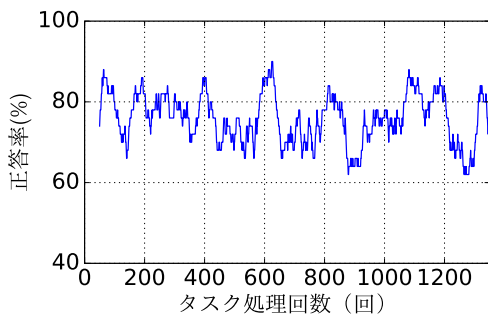
ここで、 W は任意の期間内におけるタスク処理時間が外れ値でないタスクの集合、 t は各タスク、 T_t はタスク t の処



(a) 正答率が向上する作業例



(b) 正答率が低下する作業例



(c) 正答率が変化しない作業例

図 7: 長期一連作業労働における正答率の変化の作業例

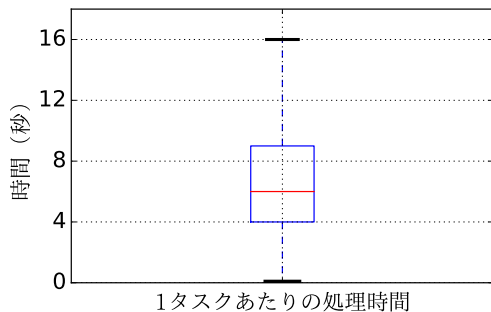


図 8: 1 タスクあたりの処理時間の分布

理時間である。

長期労働期間中における最初の 150 タスクの平均処理時間を T_{first} , 最後 10% のタスクの平均処理時間 T_{last} とす

る。我々はこの T_{first} と T_{last} の差分 T_{tranc} を処理時間の変化と考え、以下の式 (4) で定義する。

$$T_{tranc} = T_{first} - T_{last} \quad (4)$$

例えば、最初 150 タスクの平均処理時間 T_{first} が 10 秒で、最後 10% タスクの平均処理時間 T_{last} が 8 秒の場合、差分 T_{tranc} は 2 秒となる。このように $T_{tranc} > 0$ のとき、平均処理時間が向上していることを表す。逆に、 $T_{tranc} < 0$ のときは平均処理時間が低下、 $T_{tranc} = 0$ のときは平均処理時間が変化していないと考えることができる。

3.3.2 分析結果

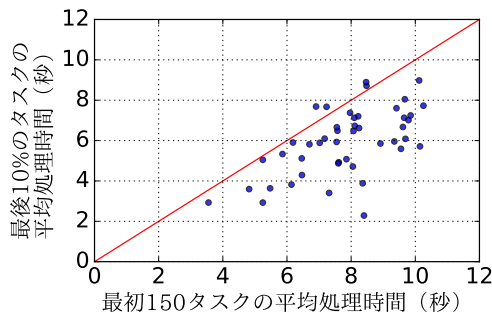
まず、長期ワーカー労働が及ぼす 1 タスクあたりの処理時間の変化についての分析を行う。長期ワーカー労働を行った 46 ワーカーそれぞれの最初 150 タスクの平均処理時間と最後 10% のタスクの平均処理時間を図 9a に示す。一つ一つの点はワーカーを表す。1 タスクあたりの処理時間が増えたワーカーは見受けられず、多くのワーカーが 1 タスクあたりの処理時間を短縮させた。

さらに、図 9b に 1 タスクあたりの処理時間の分布を示す。左側の箱ひげ図が最初 150 タスク、右側の箱ひげ図が最後 10% のタスクの 1 タスクあたりの処理時間の分布である。最初 150 タスクに比べて、最後 10% のタスクの 1 タスクあたりの処理時間の分布が下に寄っており、処理時間が少なくなっていることがわかる。また、最初 150 タスクの 1 タスクあたりの平均処理時間は 7.8 秒であったが、最後 10% のタスクの 1 タスクあたりの平均処理時間は 5.9 秒であり、これは 24.2% の短縮である。図 9c から分かるように、1 タスクあたりの処理時間が 2 秒前後短縮しているワーカーが多かった。

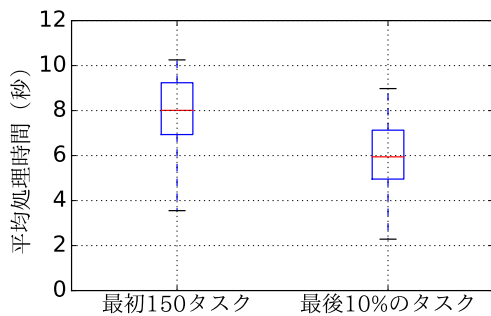
特徴的なワーカーの例を図 10 に示す。横軸がタスク処理回数、縦軸が直近 200 タスクにおける 1 タスクあたりの処理時間である。図 10a のワーカーは 1 タスクあたりの処理時間が短縮したワーカーの例である。このワーカーのように、作業開始直後から一気に 1 タスクあたりの処理時間が短縮していき、ある程度の時間に達すると 1 タスクあたりの処理時間が収束するワーカーが多く見られた。一方で、図 10b のワーカーのように、長期的な変化が見られないワーカーも数名確認できた。

次に、長期一連作業労働が及ぼす 1 タスクあたりの処理時間の変化についての分析を行う。長期一連作業労働である 255 作業それぞれの最初 150 タスクの平均処理時間と最後 10% のタスクの平均処理時間を図 11a に示す。それぞれの点は各作業を表す。一連の作業内では、1 タスクあたりの処理時間が伸びたり短くなったりとどちらかに偏ることはなく、変化が少ない作業が多かった。

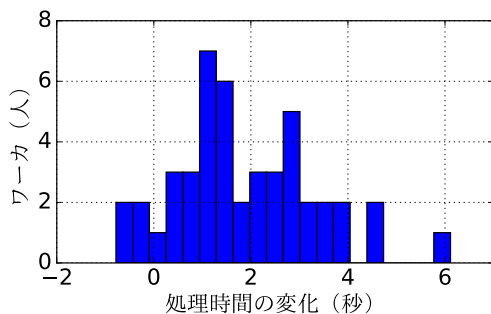
また、図 11b の左側が最初 150 タスクの平均処理時間、右側が最後 10% のタスクの平均処理時間の分布を示している。最後 10% のタスクの平均処理時間のほうが 1 秒程度中



(a) 各ワーカーの最初 150 タスクと最後 10% の 1 タスクあたりの処理時間



(b) 各ワーカーの最初 150 タスクと最後 10% の 1 タスクあたりの処理時間の分布

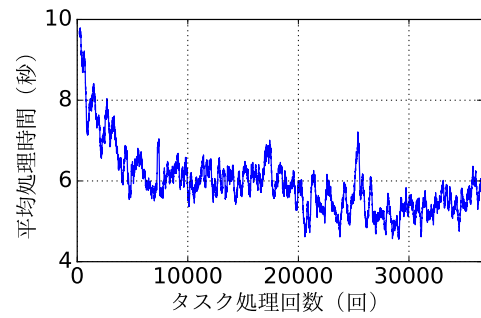


(c) 各ワーカーの最初 150 タスクと最後 10% の 1 タスクあたりの処理時間の差の分布

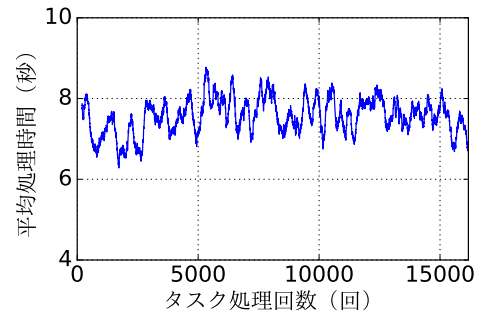
図 9: 長期ワーカー労働におけるタスク処理時間の変化

中央値が小さくなっているだけで大きな違いは見られなかった。最初 150 タスクの 1 タスクあたりの平均処理時間は 5.9 秒であったが、最後 10% のタスクの 1 タスクあたりの平均処理時間は 5.6 秒であり、これは 4.8% しか変わらない。さらに、最初 150 タスクの 1 タスクあたりの処理時間と最後 10% のタスクの 1 タスクあたりの処理時間の差の分布を示す図 11c から分かるように、ほとんどの作業で差が 0 秒付近に集まっており、一連作業内では 1 タスクあたりの処理時間が変化しないことが分かる。

特徴的な一連作業労働中の 1 タスクあたりの処理時間の変化を図 12 に示す。横軸がタスク処理回数、縦軸が直近 50 タスクにおける 1 タスクあたりの処理時間である。図



(a) タスク処理時間が向上するワーカー例



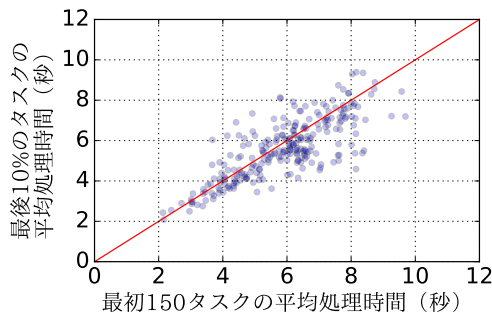
(b) タスク処理時間が変化しないワーカー例

図 10: 長期ワーカー労働におけるタスク処理時間の変化のワーカー例

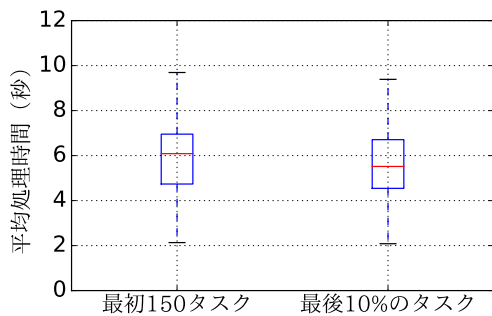
12a の作業では、400 タスクあたりから急激に処理時間が増えており、処理時間が増加している例である。図 12b は 200 タスクあたりから処理時間が減っており、処理時間が減少する例である。図 12c は長期的な傾向がなく、処理時間が安定している例であり、ほとんどの作業でこのような推移が見られた。

4. 議論

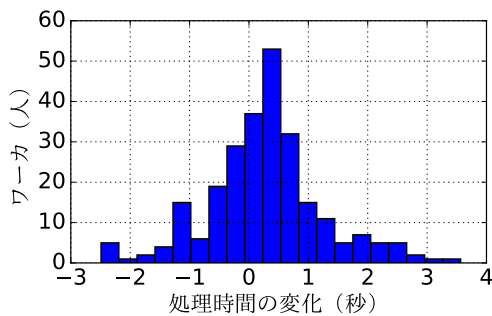
一つ目に、3.2 節の長期労働が及ぼす正答率に関する結果について議論を行う。まず、長期ワーカー労働が及ぼす正答率の変化に関する議論から進める。作業開始直後の正答率が低いと、作業後半には正答率が向上する傾向が見られた。しかし、もともと高い正答率であったワーカーはそれ以上に大きく正答率が上がることがなく、下がることもなかった。長期ワーカー労働は、正答率が低いワーカーの成長によって全体的な正答率の底上げに繋がっていると言える。また、正答率が向上する多くのワーカーは、作業開始直後から一気に正答率が向上している。正答率が悪いワーカーの作業開始直後の振る舞いを見ることは、そのワーカーがスパムワーカーかそうでないのかの判断材料になると考えられる。つまり、正答率が低いままであるワーカーは今後も正答率が向上する可能性が低く、金銭目的など悪意を持ったワーカーであるか、もしくは、そのタスクに適した能力を持ったワーカーではないと判断できる。



(a) 各作業の最初 150 タスクと最後 10%のタスクの平均処理時間



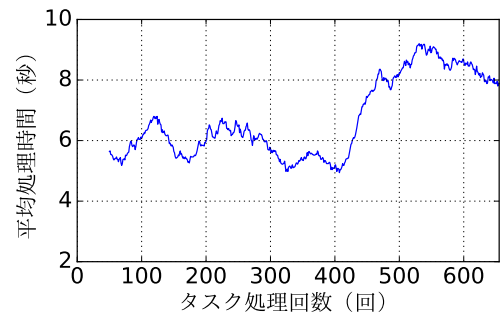
(b) 各作業の最初 150 タスクと最後 10%のタスクの平均処理時間の分布



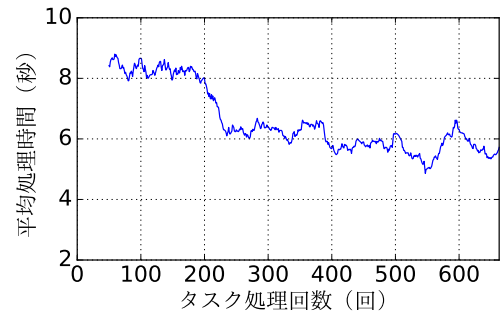
(c) 各作業の最初 150 タスクと最後 10%のタスクの平均処理時間の差の分布

図 11: 長期一連作業労働における処理時間の変化

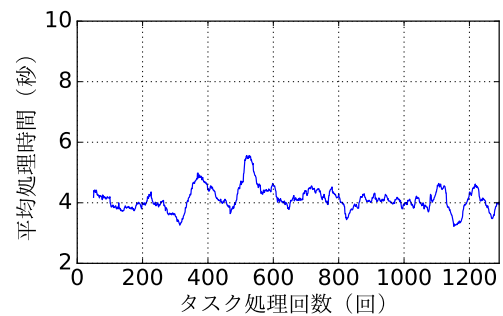
次に長期一連作業労働が及ぼす正答率の変化に関する議論を行う。長期一連作業労働では、正答率の変化にばらつきが見られるものの、全体的な変化の傾向はなく、多くのワーカーが安定した正答率で作業を続けていることが分かった。長期ワーカー労働のときと同様に、作業開始直後の正答率が低いワーカーの多くは正答率が向上した。この長期一連作業労働の多くは同一ワーカーの長期ワーカー労働の前半部分と一致していることを確認した。つまり、正答率が大きく向上するのは長期ワーカー労働の作業開始直後であり、それ以外の長期一連作業内では、正答率は向上していないと言える。また、図 7b には急激に正答率が下がったワーカーの例を示した。これは、長時間連続で作業を続けたことによ



(a) タスク処理時間が低下するワーカー例



(b) タスク処理時間が向上するワーカー例



(c) タスク処理時間が変化しないワーカー例

図 12: 長期一連作業労働におけるタスク処理時間の変化の作業例

て集中力の低下や疲労が影響していると考えられる。しかし、このようなワーカーは心理学の知見に反して少なかった。この原因としてはマイクロタスク型のクラウドソーシングの特徴が関係していると考えられる。つまり、今回のタスクのように多くのマイクロタスクでは、ワーカーは自由に作業を中断したりやめたりすることができ、強制的に作業に従事させられることはない。そのため、ワーカーは疲労を感じれば、無理に作業を続ける必要はなく休憩を挟むことができ、多くのワーカーは正答率が大きく下がる前に作業から離反したと考える。

二つ目に、3.3 節の長期労働が及ぼすタスク処理時間に関する結果について議論を行う。まず、長期ワーカー労働が及ぼす 1 タスクあたりの処理時間の変化に関する議論から進める。作業終盤の 1 タスクあたりの処理時間は作業開始直

後と比較して、平均 24.2%短縮しており、処理時間が大きく増えるワーカは存在しなかった。1 タスクあたりの処理時間が短縮するワーカは図 10a のように作業開始直後から順調に処理時間が短くなり、ある程度の処理時間に達してからは時間的に安定して作業を行った。作業開始直後はタスクに慣れていくことで次第に処理時間が短縮されるが、一旦慣れると潜在的な時間分だけ必要となり、時間的に安定した作業を行うことができるようになったと言える。今回のタスクを例に詳しく述べると、次のようなことが考えられる。作業開始直後はツイートを讀んだ後、一つ一つの選択肢の内容を讀み比べながらツイートの分類を行っていた。多くのツイートを分類していくうちに選択肢の内容を覚えたため、作業に必要な時間がツイートを讀み、選択肢を選ぶだけになった。つまり、選択肢の内容を讀む時間が短縮されたと考えることができる。

次に長期一連作業労働が及ぼす 1 タスクあたりの処理時間の変化に関する議論を行う。長期一連作業労働内では、時間的に安定した作業が行われた。処理時間が増えたり減ったりする作業は正答率の変化の議論と同様のことが言える。つまり、図 12a のように、処理時間が増えていく作業では、ワーカの集中力の低下や疲労が原因だと考えられるが、このようなワーカが少なかったのは、作業をいつでもやめることができるからだと言える。また、図 12b のように、処理時間が減る作業が少なかったのも、ワーカにとって作業に慣れていく 1 回目の作業時にしか起きないからだと言える。

最後に以上の議論をまとめる。マイクロタスク型のクラウドソーシングでは、ワーカが自由に作業をやめることができるため、長時間連続で作業を行っても、集中力の低下や疲労が現れることが少なく、正答率やタスク処理時間が悪化する例はほとんど見受けられなかった。一方で、不連続であっても長期間作業に従事し、多くの似たようなタスクを処理し慣れることが、正答率や処理時間の向上に繋がることが確認できた。

5. おわりに

本論文では、マイクロタスク型のクラウドソーシングにおいて、ワーカが長期労働を行った場合に生じるワーカの作業能力の変化に関する分析結果を論じた。全体として長時間作業を行った場合と連続的に長時間作業を行った場合に分けて、正答率とタスク処理時間の 2 種類の作業能力の分析を行った。

全体として長時間作業を行った場合の正答率は変化しないワーカが多いことが分かった。連続して長時間作業を行った場合の正答率の変化についても、同様のことが言える。これは、ワーカが作業をいつでもやめることができるために、疲労や集中力の低下が現れる前にタスクから離反したことが原因であると推測する。ワーカが作業をやめる

際にやめる理由を調査することで、この推測を検証することができる。このような結果から、マイクロタスク型のクラウドソーシングを発注する際には、時間やタスク数を強制させるのではなく、自由に取り組めるように設計することで、信頼性の高い結果を得ることができると言える。さらに、正答率が低いワーカに限ってタスクに取り組むにつれて正答率が大幅に向上するワーカが存在した。正答率が向上するワーカなのかスパムワーカなのかを判断することを今後の課題としたい。

連続した作業内ではタスク処理時間の変化は少なく、これは正答率と同様に疲労や集中力の低下が現れる前にタスクから離反したことが原因であると推測する。一方で、全体として長時間作業を行った場合には、タスク処理時間は短縮することが分かった。

今回行ったタスクの難易度が低かったために、正答率が変化しなかったとも考えられる。つまり、多少の疲労では正答率の低下に繋がらず、もともとの正答率が高いために正答率が向上する伸びしろがなかったのかもしれない。では、難易度の高いタスクとはいかなるタスクなのか、その難易度の高いタスクでは正答率の変化が起きるのか分析することが、今後のタスク設計やラベル統合において有益である。

謝辞 本研究の一部は、NAIST ビッグデータプロジェクトの助成を受けたものです。

参考文献

- [1] 芥子育雄, 鈴木 優, 吉野幸一郎, 大原一人, 向井理朗, 中村 哲: 単語意味ベクトル辞書を用いた Twitter からの評判情報抽出, 電子情報通信学会論文誌 D, Vol. 100, No. 4, pp. 530–543 (2017).
- [2] Sheshadri, A. and Lease, M.: Square: A benchmark for research on computing crowd consensus, *First AAAI Conference on Human Computation and Crowdsourcing* (2013).
- [3] Dawid, A. P. and Skene, A. M.: Maximum likelihood estimation of observer error-rates using the EM algorithm, *Applied statistics*, pp. 20–28 (1979).
- [4] Whitehill, J., Wu, T.-f., Bergsma, J., Movellan, J. R. and Ruvolo, P. L.: Whose vote should count more: Optimal integration of labels from labelers of unknown expertise, *Advances in neural information processing systems*, pp. 2035–2043 (2009).
- [5] Welinder, P., Branson, S., Belongie, S. J. and Perona, P.: The Multidimensional Wisdom of Crowds., *NIPS*, Vol. 23, pp. 2424–2432 (2010).
- [6] Krueger, G. P.: Sustained work, fatigue, sleep loss and performance: A review of the issues, *Work & Stress*, Vol. 3, No. 2, pp. 129–141 (1989).
- [7] Hata, K., Krishna, R., Fei-Fei, L. and Bernstein, M. S.: A Glimpse Far into the Future: Understanding Long-term Crowd Worker Quality, *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, ACM, pp. 889–901 (2017).
- [8] 堀江良典: VDT 作業における一連続作業時間と休憩に関する研究, 人間工学, Vol. 23, No. 6, pp. 373–383 (1987).