2M5-OS-07b-1

非定型出力をもつクラウドソーシングタスクにおける成果物の統計的品質推定

Statistical Quality Estimation for Crowdsourcing Tasks with Unstructured Response Formats

馬場 雪乃[*1]　鹿島 久嗣[*1]
Yukino Baba　Hisashi Kashima

[*1]東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo

Controlling the quality of results from crowd workers is a major challenge for requesters and platform providers of crowdsourcing. We propose an unsupervised statistical quality estimation method for crowdsourcing tasks with unstructured response formats such as article writing and logo designing, which occupy the majority on most crowdsourcing marketplaces. Our method is based on the two-stage procedure; multiple workers are first requested to work on the same tasks in the creation stage, and then another set of workers review and grade each artifact in the review stage. We model the ability of each author and the bias of each reviewer, and propose a two-stage probabilistic generative model using the graded response model in the item response theory. Experiments using several general crowdsourcing tasks show that our method outperforms popular vote aggregation methods, which implies that our method can deliver high quality results with lower costs.

1. はじめに

クラウドソーシングは，インターネットを通じて不特定多数の人々に仕事を依頼する仕組みである．Amazon Mechanical Turk などのプラットフォームの普及に伴い，画像へのタグづけや Web コンテンツの分類といった単純なものから，記事作成や翻訳，デザインといった複雑なものまでさまざまな種類の仕事がクラウドソーシングで依頼されている．

クラウドソーシングにおける重要課題のひとつが，作業者（ワーカー）が作成した成果物の品質管理である．品質管理のアプローチとして広く用いられている手法に「冗長化」，すなわち同じ作業（タスク）を複数のワーカーに依頼し成果物を統計的に統合して質の高い成果物を獲得するやり方がある [Dawid 79, Whitehill 09, Welinder 10]．しかし，多くの既存手法はワーカーの成果物が，２値（例.「はい」「いいえ」）や多値（例．５段階採点）といった定型出力である場合，あるいは重み付き平均を取ることで容易に統合できる実数出力の場合だけを対象にしており，記事作成やロゴデザインのような「非定型出力」のタスクには対応できない．クラウドソーシングにおいては，たとえば Amazon Mechanical Turk においては記事作成・改稿や Web サイトへのフィードバックといった仕事が支払額の多くを占めることが知られており [Ipeirotis 10]，また 99designs や DesignCrowds といったグラフィックデザイン専門のプラットフォームも普及し，非定型出力のタスクの需要は大きい．

非定型出力成果物の品質を推定する自然な方法は，成果物を作成する段階（「作成段階」）の後に成果物を評価する段階（「評価段階」）を導入するやり方である．作成段階では，複数のワーカー（「作成者」）がタスクに割り当てられ，それぞれ成果物を作成する．各成果物は評価段階へと送られ，複数のクラウドソーシングワーカー（「評価者」）によって採点される．採点は，多肢選択式（たとえば「悪い」「普通」「良い」）を取る事が多い．非定型出力成果物の品質を直接推定することは困難であるが，評価段階を導入することにより間接的に品質を推定し，品質の高い成果物を選ぶことが可能となる．たとえば，Zaidan らはこの２段階プロセスを導入しており [Zaidan 11]，また Dai らは多段階のプロセスを用いている [Dai 11]．しかし彼らはモデルのパラメータ推定にドメイン知識や教師情報を用いており，汎用的な手法とは言えない．

そこで我々は，２段階プロセスの上で教師情報を用いずに統計的に，非定型出力タスクにおける成果物の品質推定を行う手法を提案する．我々は，図１に示す２段階の生成モデル（「２段階モデル」）を提案する．作成段階を，作成者が自身の能力とタスクに依存する性能にもとづいてある品質の成果物を生成する過程としてモデル化する．ここでの品質が推定対象である「真の品質」であり，観測できないものとする．評価段階は，評価者が自身のバイアスと成果物に対する嗜好にもとづいて，ある真の品質をもつ成果物に対する潜在的な品質スコアを定め，そこから採点ラベルを生成する過程としてモデル化する．採点ラベルの生成モデルとしては，項目反応理論における段階反応モデル [Samejima 69] を導入する．成果物の真の品質とその他のモデルパラメータを，MAP 推定を用いて推定する．

商用クラウドソーシングサービス上でロゴデザイン，画像説明，翻訳タスクにおける成果物と採点結果のデータを収集し実験を行った．提案手法が多数決と段階ラベル統合手法 [Raykar 11] に比べて，特に一つの成果物辺りの評価者数が少ないときに，より精度良く品質を推定できることを確認した．

2. 非定型出力クラウドソーシングタスクに対する成果物の品質推定問題

まず，作成段階と評価段階から成る２段階プロセスの上で，非定型出力クラウドソーシングタスクに対する成果物の品質を推定する問題を定式化する．非定型出力をもつタスクの集合 $\mathcal{T}$ があるとする．各タスク $t \in \mathcal{T}$ について，作成者集合 $\mathcal{A}_t$ が割り当てられている．作成段階では，各作成者 $a \in \mathcal{A}_t$ がタスク t に対する成果物を作成する．成果物の（観測されない）真の品質を $q_{t,a} \in \mathbb{R}$ とする．評価段階では，タスク t に対して作成者 a が作成した成果物に対して，評価者集合 $R_{t,a}$ が割り当てられている．各評価者 $r \in R_{t,a}$ による採点ラベル $g_{t,a}^{(r)}$ は，

連絡先: 馬場 雪乃，東京大学大学院情報理工学系研究科，yukino_baba@mist.i.u-tokyo.ac.jp

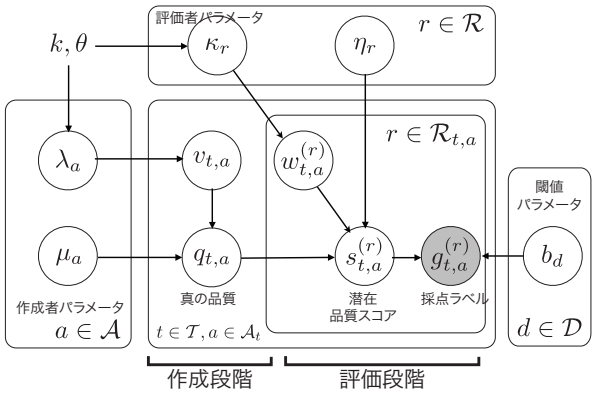

図 1: 提案する 2 段階モデルのグラフィカルモデル

段階的な採点ラベル集合 $\mathcal{D} = \{1, 2, \cdots, n\}$ から選択される．

我々の目的は，各成果物の真の品質 $\{q_{t,a}\}_{t\in\mathcal{T}, a\in\mathcal{A}_t}$ を，観測された採点ラベル集合 $\{g_{t,a}^{(r)}\}_{t\in\mathcal{T}, a\in\mathcal{A}_t, r\in\mathcal{R}_{t,a}}$ から推定することである．

3. 非定型出力クラウドソーシングタスクに対する 2 段階モデル

成果物の真の品質 $q_{t,a}$ を推定するために，我々は 2 段階の生成モデルを導入する．このモデルでは，まず作成段階において品質 $q_{t,a}$ の成果物が生成され，評価段階において評価者 r による採点ラベル $g_{t,a}^{(r)}$ が生成される．図 1 に，提案する採点ラベル生成過程のグラフィカルモデルを示す．

3.1 作成段階

我々は，高い能力をもつワーカーほど高品質の成果物を作成するという前提のもの，各作成者 $a \in \mathcal{A}_t$ が能力 $\mu_a \in \mathbb{R}$ を有するとみなした．また，作成者の性能はタスクによっても異なると考えられる．たとえば翻訳タスクにおいて，基本的な能力が低いワーカーであっても情報技術に詳しければ，情報技術に関する文章については質の高い訳文を作成できるということが考えられる．このようなタスクに依存する性能をノイズとして $v_{t,a} \in \mathbb{R}$ で表した．このノイズは，平均 0，分散 $1/\lambda_a$（すなわち，精度 λ_a）の正規分布に従うとみなす：

$$v_{t,a} \sim \mathcal{N}\left(v_{t,a} \mid 0, 1/\lambda_a\right) = \sqrt{\frac{\lambda_a}{2\pi}} \exp\left(-\frac{\lambda_a v_{t,a}^2}{2}\right)$$

作成段階においては最終的に，成果物の真の品質 $q_{t,a} \in \mathbb{R}$ が作成者の能力とタスクに依存する性能の和から決定される：

$$q_{t,a} = \mu_a + v_{t,a}$$

3.2 評価段階

我々は，各評価者 r がバイアス $\eta_r \in \mathbb{R}$ を持つとした．厳しい採点をするワーカーほどバイアスが低いものとする．我々はまた，評価者がもつ成果物に対する嗜好をモデル化した．たとえば記事執筆タスクの評価において，長文を好む評価者もいれば簡潔な文章を好むワーカーもいるだろう．このような嗜好を，成果物と評価者の組み合わせに依存するノイズ $w_{t,a}^{(r)} \in \mathbb{R}$ とみなした．このノイズは，平均 0，分散 $1/\kappa_r$（すなわち，精度 κ_r）の正規分布に従う：

$$w_{t,a}^{(r)} \sim \mathcal{N}\left(w_{t,a}^{(r)} \mid 0, 1/\kappa_r\right)$$

評価者 r が作成者 a のタスク t に対する成果物を採点するときに，評価者はまず成果物の潜在的な品質スコア $s_{t,a}^{(r)} \in \mathbb{R}$ を定めるものとする．この品質スコアは成果物の真の品質 $q_{t,a}$，評価者のバイアス η_r，成果物に対する嗜好 $w_{t,a}^{(r)}$ の和とする：

$$s_{t,a}^{(r)} = q_{t,q} + \eta_r + w_{t,a}^{(r)}.$$

評価段階において最終的に観測される採点ラベル $g_{t,a}^{(r)}$ は品質スコアによって決まる離散値であるため，その生成確率を $\Pr[g_{t,a}^{(r)} = d \mid s_{t,a}^{(r)}]$ を導入した．この確率をモデル化するため，段階反応モデル (GRM) を採用した [Samejima 69]．段階反応モデルは項目反応理論において被験者の段階的な得点回答 (例.「当てはまらない」「どちらともいえない」「当てはまらない」) をモデル化するのに用いられている．GRM において，段階反応が得られる確率は $n-1$ 個の 2 値反応モデルを用いて以下のようにモデル化されている：

$$\begin{aligned}\mathrm{GRM}\left(g_{t,a}^{(r)} = d \mid s_{t,a}^{(r)}\right) &= \Pr[g_{t,a}^{(r)} = d \mid s_{t,a}^{(r)}] \\ &= \Pr[g_{t,a}^{(r)} > d-1 \mid s_{t,a}^{(r)}] - \Pr[g_{t,a}^{(r)} > d \mid s_{t,a}^{(r)}]\end{aligned}$$

ここで，$\Pr[g_{t,a}^{(r)} > 0 \mid s_{t,a}^{(r)}] = 1, \Pr[g_{t,a}^{(r)} > n \mid s_{t,a}^{(r)}] = 0$ である．2 値反応モデルにはさまざまな種類があるが，我々は最も基本的なラッシュモデルを採用した：

$$\Pr[g_{t,a}^{(r)} > d \mid s_{t,a}^{(r)}] = \sigma\left(s_{t,a}^{(r)} - b_d\right) = \frac{1}{1 + \exp\left(-(s_{t,a}^{(r)} - b_d)\right)}$$

ここで σ はシグモイド関数，$\{b_d\}_d$ は閾値パラメータとする．提案手法を適用する際には，閾値パラメータは $(b_1, b_2, \cdots, b_{n-1}) = (1, 2, \cdots, n-1)$ と定めた．

4. 品質推定

前節で導入した 2 段階の生成モデルに対して，MAP 推定を用いて成果物の真の品質と他のモデルパラメータを推定する．作成者の能力及び評価者のバイアスの事前分布として，標準正規分布を採用した．また，タスクに依存するノイズの精度 λ_a と，評価者と成果物に依存するノイズの精度 κ_r は正数であるためガンマ分布を事前分布として用いた：

$$\lambda_a \sim \mathrm{Gamma}\left(\lambda_a \mid k, \theta\right), \kappa_r \sim \mathrm{Gamma}\left(\kappa_r \mid k, \theta\right)$$

ここで k と θ はハイパーパラメータである．

目的関数は以下で与えられる対数尤度となる：

$$\begin{aligned}\log L = &\sum_a \left(-\frac{\mu_a^2}{2} + (k-1)\log\lambda_a - \frac{\lambda_a}{\theta}\right) \\ &+ \sum_r \left(-\frac{\eta_r^2}{2} + (k-1)\log\kappa_r - \frac{\kappa_r}{\theta}\right) \\ &- \sum_{t\in\mathcal{T}}\sum_{a\in\mathcal{A}_t} \frac{\lambda_a}{2} v_{t,a}^2 - \sum_{t\in\mathcal{T}}\sum_{a\in\mathcal{A}_t}\sum_{r\in\mathcal{R}_{t,a}} \Big(\frac{\kappa_r}{2} {w_{t,a}^{(r)}}^2 \\ &\qquad + \log\left(\sigma(s_{t,a}^{(r)} - b_{g_{t,a}^{(r)}-1}) - \sigma(s_{t,a}^{(r)} - b_{g_{t,a}^{(r)}})\right)\Big) \\ &+ \sum_{a\in\mathcal{A}} \frac{|\mathcal{T}_a|}{2}\log\lambda_a + \sum_{r\in\mathcal{R}} \frac{|\mathcal{U}_r|}{2}\log\kappa_r\end{aligned}$$

ここで，$\mathcal{T}_a$ は作成者 a が成果物を作成したタスクの集合，$\mathcal{U}_r$ は評価者 r によって採点された成果物の集合を表す．

目的関数は $\{\mu_a\}_a$, $\{\eta_r\}_r$, $\{v_{t,a}\}_{t,a}$, $\{w^{(r)}_{t,a}\}_{t,a,r}$ については凸関数となる．また，$\{\lambda_a\}_a$ と $\{\kappa_r\}_r$ については閉形式で最適解を求めることができる．そこで，$\{\mu_a\}_a$, $\{\eta_r\}_r$, $\{v_{t,a}\}_{t,a}$, $\{w^{(r)}_{t,a}\}_{t,a,r}$ について勾配法での最適化と，$\{\lambda_a\}_a$ と $\{\kappa_r\}_r$ について最適解の計算を交互に行うことで目的関数の最適化を行った．

5. 実験

2段階モデルを用いた提案手法を評価するために，ロゴデザイン，画像説明，翻訳の3種類のタスクを用意しクラウドソーシングサービス Lancers[*1]にてデータを収集し，提案手法を用いて推定した品質の精度を三つの既存手法と比較した．

5.1 データセット

以下3種類のタスクをクラウドソーシングで依頼し実験用データセットを作成した．データサイズの詳細を表 1 に示す．まず作成段階におけるデータを以下のように収集した：(1) ロゴデザイン：Lancers からロゴデザインコンペティションのデータを収集し，投稿されたロゴを作成段階における成果物データとした．(2) 画像説明：SBU Captioned Photo Dataset [*2]からランダムに選んだ 20 個の画像の日本語説明文を記述するタスクを Lancers で依頼した．(3) 翻訳：Wikipedia 日英京都関連文書対訳コーパス[*3] から 20 個の英文をランダムに選択し，日本語に翻訳するタスクを Lancers で依頼した．さらに各タスクについて，Lancers において各成果物につき約 25 人のワーカーに 5 段階評価を依頼し，評価段階のデータを収集した．

5.2 対抗手法

評価段階を導入することにより，既存のラベル統合手法を非定型出力をもつタスクの成果物の品質推定に適用可能となる．我々の提案手法を，多数決と，段階ラベルに対する Dawid-Skene 法 [Raykar 11] と比較した．多数決は，単純な手法でありながらクラウドソーシングにおけるラベル統合において高い性能を示している．Dawid-Skene 法 [Dawid 79] は，ワーカーの能力を考慮してラベル統合を行う手法である．各ワーカーの能力は，真のラベルが与えられたときにワーカーが回答するラベルの条件付き確率で表されており，真のラベルと能力の推定を EM アルゴリズムを用いて交互に行う．本実験では，Rayar らが提案したラベルが段階的である場合の Dawid-Skene 法（「段階ラベル統合手法」）を対抗手法として用いる．これら対抗手法を品質推定に適用する際には，推定した採点ラベルの期待値を品質の推定値として用いた．

我々の2段階モデルと対抗手法は以下の点で異なっている．作成者 a のタスク t に対する成果物の品質を推定する際，多数決はその成果物に対する採点ラベル集合 $\{g^{(r)}_{t,a}\}_r$ しか使用しない．一方 Dawid-Skene 法と我々の手法は，データセット中の全ての採点ラベル $\{g^{(r)}_{t,a}\}_{t,a,r}$ を推定に使用する．また，提案手法は作成者の能力と評価者のバイアスを両方考慮しているが，Dawid-Skene 法は評価者の能力しか考慮していない．

さらに，作成者の能力を考慮することによる効果を確認するために，我々の2段階モデルから作成段階を取り除いたモデル（「評価段階モデル」）とも比較を行った．

5.3 評価方法

推定した品質と正解品質の相関係数を測ることで各手法の評価を行った．また，実用上は複数の成果物の中で最も質が高いものだけがわかれば十分なことも多い．そこで，nDCG@1，すなわち品質最高の成果物の，推定品質と正解品質の比でも評価を行った．品質の正解を取得することは困難であるため，多くのワーカーによる採点結果の平均値で代用した．具体的には，推定には各成果物について一部の採点ラベルだけを使用し，残りの採点ラベルの平均値を正解品質とした．

実験では，成果物辺りの評価者数を $1, 3, 5, 10$ と変化させてその影響を見た．各評価者数について，100 個の部分データをランダムに生成し推定を行った．統計的有意性を確認するためにウィルコクソンの符号順位検定を実施した．また，手法の適用に必要なハイパーパラメータは，全ての実験において $k = 16, \theta = 0.5$ に設定した．

5.4 結果

表 2 に，成果物あたりの評価者数を変化させたときの，正解との相関係数と nDCG@1 を示す．ほとんどの場合において我々の2段階モデルが，対抗手法に対して統計的に有意な精度向上を示した．特に評価者数が少ないときに，対抗手法の精度を大きく改善している．評価者が 1 人しかおらず多数決にもとづく手法が上手く働かない場合であっても，提案モデルは高い精度を示していることは特筆したい．これは，提案モデルが作成者の能力を考慮しているためだと考えられる．評価段階モデルと比較して2段階モデルの方が高精度である点からも，作成者の能力のモデル化が効果的であることが確認できる．

翻訳タスクにおいてのみ，相関係数の評価において提案手法が単純な多数決を下回った．これは，翻訳タスクにおいては評価者の能力のばらつきが小さいことが要因だと考えられる．事実，正解品質と各評価者の採点結果の相関を調べると，ロゴデザインと画像説明タスクでは大きなばらつきがあるのに対して翻訳タスクでは多くのワーカーが高い相関を示していた．このため，多数決が有効に機能したのだと考えられる．

一方 nDCG@1 は，全てのタスクで提案モデルが対抗手法よりも高い値を示していた．提案モデルのこの性質は，同じタスクに対する複数の成果物の中から最も質が高いものを選ぶというクラウドソーシングのシナリオにおいて有用なものである．

6. 関連研究

クラウドソーシングにおいて成果物の品質管理は重要な課題であり，種々の研究が行われている．特に，同じタスクを複数のワーカーに依頼し結果を統合することで質の高い成果物を獲得するという手法が広く用いられている．ワーカーの能力を考慮した Dawid と Skene の手法 [Dawid 79] 以外にも，タスクの難易度を考慮した手法や [Whitehill 09]，ワーカーごとのタスクの難易度を考慮した手法 [Welinder 10] が提案されている．これらの手法は，タスクが2値あるいは多値出力の場合だけを対象としている．

Lin らは，我々と同様に非定型出力をもつタスクを対象とした回答統合手法を提案している [Lin 12]．しかし彼らは，ワーカー間での回答の重複が期待できるタスク（例．算数の問題）だけを対象としており，本研究で対象としたロゴデザインや翻訳のような重複が望めないタスクは対象にしていない．

翻訳タスクに特化して，クラウドソーシングでの品質管理に取り組んだ研究もある [Zaidan 11]．この研究は，我々と同様に非定型出力タスクを対象にしてはいるが，彼らの手法は，

*1 http://www.lancers.jp
*2 http://dsl1.cewit.stonybrook.edu/~vicente/sbucaptions/
*3 http://alaginrc.nict.go.jp/WikiCorpus/index_E.html

表 1: 実験用データセットの詳細

	タスク数	ユニーク作成者数	総成果物数	ユニーク評価者数	成果物あたりの平均評価者数	総採点ラベル数
ロゴデザイン	34	47	710	155	25.0	17750
画像説明	20	20	200	87	25.0	5000
翻訳	20	17	190	71	24.4	4630

表 2: 正解品質との相関係数と nDCG@1 の平均と標準偏差．統計的に有意 ($p < 0.05$) に他の手法を上回る場合を太字で示している

		相関係数				nDCG@1			
	成果物あたりの評価者数	1	3	5	10	1	3	5	10
ロゴデザイン	多数決	0.263 (±0.032)	0.406 (±0.027)	0.481 (±0.023)	0.559 (±0.017)	0.895 (±0.012)	0.910 (±0.011)	0.918 (±0.010)	0.919 (±0.011)
	段階ラベル統合法	0.264 (±0.034)	0.396 (±0.028)	0.471 (±0.025)	0.547 (±0.018)	0.901 (±0.010)	0.910 (±0.012)	0.918 (±0.010)	0.917 (±0.010)
	評価段階モデル	0.270 (±0.033)	0.420 (±0.028)	0.497 (±0.023)	0.569 (±0.017)	0.898 (±0.011)	0.911 (±0.011)	0.920 (±0.010)	0.919 (±0.012)
	２段階モデル	**0.366** (±0.035)	**0.496** (±0.023)	**0.551** (±0.019)	**0.591** (±0.017)	**0.913** (±0.013)	**0.924** (±0.010)	**0.928** (±0.010)	**0.924** (±0.010)
画像説明	多数決	0.418 (±0.063)	0.614 (±0.041)	0.686 (±0.027)	0.766 (±0.020)	0.887 (±0.023)	0.922 (±0.017)	0.935 (±0.017)	0.948 (±0.013)
	段階ラベル統合法	0.394 (±0.058)	0.599 (±0.043)	0.679 (±0.029)	0.724 (±0.023)	0.893 (±0.023)	0.921 (±0.016)	0.936 (±0.015)	0.946 (±0.014)
	評価段階モデル	0.456 (±0.062)	0.648 (±0.038)	0.715 (±0.027)	0.783 (±0.020)	0.898 (±0.020)	0.926 (±0.016)	0.939 (±0.015)	0.950 (±0.012)
	２段階モデル	**0.627** (±0.054)	**0.746** (±0.025)	**0.781** (±0.021)	**0.809** (±0.015)	**0.921** (±0.019)	**0.941** (±0.014)	**0.948** (±0.015)	**0.958** (±0.010)
翻訳	多数決	0.601 (±0.041)	**0.783** (±0.025)	**0.840** (±0.016)	**0.884** (±0.011)	0.877 (±0.026)	0.936 (±0.018)	0.947 (±0.015)	0.959 (±0.011)
	段階ラベル統合法	0.563 (±0.043)	0.748 (±0.027)	0.785 (±0.020)	0.797 (±0.015)	0.883 (±0.025)	0.937 (±0.020)	0.951 (±0.016)	0.959 (±0.012)
	評価段階モデル	0.556 (±0.043)	0.751 (±0.023)	0.818 (±0.017)	0.871 (±0.012)	0.895 (±0.026)	0.939 (±0.017)	0.948 (±0.014)	0.958 (±0.011)
	２段階モデル	**0.622** (±0.032)	0.761 (±0.018)	0.813 (±0.016)	0.865 (±0.012)	**0.930** (±0.023)	**0.954** (±0.013)	**0.960** (±0.010)	0.962 (±0.012)

教師データや成果物の特徴表現に対する事前知識が必要とする点で我々とは異なるアプローチを採っている．

我々は，各ワーカーが並列作業する状況を取り扱ったが，Dai らはワーカーが前のワーカーの作業結果を改善していくという逐次作業における品質管理手法を提案している [Dai 11]．彼らの手法も教師情報を必要とする点で我々の手法とは異なっている．

7. むすび

我々は，非定型出力をもつクラウドソーシングタスクにおいて成果物の品質を統計的に推定する手法を提案した．作成段階と評価段階それぞれにおける生成過程をモデル化した２段階モデルが，ロゴデザイン，画像説明，翻訳タスクにおいて既存手法よりも高い品質精度を示すことを実験で確認した．

謝辞

本研究の一部は内閣府最先端研究開発プログラム (FIRST)「超巨大データベース時代に向けた最高速データベースエンジンの開発と当該エンジンを核とする戦略的社会サービスの実証・評価」の助成を受けたものである．

参考文献

[Dai 11] Dai, P., Mausam, , and Weld, D. S.: Artificial Intelligence for Artificial Artificial Intelligence, in *Proc. of AAAI* (2011)

[Dawid 79] Dawid, A. P. and Skene, A. M.: Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm, *Journal of the Royal Statistical Society. Series C (Applied Statics)*, Vol. 28, No. 1, pp. 20–28 (1979)

[Ipeirotis 10] Ipeirotis, P. G.: Analyzing the Amazon Mechanical Turk marketplace, *ACM XRDS*, Vol. 17, No. 2 (2010)

[Lin 12] Lin, C., Mausam, M., and Weld, D.: Crowdsourcing Control: Moving Beyond Multiple Choice, in *Proc. of UAI* (2012)

[Raykar 11] Raykar, V. C. and Yu, S.: Ranking annotators for crowdsourced labeling tasks, in *Proc. of NIPS* (2011)

[Samejima 69] Samejima, F.: Estimation of latent ability using a response pattern of graded scores., *Psychometrika Monograph Supplement* (1969)

[Welinder 10] Welinder, P., Branson, S., Belongie, S., and Perona, P.: The Multidimensional Wisdom of Crowds, in *Proc. of NIPS* (2010)

[Whitehill 09] Whitehill, J., Ruvolo, P., Wu, T., Bergsma, J., and Movellan, J.: Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise, in *Proc. of NIPS* (2009)

[Zaidan 11] Zaidan, O. F. and Callison-Burch, C.: Crowdsourcing translation: professional quality from non-professionals, in *Proc. of ACL-HLT* (2011)