1J5-OS-18b-4

プライベートクラウドソーシングにおける精度向上手法

Construction and Management of High-quality Private Crowdsourcing Platform

芦川　将之 *1　　川村　隆浩 *1　　大須賀　昭彦 *2
Masayuki ASHIKAWA　Takahiro KAWAMURA　Akihiko OHSUGA

株式会社東芝 研究開発センター *1
Corporate Research and Development Center, Toshiba Corporation

電気通信大学大学院情報システム学研究科 *2
Graduate School of Information Systems, The University of Electro-Communications

Open Crowdsourcing platforms like Amazon Mechanical Turk provide an attractive solution for process of high volume tasks with low costs. However problems of quality control is still of major interest. In this paper, we design a private crowdsourcing system, where we can devise methods for the quality control. For the quality control, we introduce four worker selection methods, each of which we call preprocessing filtering, real-time filtering, post processing filtering, and guess processing filtering. These methods include a novel approach, which utilizes a collaborative filtering technique in addition to a basic approach of initial training or gold standard data.

1. はじめに

クラウドソーシングは，2006 年に Wired 誌の Jeff　Howe によって提唱された．Crowd(群衆) + Sourcing(アウトソーシング) の造語であり,「企業，組織が，自社もしくはアウトソースの人材により実施していた業務を，よりオープンかつ不特定多数の Crowd(群衆) から人材を集め実施すること」と定義されている．

我々はこのクラウドソーシングの技術を様々な研究データの解析に用いている．研究データの作成は精度的な問題から自動化出来ないケースが多く，研究者，もしくは専門の技術を持った外部の業者といった人手による作業が必要になる．しかし，昨今の研究に用いられるデータはビッグデータと称される巨大なデータであることが多く，従来の人手による作業では巨大データを扱うにはコスト，速度の面から難しくなってきている．そこで，我々はクラウドソーシングを用いている．

既存のクラウドソーシングサービスとして Amazon Mechanical Turk[AMT] や Yahoo!クラウドソーシング [Yahoo!] などの様々なサービスが存在する．しかしこれらの外部サービスを研究データの作成に利用するには精度の面から問題があった．我々は作業 (タスク) の処理結果を研究データとして用いるため作業結果の品質を高く維持しなくてはならないが，そのためには外部のサービスが提供している機能の範囲では十分ではなく，さらに外部のサービスに新規の機能を追加することも難しい．我々はこれらの問題を解決するために，独自のクラウドソーシングシステムを構築し，システム内にて様々な精度向上手法を適用することで問題の解決を試みている．

本稿では我々が研究対象としているマイクロタスク型のクラウドソーシングに関して述べ (2 章)，マイクロタスク型のクラウドソーシングにおける精度向上に関する既存の研究に関して紹介し (3 章)，さらに我々が構築した PCSS における精度向上手法に関して紹介する (4 章)．

連絡先: 芦川将之，(株) 東芝研究開発センター知識メディアラボラトリー，〒 212-8582 川崎市幸区小向東芝町 1，044-549-2243，masayuki.ashikawa@toshiba.co.jp

2. マイクロタスク型クラウドソーシング

クラウドソーシングの定義は非常に緩やかなものであり，特定の目標に対して不特定多数の人間が関わって作業をしていればクラウドソーシングとして扱われている．その中でも企業や組織が用意した大量のタスクを，数多くの不特定のワーカーが処理する形式のクラウドソーシングはマイクロタスク型クラウドソーシングと言われている．我々は大規模な研究データの構築，解析のためにクラウドソーシングを用いており．そのためにはこのマイクロタスク型のクラウドソーシングが最適である．しかし外部のマイクロタスク型のクラウドソーシングサービスが提供している精度向上のための機能の範囲では十分ではないことが多く，また外部のサービスに精度向上のための新規機能を追加することも難しいという問題がある．

そのため，我々はシステム側を自由に変更することが可能なプライベートな環境下におけるクラウドソーシングシステム (PCSS) を構築し、様々な精度向上手法を適用している．

3. 関連研究

マイクロタスク型のクラウドソーシングはその特性上「安価で大量の処理が可能」という点に注目されることが多く，精度は優先度を低く設定されがちである．また，マイクロタスク型は一つ一つの作業の難易度が低いことも多く，精度を軽視させる要因の一つとなっている．しかし，我々はマイクロタスク型のクラウドソーシングを研究データの構築に用いており，精度に関しても高レベルでなくてはならない．

これまでもマイクロタスク型のクラウドソーシングの精度を向上させる方法に関して様々な研究がなされている．我々はこれらの研究を以下の 3 つのカテゴリに分類した．

1. タスクに対する精度向上手法
2. 作業者 (ワーカー) に対する精度向上手法
3. 作業出題者 (リクエスタ) に対する精度向上手法

PCSS では主に (2) のワーカーに対する精度向上手法を中心に行っている．(1) に関してはシステム外の精度向上手法に関する事項であるため，タスク内容に依存することが多くシス

テム側で対応しにくいという問題がある．実際に PCSS を運用するにあたってはリクエスタのタスクの内容に応じて対策を行っているが，PCSS における機能とは異なるため本稿では触れない．また，(3) に関してはプライベートなクラウドソーシングという特性上リクエスタが明確であるため，不正なリクエスタは存在せず対策は不要である．

(2) に関する研究として，ワーカーに信頼度の高いワーカーを紹介させる研究 [西 13]，作業結果を学習データとしてスパムワーカーを排除する研究 [Halpin 12]，ワーカーのタスクに非依存な行動からワーカーの能力を予測する研究 [Kilian 12]，ワーカーのランキングを行うことで低品質ワーカー，スパムワーカーを排除する研究 [Raykar 11] などが行われている．既存のサービスにおいても，ワーカーに事前テストを受けさせてリクエスタが必要に応じてワーカーを選別する手法 [AMT] などが行われている．

4. PCSS における精度向上手法

本章では我々が構築した PCSS における精度向上手法に関して述べる．

4.1 PCSS の構築

PCSS では，ワーカーの募集をネットワークリサーチを行なっているポイント業者 *1 へと委託した．ポイント業者は既にリサーチ対象となるユーザを数百万規模で管理しており，これらのユーザを PCSS のワーカー候補とし，そこから我々が望む条件に合致するワーカーの絞り込みをおこなった．これにより我々はポイント業者のユーザをワーカーとして作業を提供し，Web 経由で作業可能とし，さらにポイント業者を経由してワーカーに報酬を支払うという図 1 の構成を構築している．本システムは 2011 年 11 月から運用を継続しており，表 1 に示す運用実績を持っている．[芦川 12, 芦川 13]

表 1: PCSS の運用実績

運用開始	2011 年 11 月
ワーカー総数	1568 人
毎月実績のあるアクティブなワーカー	150 人
問題数	570 万件

4.2 PCSS における精度向上手法

PCSS における精度向上手法は主にワーカーに対する管理を中心に行っている．クラウドソーシングは「不特定多数の外部の人間」に作業を委託する仕組みであるため，ワーカーの品質は様々であり，優秀なスキルを持ったワーカーの存在に対して，タスク結果の品質を考慮しない低品質ワーカーや，スクリプトなどを使用して処理するスパムワーカーと呼ばれるワーカーも存在する．既存のクラウドソーシングサービスでは数多くのリクエスタから数多くのタスクを受け入れているため，ワーカーが行うタスクは多種多様となり，結果としてタスク単位におけるワーカーの行動情報が少なくなり，ワーカーのコントロールが難しくなっている．PCSS ではプライベートという特徴上タスクのカテゴリが限られているため，タスクカテゴリに対するワーカーの行動情報は相対的に多くなっており，その

*1 自社の会員に対して他社のアンケート入力作業やサービスなどを紹介し，作業結果やサービス利用の対価として一定の条件で計算されたポイントを与えるサービス．ポイントは商品や現金と交換することが可能．

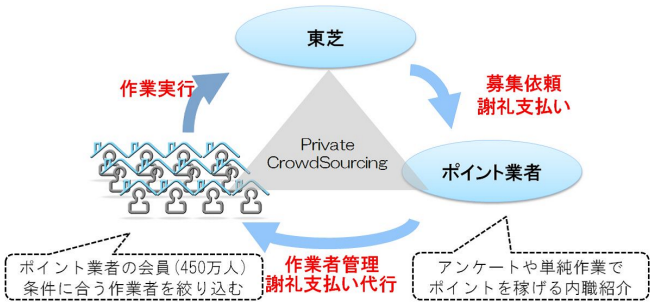

図 1: ポイント業者を経由したクラウドソーシングシステムの構築

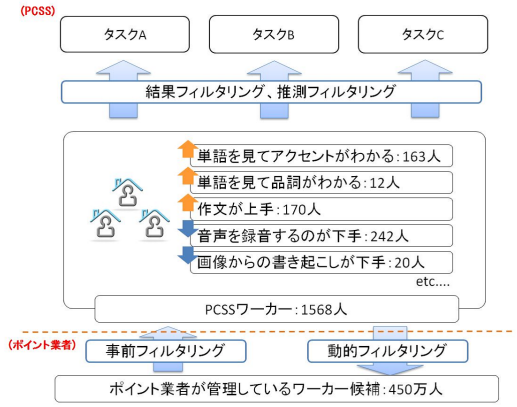

図 2: PCSS におけるワーカーに対する精度向上手法

ワーカーの行動情報を活かすことでワーカーの特性に応じた適切なタスクを与え，低品質ワーカーおよびスパムワーカーを排除することを可能としている．

以下にワーカーに対する PCSS の精度向上手法を (1) 事前フィルタリング，(2) 動的フィルタリング，(3) 結果フィルタリング，(4) 推測フィルタリング，の 4 つのカテゴリに分類した．それぞれの手法は PCSS の運用における図 2 に示したタイミングで行われる．それぞれの手法に関して詳細を述べる．

4.2.1 事前フィルタリング

ポイント業者からワーカーを募集する際に行うフィルタリングである．ポイント業者は数百万人の会員を有しており，これらのすべての会員をワーカーとして扱うのはコスト的に現実的ではなく処理能力的にも過剰である．また，これらの会員には ICT の素養が低い，Web における継続的な作業を望んでいない，などの PCSS に不適である会員も多く存在しており，このような明らかに高品質なワーカーになりえないワーカー候補を排除するために事前のアンケートを用いてフィルタリングを実施している．アンケート内容は「作業可能な時間」「熱意」「希望時給」「学歴」「基本的な IT スキル」などの基本的な設問に加えて，ワーカー募集の目的に応じた設問を追加して実施している．例として文法に関する技術を有するワーカーを募集したい場合は文法に関する設問を追記し，音声の収集を行いたい場合は保持するマイクの種類に関する設問を追記するなどの対応ができる．

4.2.2 動的フィルタリング

ワーカーがタスク処理をしている際に行うフィルタリングである．(1) 事前フィルタリングにて最低限の品質を確保できたワーカーであるが，すべての低品質なワーカーを排除できたわけではない．また，人間は時間の経過に応じて能力が上下するため，初期の品質判定が継続するとは限らない．そのため，タスク処理を進めていく課程で動的にワーカーのフィルタリングを行うために精度と経験値という 2 点の項目を設けている．

正解率は「正解数/総作業数」で算出し，一定値以下のワーカーは低品質ワーカーとみなし，以降の PCSS におけるタスク処理を禁止する．また同様に,「正解数 - 不正解数」で算出される経験値を設定し，一定の経験値を持つワーカーに対して高報酬，高難易度のタスクを提供している．これらの数値は作業中に画面に常に表示している．正解率が一定値以下になることでタスク処理ができなくなることはワーカーに明言してあり，ワーカーはこの数値表示によって精度に対する注意を喚起されるため，結果としてモチベーションを高めるゲーミフィケーション的な効果を持つ．一方，これらの数値を算出するためには正解率が必要であり，ワーカーによって入力された結果の合否判定を行わなければならない．合否判定に用いる手段としては多数決を用いる手法が提案されている [Snow 08]．我々も主に多数決にて正解を決定しており，アンケートなど正解がない場合にはタスク説明に正解が無い旨を明記し，正解率は変動させない．

ワーカーが確認することが出来るのはすべてのタスクの全体平均正解率である．しかし，動的フィルタリングをこの全体平均正解率のみで行うとフィルタリング効果が低いことがわかっている．例えば,「賃金が高く難易度も高いタスク A」と「賃金が低く難易度も低いタスク B」があった場合，ワーカーはタスク A を処理し，全体平均正解率が下がるとタスク B を行なって全体平均正解率を回復させるという行動をとることが多く，結果としてタスク A の結果品質が低下してしまう場合がある．このようなワーカーの行動に対応するため，我々はタスクのカテゴリごとに正解率をワーカーに明示せず別途管理している．特定のカテゴリの精度が一定値以下になった場合は，そのカテゴリに属するタスクを隠し，処理をさせないようにすることでワーカーの行動コントロールを行っている．

我々はこの動的フィルタリングを用いて 1630 人のワーカーから 62 人のワーカーを低品質，スパムワーカーとして排除している．

4.2.3 結果フィルタリング

ワーカーのタスク処理結果からワーカーの特徴を判別するフィルタリングである．(2) 動的フィルタリングは正解を判定することが出来る作業に対してのみ有効であり，アンケートや文章作成のような明確な正解がなく，多数決も実施しにくいタスクにおいては適用できない．しかし，明確な正解がないタスクでも，リクエスタの意図に沿った内容か否かという判定は存在しており，この判定をリクエスタにタスク毎に行わせるには大きなコストがかかる．このようなタスクに関して，リクエスタは他のリクエスタの類似したタスクの結果や，小規模のテスト用タスクを実施した結果などから，出題意図に沿った回答をしているワーカーを選別し，以降のタスクは条件に該当するワーカーのみに出題することで結果精度を向上させることができる．これらのワーカーの情報を我々は「スキル」と呼称している．例えば「品詞」のカテゴリのタスクの正解率が高いワーカーには「品詞」のスキルを付与し,「品詞」のタスクは「品詞」スキルを持つワーカーにのみ出題することで精度向上を行っている．これらのスキルはリクエスタ間で共有して使用することが出来るため，新規のリクエスタも初回から高品質なワーカーにタスクを処理させることが可能である．

我々は結果フィルタリングを用いて 163 人のアクセントスキル保持ワーカー，12 人の高難易度品詞スキル保持ワーカー，170 人の文書作成スキル保持ワーカー，242 人の音声処理が苦手な負スキル保持ワーカー，20 人の画像判定が苦手な負スキル保持ワーカーなどの絞り込みを行い，実際にタスクを振り分けることで高品質な処理結果を得ることができている．

4.2.4 推測フィルタリング

(2) 動的フィルタリングや (3) 結果フィルタリングは何らかのタスクの処理結果をワーカーの行動コントロールに流用したものであり，ワーカーが低品質ワーカーであった場合はワーカーの行動コントロールが出来る段階に達した時点で低品質な処理結果を残してしまっている事が多い．これらのデータは再処理が必要であり，大量のワーカーによって短時間で大量のタスク処理が行われるマイクロタスク型のクラウドソーシングでは時間，賃金ともに再処理のコストが大きくなってしまう．そこで，我々は更に低品質なタスク処理結果を削減するために，ワーカーの特性から行動を推測し，事前にタスクに不適切なワーカーをフィルタリングすることで精度向上を試みている．そのために我々はワーカーの類似性，及びタスクの類似性を利用した協調フィルタリングを用いて，ワーカーが未作業のカテゴリのタスクの結果精度の推測を行い，精度が低いと推測されるカテゴリのタスクは最初から処理させないという方法を用いている．協調フィルタリングとは多くのユーザの嗜好情報を蓄積し，あるユーザと嗜好の類似した他のユーザの情報を用いて自動的に推論を行う方法である．我々はユーザの嗜好情報の代わりにワーカーを特徴づける情報として，タスクのカテゴリ毎の結果精度を用いている．ワーカーをカテゴリ毎の結果精度のパターンで比較し，類似したワーカーの情報を用いて，未作業のカテゴリのタスクの結果精度の推測を行う．

実際に推測フィルタリングを行うにあたって推測精度を調査するため，今までの PCSS の運用データを用いて実験を行った．対象となったのは 2013 年 11 月時点で正解判定がある何らかのタスクを実施した経験のあるワーカー 792 人である．各ワーカーの結果精度をカテゴリ毎に集計し，その集計結果を元にピアソン相関係数を用いてワーカーの類似度を計算した．

ピアソン相関係数は協調フィルタリングにて類似度を判定する際に用いられることの多い値である．全ワーカーの集合を W，その要素を u, v，全タスクカテゴリの集合 T，その要素を i, j とする．この時あるワーカー u のタスクカテゴリ i における結果精度を $r_{u,i}$，ワーカー u の結果精度の平均を $\bar{r_u}$ とした場合，ワーカー u とワーカー v の類似度 $S_{u,v}$ は式 1 のようになる．

$$S_{u,v} = \frac{\sum_{i \in T}(r_{u,i} - \bar{r_u})(r_{v,i} - \bar{r_v})}{\sqrt{\sum_{u \in W}(r_{u,i} - \bar{r_u})^2}\sqrt{\sum_{v \in W}(r_{v,i} - \bar{r_v})^2}} \quad (1)$$

式 1 を用いて各ワーカーの類似度を計算した結果は図 3 のようになった．この結果よりワーカー間の類似度は一定ではなく，類似しているワーカーと類似していないワーカーが存在することがわかる．得られたワーカー間の類似度を元に，ワーカー u がまだ作業していないタスク i における予測タスク結果精度 $P_{u,i}$ は式 2 のように計算することができる．

$$P_{u,i} = \bar{r_u} + \frac{\sum_{v \in W}(r_{v,i} - \bar{r_v})S_{u,v}}{\sum_{v \in W} \mid S_{u,v} \mid} \quad (2)$$

ワーカーID	101	102	103	104	105
101	1	0.43	−0.4	0.13	0.59
102	0.43	1	−0.07	0.76	0.58
103	−0.4	−0.07	1	−0.38	0.79
104	0.13	0.76	−0.38	1	0.51
105	0.59	0.58	0.79	0.51	1
106	0.31	0.92	0.24	0.62	0.51
107	−0.27	0.77	−0.54	0.86	−0.1
108	0.36	0.93	−0.26	0.68	0.11
109	0.73	0.86	−0.36	0.97	0.18
110	0.69	0.93	−0.23	0.82	0.59
111	−0.61	0.1	0.61	−0.39	0.24
112	0.18	0.56	0.21	0.38	0.23
113	0.1	0.46	0.16	0.04	−0.04
114	0.79	0.82	−0.44	0.97	0.49
115	0.11	0.07	0.67	0.29	0.78

図 3: ワーカー間類似度 (一部)

式 2 で得られた予測タスク結果精度 $P_{u,i}$ の精度を確かめるために既存のデータを用いて検証を行った．既に実際の解答履歴から算出されているタスク i におけるワーカー u の実測タスク結果精度 $M_{u,i}$ と，他のワーカーとの類似度から推測した予測タスク結果精度 $P_{u,i}$ を比較検証した．検証の対象とするタスクはワーカーごとに精度差が大きく出ている「品詞判定に関するタスク」を例に用いた．得られた実測タスク結果精度 $M_{u,i}$ と予測タスク結果精度 $P_{u,i}$ の値の差の平均は 4.45 ポイントとなった．予測タスク結果精度を元に 90%以上の精度のワーカーをこのタスクにおける高品質ワーカーとしたところ，「品詞判定に関するタスク」を行った 127 名中 23 名が高品質ワーカーと推測された．実測値で調査したところ推測された 23 名全員が実際に 90%以上の高品質ワーカーであった．

また，クラウドソーシングのリクエスタは多数であるため，既存のカテゴリに属しないタスクが発生する場合も多い．それらのタスクに対してはタスク間の類似度を利用し，類似したタスクが属するカテゴリにおけるワーカー精度を元にワーカーのフィルタリングを行っている．タスク間の類似度はタスクにおけるタイトル，説明文，練習画面における文章，作業画面における文章に対して 3-gram で解析して算出した．

例として表示された二つの単語が意味的に同じかどうかを判定する，単語の意味の判定に関するタスクを挙げる。単語の意味の判定に関するタスクは数が少なく，既存のカテゴリに属していないためタスク間の類似度を用いた推測フィルタリングを用いる．そのために既存のカテゴリに属している各タスクとタスク間の類似度を計算した結果，単語の読み入力に関するタスクが類似度 0.6 で類似したタスクとして該当した．単語の読み入力に関するタスクは既存の「読み付け」カテゴリに属している.「読み付け」カテゴリにて平均精度 90%以上の高品質ワーカーを，単語の意味の判定に関するタスクを処理するワーカーとしたところ，単語の意味の判定に関するタスクにおいても精度 90%以上の結果を出すことが出来たワーカーは 14 人中 10 人，残りの 4 人のワーカーも 80%以上と高品質なワーカーであった．

5. まとめと今後の課題

本研究ではマイクロタスク型における精度向上手法を導入したプライベートなクラウドソーシングシステムを構築した．既存のクラウドソーシングサービスを利用するのではなく，ポイント業者の会員をワーカー候補としたクラウドソーシングをプライベートな環境下に構築することで独自の精度向上手法の適用が可能となっている．精度向上のための手法として，事前フィルタリングでワーカー候補を絞込み，タスク処理過程による動的フィルタリング，結果フィルタリング，推測フィルタリングを繰り返すことで高精度なワーカーを維持し，研究データに利用可能な精度を持つタスク処理結果を得ることが出来ている．

PCSS では精度向上のために様々なフィルタリングを用いているが，低品質な処理結果がなくなったわけではない．研究データの構築には常に高品質なデータが求められるため，引き続き精度向上のための手法を考案，適用していかなくてはならない．また，今後クラウドソーシングを用いた就労形態が一般的になった際に，簡易にワーカーを排除することは効率的な面からも社会的な面からも問題がある．そのため，低品質ワーカーに対しては排除だけではなく低品質ワーカーを高品質ワーカーにするための手法を検討するなど新たな精度向上施策を検討していくことが今後の課題である．

参考文献

[AMT] Amazon Mechanical Turk, https://www.mturk.com/mturk/

[芦川 12] 芦川 将之, 西山 修, 下郡 信宏, “CrowdSourcing を用いた単語への読み付け, アクセント付け手法の提案”, 電子情報通信学会技術研究報告 ,111(447), pp. 11-16, (2012).

[芦川 13] 芦川 将之，宮村祐一，有賀康顕,“PrivateCrowdSourcing を用いた言語, 音声資源の収集 〜システムの構築と言語収集〜”，人工知能学会全国大会,(第 27 回), (2013).

[Halpin 12] Halpin, H., Blanco, R.,“Machine-Learning for Spammer Detection in Crowd-Sourcing”, HCOMP, (2012)

[Kilian 12] Kilian, N., Krause, M., Runge, N., Smeddinck, J., “Predicting Crowd-Based Translation Quality with Language-Independent Feature Vectors”, HCOMP, (2012)

[Raykar 11] Raykar, V., Yu, S., “Ranking annotators for crowdsourced labeling tasks”, NIPS, (2011).

[西 13] 西 智樹, 小出 智士, 大野 宏司, 長屋 隆之, “ソーシャルネットワークを用いたクラウドソーシングの品質向上”, 人工知能学会全国大会,(第 27 回)(2013).

[Snow 08] Snow, R., O’Connor, B., Jurafsky,D., Ng, A.Y., “Cheap and Fast But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks”, EMNLP, (2008).

[Yahoo!] Yahoo!クラウドソーシング, http://crowdsourcing.yahoo.co.jp/