マウスの動きによる低品質ワーカーの検出

三上智也
融合知能デザイン研究室
指導　堀川聡

発表の要点

1	低品質ワーカーの検出方法について、正答率や、特定の問題を正解しているかなど回答による検出は本当に正しいかどうかわからない
2	散布図では特徴があるように見えるが、統計的な検定では有意差が見られなかった
3	反省点・今後の検証について

クラウドソーシングが抱えている問題

- クラウドソーシングには、不特定多数の人々にタスクを行ってもらうため、品質が保証されないという問題点がある

- Gold Standard法や、多数決法のようなタスクの結果から低品質なワーカーを検出する方法があるが、適当に行ったか本当に間違えただけかわからないという問題がある

ワーカーの振る舞いを用いた低品質ワーカーの検出

- 最近ではタスクの結果を用いたものではなくタスクを行っている振る舞いを用いて検出する方法について研究されているが、複数の要因を使用するために特別な形式のタスクを準備する必要がある

- 例えば、マウスの動きのように単純で汎用的な振る舞いのみで低品質ワーカーを検出することはできないのだろうか？

クラウドソーシングの分類

- クラウドソーシングの結果そのものから問題の検証・解決を行うのではなく、あくまでデータを集めるために用いているため分類は「集める」とする

発表の要点

1. 低品質ワーカーの検出方法について、正答率や、特定の問題を正解しているかなど回答による検出は本当に正しいかどうかわからない
2. 散布図では特徴があるように見えるが、統計的な検定では有意差が見られなかった
3. 反省点・今後の検証について

実験の目標

- 今研究で注目するワーカーの振る舞いは回答中のマウスの動きとする
- 低品質ワーカーのマウスの動きの教師データが公開されていない、または存在していない
- そのため、実験の流れとしては、既存の方法で低品質ワーカーの検出を行い、特徴があるかどうかを考察していく

手法

- ヤフークラウドソーシング上でPCからアクセスしている200人を対象とする
- タスクの内容は、画像を見て、猫の血統種を答える問題を50問連続で行うもの
- タスク実施中のマウスの座標を逐次収集し、分析を行う

タスク画面

Project: Tagging_Task

同じ血統種の猫を選択してください 10 / 50

Abyssinian

Birman

Bengal

Bombay

next question

※上記の問題が研究室内で行った動作確認のデータを参考に決めたGold Standardの問題

仮説

- マウスの動きの分布を調べる際、真剣に答える人は一様分布になり、適当に答えた人は偏った分布となる
- マウスの動きの分布の散らばりから低品質ワーカーかどうかを検出することができる

クラウドソーシングの結果

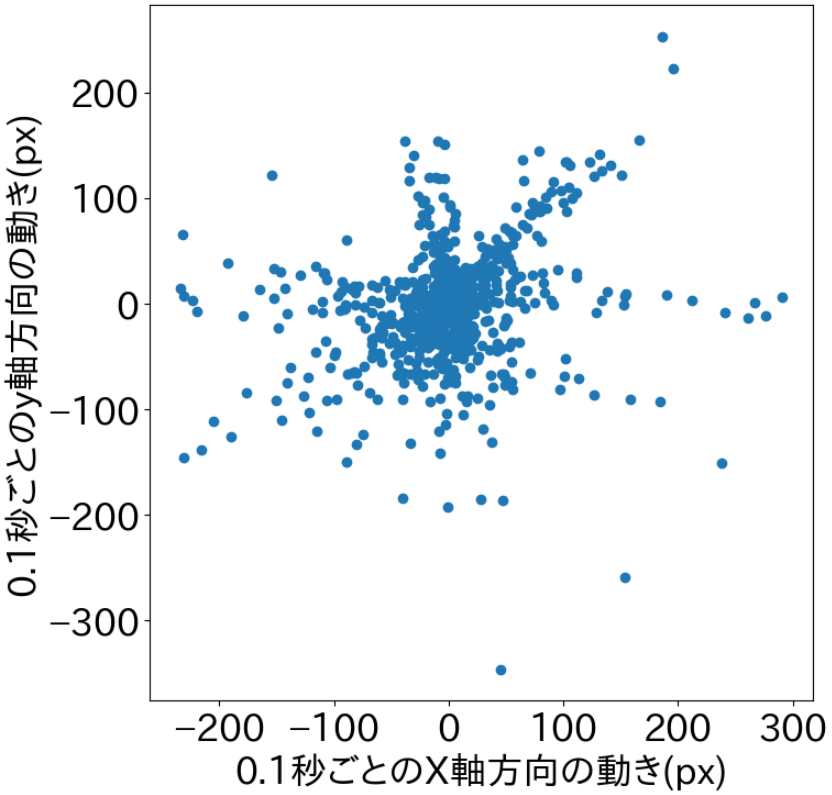
- 有効なデータは200中173
- 173中Gold Standardの設問を間違えていたデータは4つ
- 全回答者の中で正答率が30%以下であったものはID82番の回答者のみ

回答者ID	正答率
209	62
210	79.59
265	80
82	30

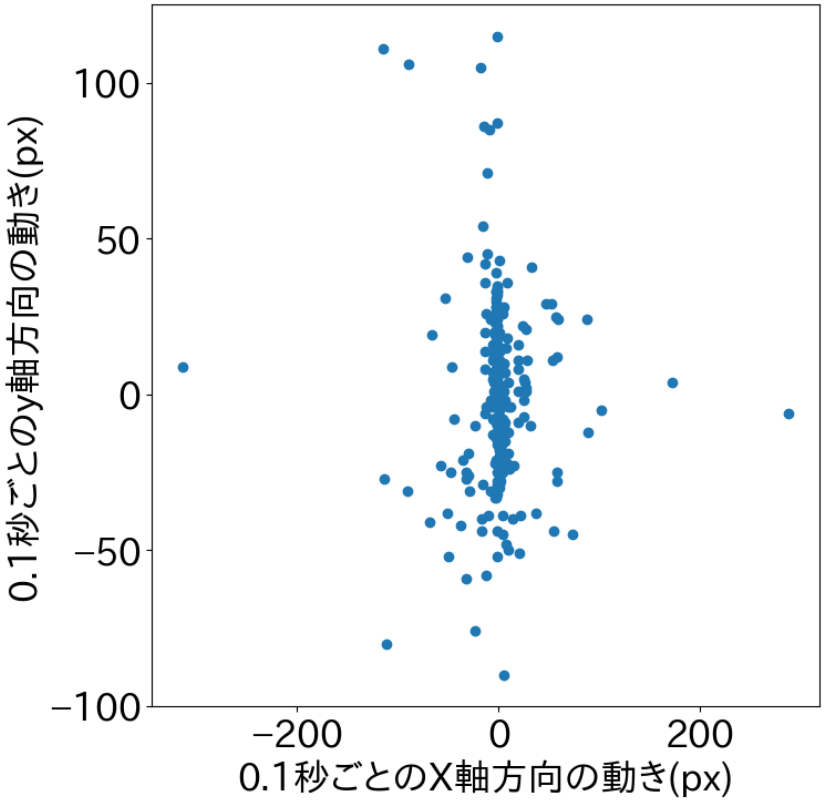
Gold Standardの設問を間違えていた回答のデータ
IDはcrowd4uで割り当てられたID

収集したデータのマウスの動きの散布図

回答者ID1の
マウスの動きの散布図



回答者ID82の
マウスの動きの散布図



- 適当に行った結果と真剣に行ったタスクでは移動量に差が出ているということが分かった

統計的な検定の結果

- ID82の回答者と他の回答者でマウスの移動量の平均値に差があるかどうかを検証するためにx方向, y方向についてそれぞれt検定を行った。有効なデータのすべての組み合わせで検定を行う

- 検定はデータ数が異なるためWelchのt検定をおこなった。

- その結果、平均値間に統計的な有意差は認められなかった(p<=0.05以下なし)

散布図と統計的な検定についての考察

- 散布図から差があるように予想はできるが数値として何が違うかを検証することができなかった。

- 先行研究においても、複数の要因からランダムフォレストなどの方法で特異なものを検出しているため、具体的にどのような振る舞いを行ったら低品質化を数値で表すのは難しい

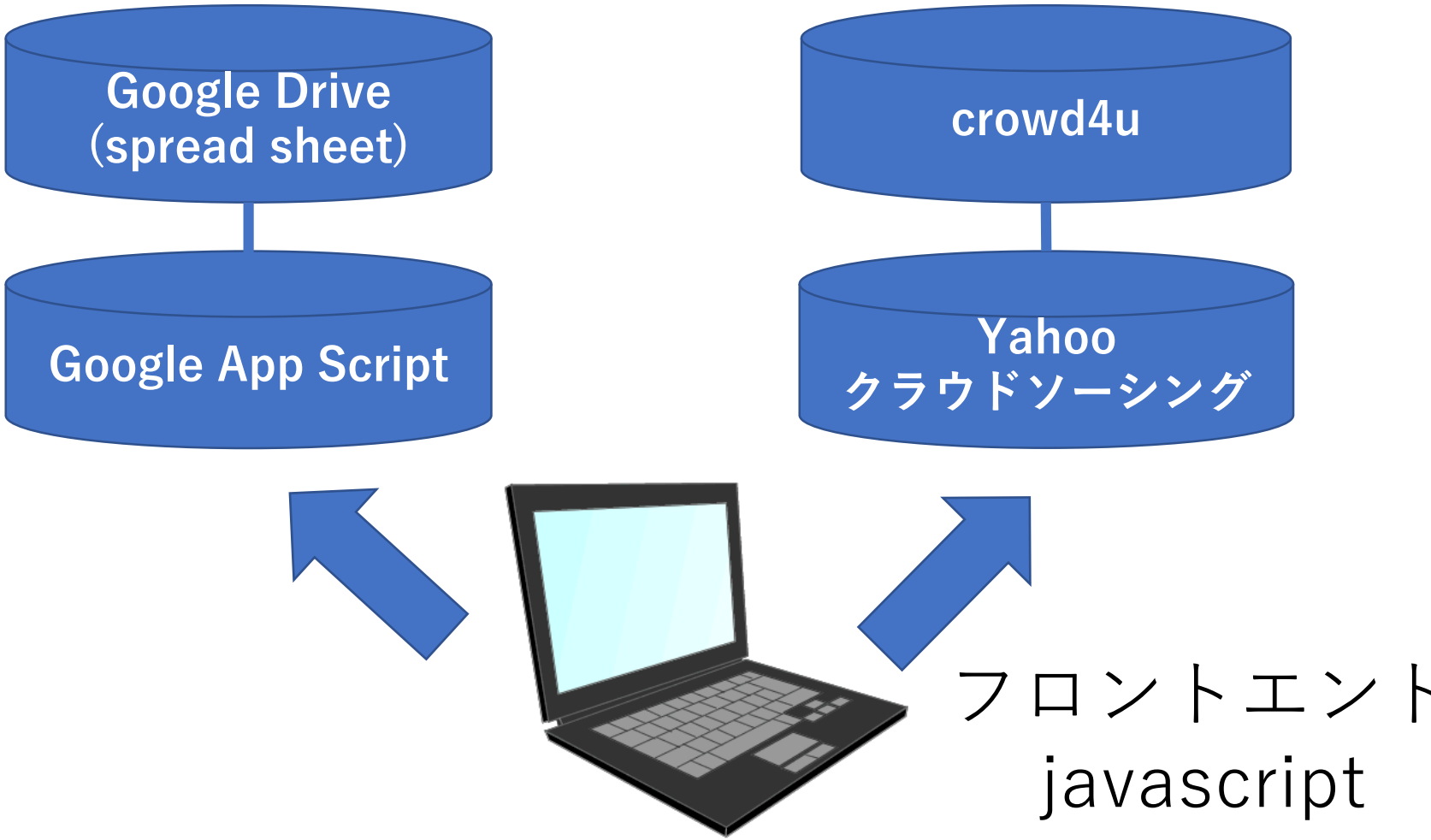
発表の要点

1	低品質ワーカーの検出方法について、正答率や、特定の問題を正解しているかなど回答による検出は本当に正しいかどうかわからない
2	散布図では特徴があるように見えるが、統計的な検定では有意差が見られなかった
3	反省点・今後の検証について

ワーカーの行動を取得するためのコストの問題

- ワーカーの振る舞いを逐次記録していくため、すべてをまとめて送ろうとすると膨大なデータが必要となる
- ブラウザの負担を減らすため、逐次データを送り、それを保存するシステムを準備する必要がある

今実験のシステムアーキテクチャ



ユーザーの振る舞いを取得するために
別途システムが必要になる

ユーザーの振る舞いについて

- マウスの動きのようにある程度連続したデータを取得するものはマイクロタスクと相性が悪い
- また、スマートフォンで行った場合とPCで行った場合ではユーザーの振る舞いが変わってくることも考えなくてはいけない

関連研究

- Aroyo, Lora and Welty, Chris. Crowd Truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. WebSci2013. ACM. 2013.

- Rzeszotarski, Jeffrey M and Kittur, Aniket. "Instrumenting the crowd: using implicit behavioral measures to predict task performance". Proceedings of the 24th annual ACM symposium on User interface software and technology. ACM, 2011, p.13-22.

- Oleson, David and Sorokin, Alexander and Laughlin, Greg P and Hester, Vaughn and Le, John and Biewald, Lukas. Programmatic Gold: Targeted and Scalable Quality Assurance in Crowdsourcing. Human computation. 2011, 11(11).

- 松田義貴, 鈴木優, 中村哲. タスク介入によるクラウドワーカの品質推定精度の改善 第10回データ工学と情報マネジメントに関するフォーラム (DEIM2018), Mar. 2018