3K3-OS-20a-2

クラウドソーシングワーカーの段階的育成方法の提案

Proposal of Worker Resources Development Method in 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

Current crowdsourcing platforms such as Amazon Mechanical Turk provide an attractive solution for processing of high-volume tasks at low cost. However, problems of quality control remain a major concern. We developed a private crowdsourcing system (PCSS) running in a intranetwork, that allow us to devise for quality control methods. In the present work, we designed a novel task allocation method to improve accuracy of task results in PCSS. PCSS analyzed relations between tasks from workers' behavior using Bayesian network. PCSS created learning tasks according to analyzed relations. PCSS increased quality of task results by allocating learning tasks to workers before processing difficult tasks.

1. はじめに

Wired 誌の Jeff Howe によって 2006 年に提唱されたクラウドソーシング技術は，大規模データの解析や構築など様々な分野や用途で利用されている．その利用範囲の拡大に従い実際に作業 (タスク) を処理する作業者 (ワーカー) の数も増大しており，将来的にクラウドソーシングにおける作業が社会における一つの就労形態となることが予想される．しかしそのような傾向にあるにもかかわらず，現状のクラウドソーシングではワーカーに対する育成や労働環境の改善と言ったサポートが十分であるとは言い難い．これはワーカーが不特定多数であり，補充や変更が容易であることが原因であると予想されるが，このようなワーカーの安易な変更は，ワーカーの経験不足による全体の精度低下やワーカーの不当解雇という問題につながりかねない．

そのため今後のクラウドソーシング運用では通常の労働環境と同様に人材 (ワーカー) の育成が重要になってくると予想される．しかしクラウドソーシングにおける人材育成には様々な問題がある．特にマイクロタスク型クラウドソーシングではワーカーの数の多さ，ワーカーの匿名性からワーカー個人への対応が難しい．また「高速」「低コスト」が利点であるため，コストや時間をかけて人材を育成するのはその利点を失わせてしまう可能性があるなどの問題がある．

我々はこのようなマイクロタスク型クラウドソーシングにおける人材育成の問題に対し，ワーカーがタスクを処理する過程で適切な学習タスクをこなすことで技術を向上させる段階的な学習方式を提案する．具体的な手法としては全てのワーカーと全てのタスク処理結果からベイジアンネットワークを用いてタスク間の関係性を解析し，タスクを処理することで段階的な学習が可能となるような学習用タスクを自動生成することでワーカーの能力の育成を狙う．

本稿では我々が構築したプライベート環境下における独自のクラウドソーシングプラットフォーム (PCSS) に関して紹介し (2 章), マイクロタスク型のワーカーベースの精度向上に関する既存の研究に関して紹介し (3 章)，さらに我々が提案するワーカーの能力向上のためのタスクの出題コントロールに関して述べる (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. PCSS

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難易度の低いタスクを，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．一般にマイクロタスク型のクラウドソーシングにおけるタスクの難易度は低く，一つのタスクにかかる時間は数秒から数分と非常に短いが，支払われる単価も低く設定されており大量にタスクを処理することが前提となっている．Amazon Mechanical Turk[AMT] などがこの形式のクラウドソーシングを行っている．

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いる．そのために用いるクラウドソーシング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では外部の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った．

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 [芦川 14]．PCSS は表 1 に示す運用実績を持っている．

PCSS における精度向上手法はワーカーのコントロールが中心である．我々はこれをフィルタリングと呼んでおり，ワーカーを募集する際に条件でフィルタリングする事前フィルタリング，タスク処理中に一定値以下の精度となったワーカーを排除する動的フィルタリング，タスク処理結果からワーカーの特徴を解析して得意分野のタスクを処理させる結果フィルタリン

表 1: PCSS の運用実績

運用開始	2011 年 11 月
ワーカー総数	2454 人
毎月実績のあるアクティブなワーカー	150 人
問題数	1135 万件

グ，ワーカーの行動履歴からワーカー間の類似度を計算し，類似したワーカーの特徴から得意分野を推測する推測フィルタリングの４つのフィルタリングから構成されている．これらのフィルタリングでは，全て，もしくは特定のタスクにおいて低品質と判定されたワーカーは全て，もしくは特定のタスクの作業ができなくなる．しかしこのようなワーカーの安易な排除はワーカーの不足という別の問題の原因となった．ワーカー不足に対処するためにはワーカーを新規に追加する方法が最も容易だが，追加するワーカーは未経験のワーカーであるため結果としてタスク処理結果の精度低下につながってしまう．また安易なワーカーの排除によりクレームが発生するなどの問題もあった．そのため本研究では低品質と判断されたワーカーを単純に排除するだけではなく，低品質なワーカーを高品質なワーカーへと変化させるための手法に関して検証している．

3. 関連研究

マイクロタスク型のクラウドソーシングの精度向上手法としてワーカーに対する精度向上手法に関しては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. 低品質なワーカーを排除する精度向上手法
2. ワーカーへ出題するタスクをコントロールする精度向上手法
3. ワーカー間のコミュニケーションを助長する精度向上手法

従来の PCSS では主に (1) の低品質なワーカーを排除する精度向上手法を中心に行ってきた．しかし安易なワーカー排除は前述のように様々な問題の原因となった．そのため本研究では (2) のワーカーへ出題するタスクをコントロールする精度向上手法を用いて低品質ワーカーを高品質ワーカーに変化させることを試みている．(3) に関してはワーカー間のコミュニケーショントラブルによる炎上やクレームなどでメンテナンスコストが向上することが予想されるため PCSS では採用していない．

(2) に関する研究として，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]，ワーカーとタスク出題者 (リクエスタ) の関係を解析しタスクの出題をコントロールする研究 [Martin 14]，難易度の高いタスクから単純化したタスクを作成して先に処理させる研究 [Andre 14]，タスク難易度に応じてタスクの出題方法をコントロールする研究 [Bragg 14]，タスクの内容やワーカーのタスクに対する完遂率をベースにタスクの推薦を行なう研究 [Ambati et al., 2011]，ワーカーの行動履歴，ワーカーのタスクに対する嗜好からワーカーにタスクの推薦を行なう研究 [Yuen et al., 2012]，タスクの難易度レベル，ワーカーのスキルのレベルを推測した結果からワーカーにタスクの推薦を行う研究 [Ho et al., 2013] などが行われている．

4. PCSS における段階的学習法

ある業務において作業者が作業内容を学習する方法として，最初から難易度の高い作業を行うのではなく，目的の作業に関連する難易度の低い作業から開始して訓練し，段階的に難易度を上げていくことで作業者の能力を向上させていくという手法があげられる．この手法は学校教育の仕組みと同じであり，有効であることは示されている．この段階的な学習方法をクラウドソーシング環境で実施するためにはタスクを難易度別に段階的に用意しなくてはならない．しかしタスク内容はリクエスタによって千差万別であり，システム側で汎用的な学習用タスクを作成することは困難である．一方リクエスタがタスクごとに段階的に学習用タスクを作成するのはコストの面で現実的ではない．そのため多くの場合は目的のタスクを説明するための単純な練習画面を手動で作成するにとどまっている．我々はこの問題を解決するためにタスクを流用して学習用タスクを自動生成する手法を提案する．タスク A を実施してからタスク B を実施した場合と，タスク A を実施せずにタスク B を実施した場合で，多くのワーカーが前者のケースでタスク B の処理結果が向上していた場合，タスク A はタスク B の学習用タスクとして扱うことができるという手法である．この方法は大量のワーカーの行動履歴とタスクが必要となるため難易度が高かったが，PCSS では表 1 に示すような運用実績を持っており，これらの大規模データを利用することで実現が可能となった．

4.1 タスクの自動カテゴライズ

学習用タスクを他のタスクから自動生成するためには，タスク間の関係性をタスクの内容に応じて解析する必要がある．しかし PCSS は表 1 に示しているように大量のタスクを扱っており，個々のタスクを解析するには多大なコストが発生する．タスクの解析を効率化するためにタスクを内容に応じてカテゴリ分類している．

PCSS ではタスクを出題する際にはリクエスタがタスクのタイトル，説明文を記述している．タスクをカテゴリ化するためにこのタイトルと説明文を形態素解析し，得られた単語を元に各タスクの TFIDF 値を計算した．単語 i の出現回数を W_i，全タスクにおけるすべての単語の種類数を W_{all}，全てのタスク数を T_{all}， 単語 i の出現するタスク数を T_i とした場合，タスク t における単語 i の TFIDF 値 $TFIDF_{t,i}$ は式 1 のように計算することができる．

$$TFIDF_{t,i} = \frac{W_i}{W_{all}} log \frac{T_{all}}{T_i} \tag{1}$$

得られた各タスクにおける各語彙の TFIDF 値を用いて，各タスク間における類似度の計算を行った．類似度の計算にはコサイン類似度を用いており，タスク t における単語 i の TFIDF 値を $TFIDF_{t,i}$，全単語の集合を W とした場合，タスク $t1$ とタスク $t2$ 間のコサイン類似度 $cos(t1, t2)$ は式 2 のように計算することができる．

$$cos(t1, t2) = \sum_{i \in W} TFIDF_{t1,i} \cdot TFIDF_{t1,i} \tag{2}$$

PCSS で扱っているタスクの内，882 タスクに対して相互のタスク間類似度を計算した．882 タスク間の相互組み合わせ 777,924 通りにおけるタスク間類似度の計算をおこなった．算出されたコサイン類似度の精度を確認するために，類似度の一定の範囲ごとに 100 件のタスクの組み合わせを抽出し目視で確認したところ，コサイン類似度が 0 以上 0.1 未満の場合は 100 件中 88 件，0.1 以上 0.2 未満の場合は 100 件中 43 件，

0.2 以上 0.3 未満の場合は 17 件，0.3 以上 0.4 未満の場合は 4 件が別カテゴリに所属すると思われるタスクが存在し，0.4 以上の場合は別カテゴリに所属すると思われるタスクは存在しなかった．そのためコサイン類似度 0.4 を閾値としてタスクのカテゴリ分類を行っている．

さらにこの類似度を用いてタスクのカテゴリ分類を行った．カテゴリ分類のアルゴリズムは 1) 各カテゴリ所属のタスク全てと比較し，最も類似しているタスクが所属するカテゴリに分類，2) 閾値を類似度 0.4 とし，どのカテゴリのどのタスクとも類似度が 0.4 以下なら新カテゴリを割り当てる，の繰り返しである．その結果 882 タスクを 50 カテゴリに分類することができた．

また，どれにも類似していないタスク，すなわち最大類似度が 0.4 以下のタスクは 882 タスク中 9 タスク存在したが，それぞれがテスト系などの独立したタスクであった．そのためこの 9 タスクは除外している．

4.2 タスクカテゴリ間の関係性の解析

PCSS では表 4.1 で得られたタスクカテゴリ間の関係性を解析するためにベイジアンネットワークを用いている．ベイジアンネットワークは因果的な特徴を有向グラフによるネットワークとして表す，現象の因果性，連関性を計算的に推論する理論・技術である．ベイジアンネットワークの操作は大別して 1) ベイジアンネットワークの学習，2) ベイジアンネットワークを用いた推論の二つで行われる．

ベイジアンネットワークをクラウドソーシングに適用した具体的な例をあげる．クラウドソーシングにおけるタスクとワーカーの行動履歴からベイジアンネットワークの学習を行い図 1 のような有向グラフが得られたと仮定する．タスク A はタスク B に影響し，タスク B，C はタスク D に影響することがわかる．この影響は以下の 2 つのパターンが考えられる．1) タスク B，C を処理した後にタスク D を処理したタスク処理結果精度と，タスク B，C を処理せずにタスク D を処理したタスク処理結果精度を比較した場合，前者の方がタスク処理結果の精度が高かった場合，タスク B，C をタスク D の学習用タスクとして取り扱うことで結果精度の向上を狙うことができる．2) タスク B，C を処理した後にタスク D を処理したタスク処理結果精度と，タスク B，C を処理せずにタスク D を処理したタスク処理結果精度を比較した場合，後者の方がタスク処理結果の精度が高かった場合，タスク D を処理させるワーカーにはタスク B とタスク C を処理させないことで結果精度の向上を狙うことができる．

PCSS にベイジアンネットワークを適用するにあたって，PCSS の今までの運用実績からベイジアンネットワークの学習を行うため，各タスクカテゴリにおけるワーカーの結果精度平均値を算出した．

算出したワーカーの結果精度平均値からベイジアンネットワークの学習を行い有向グラフの作成を行う．ベイジアンネットワークの学習と有向グラフの作成には Weka を用いた．精度向上させたいタスクカテゴリとして今までの処理結果から平均精度が低い順に表 2 に示した．結果として「3. Ｗｅｂページのジャンルが似ているか判定」「11. 文節の切れ目を判定」のタスクカテゴリに関して図 2，図 3 に示すような有向グラフを得ることができた．

得られたタスクの相関関係の有効性を確認するために各タスクに関して以下の様な実験を行った．

1. 初期の状態確認のために精度向上対象タスクカテゴリのタスクを処理させる．

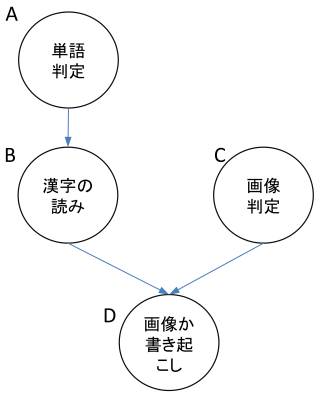

図 1: ベイジアンネットワークをクラウドソーシングに用いた例

表 2: 制度改善タスクカテゴリ一覧

ID	タスクカテゴリ名	平均精度 (%)
1	キーワード分類	85.8
2	品詞判定	86.6
3	Ｗｅｂページのジャンルが似ているか判定	90.1
4	アクセントを評価する	90.4
5	単語の読み方確認	92.3
6	アクセントを含めた読み方確認	93
7	アクセント選択	93.2
8	言葉の読み方を判定	94.1
9	言葉の読み方の入力	94.4
10	自然文判定	94.7
11	文節の切れ目を判定	95.8

2. 学習タスクを処理するワーカーグループと (1) と同じ内容のタスクを処理するワーカーグループに分ける．
   (a) (1) の作業を行ったワーカーの内半分のワーカーに学習タスクの処理をさせる．
   (b) もう半分のワーカーに (1) と同じ内容のタスクを処理させる
3. (2) を実施後，(2) を処理した全員のワーカーにもう一度 (1) と同じ内容のタスクを処理させて，(1) と (3) の精度変化を確認する．

この実験によって得られた効果を表 3 に示す．

以上のようにベイジアンネットワークの学習から導出された学習用タスクを実施することによって，精度向上対象タスクの処理結果精度が大きく向上していることがわかる．学習用タスクを実施しない場合でもわずかながらの改善が見られるが，これは同一のタスク処理を継続することで作業に慣れた結果であると推測している．

上記 2 つ以外のタスクカテゴリに関して，タスクカテゴリ「5. 単語の読み方確認」は有向グラフを得ることができたが，実験時に初期の状態確認用作業で全員の平均正解率が 97.5%となってしまい改善効果を測定することが難しくなったため実験対象から外している．またタスクカテゴリ「8. 言葉の読み方を確認する」も有向グラフを得ることができたが，対象となる

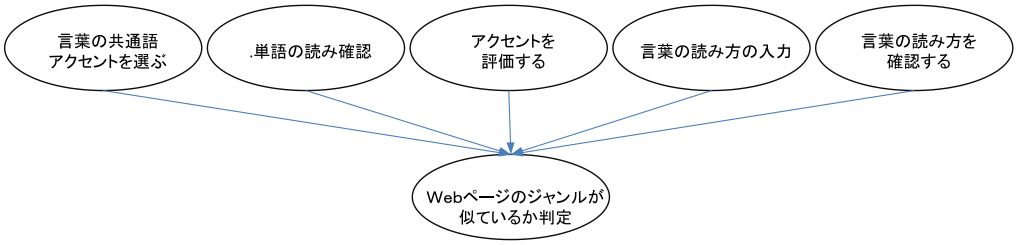

図 2: 3. Ｗｅｂページのジャンルが似ているか判定

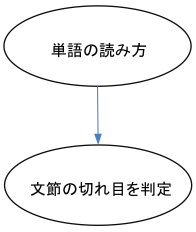

図 3: 11. 文節の切れ目を判定

表 3: 学習用タスク実施のの有無によるタスク改善効果

対象タスクカテゴリ	学習内容	ワーカー数	改善した人数	改善効果 (ポイント)
Web ページのジャンルが似ているか判定	学習用タスク	53	47	15.9
	同じタスク	4	1	1.9
文節の切れ目を判定	学習用タスク	10	10	9.65
	同じタスク	10	5	2.86

タスクカテゴリに影響しているタスクカテゴリが存在しない有向グラフであったため実験対象から外している．またタスクカテゴリ「2. 品詞判定」も有向グラフを得ることができたが，対象となるタスクカテゴリに影響しているタスクカテゴリが悪影響を与えるタスクカテゴリのみであったため実験対象から外している．またタスクカテゴリ「4. アクセントを評価する」「6. アクセントを含めた読み方確認」「7. アクセント選択」は有向グラフを得ることができなかったが，アクセント系のタスクカテゴリは作業可能なワーカーを限定していることが原因であると予測している．またタスクカテゴリ「9. 言葉の読み方を入力」「10. 自然文判定」も有向グラフを得ることができなかったが，これらのタスクカテゴリは正解率が高いワーカーが多く行動履歴の特徴が解析できなかったためと予測している．

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において，ワーカーの能力を向上させるための段階的な学習法を提案した．今までの運用履歴からタスクを自動でタスクカテゴリ化し，得られたタスクカテゴリにおけるワーカーの行動履歴をベイジアンネットワークの学習に用い，得られた有向グラフからタスクカテゴリ間の関係性を学習用タスクの作成に用いることで，有効な学習用タスクの作成を自動に行うことができている．

このように学習用タスクの自動生成を行うことでワーカーの能力向上につなげることができたが，タスクカテゴリの中にはワーカーの行動履歴の不足などから明確な関連性を得られないタスクカテゴリも存在する．そのタスクカテゴリにおけるワーカーの行動履歴を増やし，学習量を増やすことで将来的に関連性が得られると予測しているが，それまでは学習用タスクを手動で作らなければならないという問題がある．少量のタスクしか持たないタスクカテゴリに関しても効果的な精度向上手法を検討することが今後の課題である．

本論文に掲載のサービス等の名称は，それぞれ各社が商標として使用している場合があります．

参考文献

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

[Andre 14] Andre, P., Aniket, K., Dow, S., “Crowd Synthesis:Extracting Categories and Clusters from Complex Data”, CSCW, (2014)

[Ambati et al., 2011] Ambati, V. et al., “Towards task recommendation in micro-task marketss”, In proc. of HCOMP, (2011).

[芦川 14] 芦川 将之, 川村 隆浩, 大須賀 昭彦, “マイクロタスク型クラウドソーシングプラットフォーム環境における精度向上手法の導入と評価”, 人工知能学会論文誌 ,29(6), 503-515, (2014)

[Bragg 14] Bragg, J., Kolobov, A., Mausam, Weld, D., “Parallel Task Routing for Crowdsourcing”, HCOMP, (2014)

[Ho et al., 2013] Ho, C. J., et al., “Adaptive Task Assignment for Crowdsourced Classification, In proc. of ICML, (2013).

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

[Martin 14] Martin, D., Hanrahan, B., O’Neill, J., Gupta, N., “Being A Turker”, CSCW, (2014)

[Yuen et al., 2012] Yuen, M. C., et al., “TaskRec: probabilistic matrix factorization in task recommendation in crowdsourcing systems”, In proc.of ICONIP, (2012).