1L5-OS-09b-4in1

一対比較を用いたクラウドソーシングの成果物品質推定法

Quality estimation of crowdsourced artifacts using pairwise comparisons

砂長谷健 Takeru Sunahase　馬場雪乃 Yukino Baba　鹿島久嗣 Hisashi Kashima

京都大学大学院情報学研究科知能情報専攻
Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University

Crowdsourcing offers a new working style and the application of crowdsourcing has been expanded. One of the important problems in the use of crowdsourcing is the quality of results. The common approach to tackle this problem is to introduce redundancy, that is, to request multiple workers to work on the same tasks and aggregate their results to improve the quality. However, the approach cannot be applied to the tasks whose artifacts are unaggregatable, such as article writing and logo designing. In previous work, the two-stage procedure was introduced; the artifacts gathered in creation stage are evaluated by multiple workers in review stage. Then, the quality of the unaggregatable artifacts can be estimated. In this study, we especially focus on pairwise comparison in the two-stage procedure. We apply the Kleinberg's HITS algorithm to the pairwise quality estimation, and investigate the performance of our methods.

1. はじめに

クラウドソーシングとは、インターネット上の不特定多数の人々に作業を外注する仕組みである。依頼者は必要に応じて安価な人的資源が利用できる。受託者は時間を選ばず仕事ができる。これらの要因によりクラウドソーシングの利用は拡大し、様々な種類の仕事が依頼されている。

クラウドソーシングには**作業結果の信頼性の担保**という重要な課題が存在する。作業を行うワーカーには作業能力や意欲にばらつきがあるため、正しい結果が得られるとは限らず、信頼性を高める工夫が必要となってくる。信頼性を高める方法には**作業の冗長化**がある。これは同じ作業に対して複数の結果を集め、集めた複数の結果を何らかの方法で統合することで作業結果の信頼性を高める方法である。しかし、作業の冗長化にはいくつか問題がある。その一つは統合できない結果を返す作業が存在することである。クラウドソーシングには一意に統合できない文章やデザインといった**成果物**が集まる作業が多く存在する [Ipeirotis 10]。

本研究の目的は統合が行えない成果物の品質を効率的に推定することにある。既存研究では、成果物を作る作業 (作成段階) に続いて、成果物の良さを評価する作業 (評価段階) を行うという**二段階過程**を取り入れた [Baba 13]。二段階過程を用いることで、評価段階で与えられる評価値から品質を測る指標が得られ、品質の推定が可能となる。しかし、取り入れられた評価段階の信頼性確保に再び冗長化が必要となる (図 1)。二段階のそれぞれで作業の冗長化を行うと大きな作業コストがかかる。このコストをワーカーの作業能力を考慮することで削減することを目指す。これまでの研究においてワーカーの作業能力を考慮することで効率的に結果を統合できる手法が提案されてきた [Dawid 79, Whitehill 09, Welinder 10]。二段階過程においても、新たに作成者能力を加味することで良い成果を得ている [Baba 13]。ところがこの既存研究では、成果物の評価方法は絶対評価に限られていた。絶対評価は品質の差が小さい成果物どうしの区別、細かな差が品質に大きく関わってくる成果物の評価には向いていないと考えられる。そこで本研究では相対評価を、そのなかでも一対比較を用いる。一対比較の品質推定には、これまでに評価者能力を利用した手法が存在するが [Chen 13]、二段階過程の作成者能力を考慮した手法は存在しない。

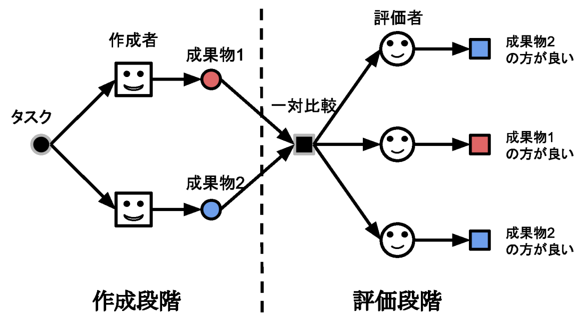

図 1: 作成と評価の二段階過程における作業の冗長化

本研究では一対比較において、作成と評価の二段階過程を用いた成果物の品質推定法を提案する。まず、Web ページのリンク解析アルゴリズムとして提案された HITS アルゴリズム [Kleinberg 99] のモデル構造を評価段階のみの品質推定に応用する (3 章)。この手法を Pairwise HITS と名付ける。続いて、品質推定のモデルに作成者の能力を取り入れ二段階過程の場合に拡張する (4 章)。この手法を二段階 Pairwise HITS と名付ける。さらに、本研究の提案手法が有効に機能することを実データを用いた実験により確認した (5 章)。

本研究の貢献は以下の三つである。

- HITS アルゴリズムのモデルを利用した一対比較の品質推定法を提案した
- 一対比較の品質推定において二段階過程の作成者能力を取り入れた品質推定法を提案した
- 実データを用いた実験により、提案手法が一対比較における品質推定の効率向上に有効であることを確認した

連絡先: 砂長谷健, stakeru@ml.ist.i.kyoto-u.ac.jp

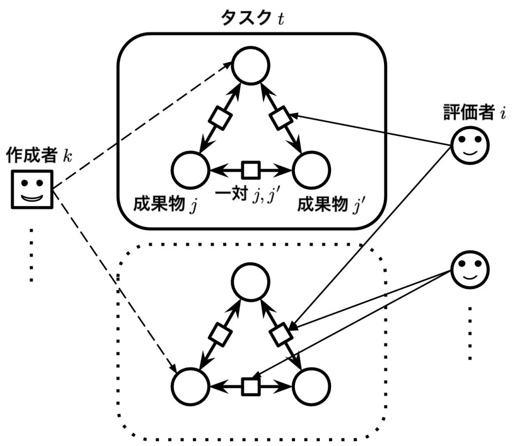

図 2: 問題設定のイメージ図

2. 問題設定

本研究では評価段階で得られる複数の一対比較を用いて、作成段階で作成された複数の成果物の品質値を推定する問題を扱う。作成段階では m 個の成果物作成タスクから n 個の成果物が作られる。作成タスクを $t \in \{1, \ldots, m\}$ で、成果物を $j \in \{1, \ldots, n\}$ で表す。各作成タスク t では n_t 人の異なる作成者により成果物が作られる。作成タスク t の成果物を $j_t \in \{1, \ldots, n_t\}$ で表す。ここで $n = \sum_{t=1}^{m} n_t$ である。続く評価段階では比較ペアそれぞれに複数の評価者により優劣をつけられる。一対比較は同じ作成タスクで作成された成果物間に発生する。一対比較の比較対象は一対の成果物なので、作成タスク t における比較ペア数は $\binom{n_t}{2}$ 個となる。成果物 j と成果物 j' の一対比較結果は、「成果物 j の方が成果物 j' より良い」と「成果物 j' の方が成果物 j より良い」の二種類があり、それぞれの意見を (j, j')、(j', j) と表す。また、作成者は合計 p 人, 評価者は合計 q 人とし、それぞれ $k \in \{1, \ldots, p\}$, $i \in \{1, \ldots, q\}$ で表す。問題設定の概要を図 2 に示す。

本研究の目的は観測された一対比較集合から、作成タスクごとの成果物の真の品質ランキングを推定することである。

3. Pairwise HITS

まず評価段階のみに着目して、HITS アルゴリズム [Kleinberg 99] のハブとオーソリティの概念を、クラウドソーシングの文脈に応用して、評価段階における評価者能力と成果物品質をモデリングする (3.1 節)。続いて、提案手法 Pairwise HITS のアルゴリズムについて述べる (3.2 節)。

3.1 能力と品質のモデリング

HITS アルゴリズムにおける、リンクをはる、はられるという関係はクラウドソーシングで評価をする、されるの関係と類似する点がある。ページを成果物と評価者、ハブを評価能力、オーソリティを品質に置き換える。すると以下のような、評価能力と品質の関係性と、ハブとオーソリティの関係性との共通点が存在する。

- 能力が高い評価者ほど高品質な成果物を正しく良いと評価する
- 高品質な成果物ほど能力が高い評価者に正しく良いと評価される

モデリングはこの二つの仮定をもとに行う。本研究では成果物の評価方法に一対比較を用いるため、従来の HITS 関連アルゴリズムとは異なる新たな定式化を行う。ここで、各評価者の能力を r_i とし、各成果物の品質を a_j とする。

「能力が高い評価者ほど高品質な成果物を正しく良いと評価する」という仮定に従って、評価者の能力をモデリングする。各成果物の品質 a_j が決まっているとする。仮定より、高い能力を持つ評価者は、$a_j > a_{j'}$ である成果物 j と j' に対し「成果物 j の方が j' よりも良い」と正しく比較する回数が多い。評価者はそれぞれ比較回数が異なるので、評価者ごとの全比較回数を考慮して正しく能力を表現する必要がある。つまり、評価者 i の能力 r_i は評価者の全比較のなかで正しく比較した割合とし、次式で表す。

$$r_i = \frac{|\{(j, j') \in V_i \mid a_j > a_{j'}\}|}{|V_i|} \tag{1}$$

ただし、V_i は評価者 i の与えた比較の集合とする。

「高品質な成果物ほど能力が高い評価者に正しく良いと評価される」という仮定に従って、成果物の品質をモデリングする。各評価者の能力 r_i が決まっているとして、評価者の判断は評価者の能力 r_i に重み付けされるとする。一対比較の場合、例えば成果物 j と j' の一対は「成果物 j の方が j' よりも良い」という判断と、「成果物 j' の方が j よりも良い」という判断の、相反する二種類の意見を集める。二種類の意見は集めた判断から、その重みの合計によってそれぞれの信憑性を測ることができる。仮定から、一対の成果物のうち高品質な成果物ほど多くの評価を集めると考えられる。さらに能力の高い評価者の評価ほど高品質な成果物に集まると考えられる。本研究では以上の議論を踏まえ、「成果物品質の差は集まる評価の差に等しい」という仮説を立てる。この仮説を次式で表す。

$$a_{j_t} - a_{j'_t} = \sum_{i \in R_{(j_t, j'_t)}} r_i - \sum_{i \in R_{(j'_t, j_t)}} r_i \tag{2}$$

ただし、$R_{(j,j')}$ は「j よりも j' の方が良い」といった評価者の集合とする。成果物 j の品質 a_j は式 (2) の解である。

3.2 パラメータ推定

ハブとオーソリティに変わる、評価能力と品質のモデルを設計したので、次にモデルのパラメータ r_i, a_j の推定法を示す。パラメータ推定には HITS と同様に評価能力 r_i の更新と品質 a_j の更新を交互に行う。品質 a_j の更新に関しては、式 (2) のモデルが制約式の形を取っているので全ての制約を二乗誤差最小で満たす解 a_j^* を更新する値とする。

提案手法 Pairwise HITS のアルゴリズムについてまとめる。

1. パラメータ $\{r_1, \ldots, r_q\}$ に初期値を与える
2. a_j を制約式 (2) から求められる解 a_j^* に更新する
3. r_i を式 (1) で更新する
4. $||\boldsymbol{r}||_2^2 = 1$ となるように $\boldsymbol{r} = \{r_1, \ldots, r_q\}$ を正規化する
5. 2〜4 をパラメータが収束するまで、または決められた回数行う

4. 二段階 Pairwise HITS

前章で導入した手法 Pairwise HITS をさらに二段階過程に拡張するために、成果物品質モデルの一部を修正し、作成者能力をモデリングする (4.1 節)。続いて、提案手法の二段階 Pairwise HITS のアルゴリズムについて述べる (4.2 節)。

4.1 能力と品質のモデリング

作成段階と評価段階の二段階過程で品質推定をする場合、Pairwise HITS には新たに作成者の能力を考える。作成能力も Pairwise HITS と同様に、以下のような仮定が置ける。

- 能力が高い作成者ほど高品質な成果物を作る
- 高品質な成果物ほど能力が高い作成者に作られる

二段階過程でも、ワーカーの能力は HITS のモデルに組み込むことができ、評価能力と同様にハブを作成能力に置き換える。新たにモデルに加わるのは作成能力で、作成能力に関係がある成果物品質モデルの修正も必要となる。評価者の能力モデルは変わらない。ここで、各作成者の能力を c_k とする。

「能力が高い作成者ほど高品質な成果物を作る」という仮定に従って、作成者の能力をモデリングする。各成果物の品質 a_j が決まっているとする。仮定から、作成者能力は作った成果物から判断される。作成者はそれぞれ作成回数が異なるので、作成者ごとの全作成回数を考慮して正しく作成能力を表現する必要がある。つまり、作成者 k の能力 c_k は作った成果物の品質の平均値とし、次式で表す。

$$c_k = \frac{1}{|A_k|} \sum_{j \in A_k} a_j \tag{3}$$

ただし、A_k は作成者 k が作った成果物の集合とする。

「高品質な成果物ほど能力が高い評価者に正しく良いと評価される」という仮定に加え、「高品質な成果物ほど能力が高い作成者に作られる」という新たな仮定に従って、成果物品質のモデルを修正する。二段階過程に拡張することによって成果物の品質モデルは二つの仮定が混在する。本研究では、二つの仮定をハイパーパラメータ λ を用いて足し合わせる。前者の仮定は、前章の式 (2) で表現される。後者の仮定は、成果物は作成者の能力に依ることを示している。以上より、成果物の品質モデルを次式で表す。

$$a_j = (1-\lambda)a_j^* + \lambda c_k \tag{4}$$

ただし、a_j^* は式 (2) の制約式を二乗誤差最小で満たす a_j の解であり、c_k は成果物 j の作成者、$0 < \lambda < 1$ とする。

4.2 パラメータ推定

作成能力と評価能力、品質のモデルを設計したので、次にモデルのパラメータ c_k, r_i, a_j の推定法を示す。パラメータ推定には HITS と同様に c_k, r_i の更新と a_j の更新を交互に行う。

提案手法二段階 Pairwise HITS のアルゴリズムについてまとめる。

1. パラメータ $\{c_1, \ldots, c_p\}, \{r_1, \ldots, r_q\}$ に初期値を与える
2. 決められた値 λ を用いて、a_j を式 (4) で更新する
3. c_k を式 (3) で、r_i を式 (1) で更新する
4. $||\boldsymbol{c}||_2^2 = 1$、$||\boldsymbol{r}||_2^2 = 1$ となるように $\boldsymbol{c} = \{c_1, \ldots, c_p\}$ と $\boldsymbol{r} = \{r_1, \ldots, r_q\}$ をそれぞれ正規化する
5. 2〜4 をパラメータが収束するまで、または決められた回数行う

5. 実験

二つの提案手法の性能を評価するためにベースライン手法との比較を行った。比較には二種類のデータセットを用いて、提案手法の有用性を検証した。

5.1 データセット

実験には既存研究 [Baba 13] で用いられたデータ *1(画像説明タスク) と、クラウドソーシングプラットホームであるランサーズ *2 で集めたデータ (記事要約タスク) を用いた。データサイズの詳細は表 1 に示す。

5.1.1 画像説明タスク

画像説明タスクでは、ある画像の説明文に複数の五段階評価値が与えられる。このタスクの評価データは五段階評価であるため、同じ評価者の二つの評価値を比べることで擬似的に一対比較評価を生成した。本実験では、二つの五段階評価値を比べた結果が同値だったときランダムに片方が良いと選択したとする。利用する評価データは一対比較の実データではないが、二段階過程での評価をもとに生成したデータであるので二段階過程特有の偏りがデータにあると考えられる。

5.1.2 記事要約タスク

記事要約タスクの目標は、与えられたニュース記事を要約することである。まず、成果物として記事の要約文を集めた。要約対象となるニュース記事はハフィントンポスト *3 より選んだ。記事の内容は一定のジャンルに固定されないように選択し、記事の文字数は 600 字から 2000 字まである。要約はそれぞれ日本語 80 字以上 110 字以内で記述される。評価段階では評価者に同じ記事に関する二つの要約を比べて優劣をつける作業をランサーズ上で依頼した。要約の評価指標には特に基準を設けなかった。

5.2 比較手法

本研究の提案手法 Pairwise HITS と二段階 Pairwise HITS を Bradley-Terry モデル (BT)[Bradley 52] と比較した。BT は一対比較を比較対象の真の品質に関する確率モデルとして表すことで、最尤法を用いて品質を求める。一対比較評価データのみを用いて評価対象の品質を測るため、本実験では BT をベースライン手法に位置付ける。

5.3 評価方法

推定した品質から成るランキングと正解品質から成るランキングの類似性を二つの指標で測ることで各手法を評価した。一つ目は推定と正解ランキングの全順序の近さを測るために、スピアマン順位相関を用いた。もう一つは、正解と推定のランキングの上位の近さを測るために、nDCG@1 を用いた。真の品質を得ることは難しいので、集めた全評価データをもとに BT で推定した品質値を正解品質とし、評価データの一部に各手法を適用した。

実験では集めた評価データの一部をもとに、各手法で品質を推定した。1 比較ペアごとに決まった評価者数となる部分評価データ 100 個をランダムに生成し推定を行った。そして、1 比較ペアごとの評価者数を変化させ、手法による違いを計測した。さらに、統計的有意性を確認するためにウィルコクソンの符号付順位和検定を行った。本実験においてハイパーパラメータ λ は、画像説明タスクで 0.2、記事要約タスクで 0.1 に設定し、能力パラメータの初期値は全て 0 に設定した。

5.4 結果

1 比較ペアごとの評価者数を変化させたときの、正解との順位相関と nDCG@1 をそれぞれ表 2 と表 3 に示す。画像説明タスクにおいて順位相関、nDCG@1 どちらを見ても、二段階 Pairwise HITS が他手法に比べ統計的に有意な精度向上を示し

*1 http://yukino.moo.jp/SQEGCT/
*2 http://www.lancers.jp/
*3 http://www.huffingtonpost.jp/

表 1: 実験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	比較一対あたりの平均評価者数	総比較数
画像説明	20	20	200	87	17.3	16314
記事要約	14	11	70	45	26.3	3678

表 2: 画像説明タスクの比較結果: 数値は全作成タスク平均の平均値と標準偏差である。各比較者数についてウィルコクソン符号付順位和検定で有意水準 5% の有意差が認められた最良の結果を太字で表記している。また、各比較者数ごとに BT と比べ有意差が認められた結果に対して † を付けた。

	順位相関				
比較者数	1	2	3	4	5
BT	0.650 ± 0.045	0.767 ± 0.032	0.820 ± 0.024	0.859 ± 0.020	0.879 ± 0.022
PairwiseHITS	†0.660 ± 0.042	†0.774 ± 0.032	†0.824 ± 0.024	†0.862 ± 0.021	0.879 ± 0.021
二段階 Pairwise HITS	†$\mathbf{0.733} \pm 0.030$	†$\mathbf{0.804} \pm 0.025$	†$\mathbf{0.842} \pm 0.021$	†$\mathbf{0.869} \pm 0.018$	†$\mathbf{0.885} \pm 0.018$
	nDCG@1				
比較者数	1	2	3	4	5
BT	0.837 ± 0.035	0.888 ± 0.030	0.911 ± 0.027	0.930 ± 0.026	0.941 ± 0.022
PairwiseHITS	0.838 ± 0.036	0.891 ± 0.032	0.911 ± 0.028	†0.933 ± 0.024	0.938 ± 0.022
二段階 Pairwise HITS	†$\mathbf{0.860} \pm 0.032$	†$\mathbf{0.898} \pm 0.030$	†$\mathbf{0.919} \pm 0.025$	0.932 ± 0.025	0.939 ± 0.018

表 3: 記事要約タスクの比較結果: 表記は表 2 と同様。

	順位相関				
比較者数	1	3	5	7	9
BT	0.561 ± 0.092	0.714 ± 0.061	0.770 ± 0.057	0.822 ± 0.046	0.846 ± 0.043
PairwiseHITS	0.550 ± 0.095	0.712 ± 0.065	0.773 ± 0.052	0.823 ± 0.045	0.846 ± 0.041
二段階 Pairwise HITS	†$\mathbf{0.580} \pm 0.091$	0.713 ± 0.065	0.768 ± 0.052	0.822 ± 0.044	0.845 ± 0.040
	nCDG@1				
比較者数	1	3	5	7	9
BT	0.858 ± 0.051	0.912 ± 0.041	0.940 ± 0.032	0.955 ± 0.024	0.964 ± 0.023
PairwiseHITS	0.854 ± 0.053	0.911 ± 0.042	0.941 ± 0.030	0.955 ± 0.025	0.963 ± 0.022
二段階 Pairwise HITS	0.859 ± 0.051	0.911 ± 0.044	0.940 ± 0.029	0.954 ± 0.024	0.964 ± 0.022

た。特にワーカー数が少ないときに大きな精度改善が見受けられる。BT に対する Pairwise HITS と二段階 Pairwise HITS の精度向上を比べると、作成者の能力を加味したことが精度向上に大きく繋がっていることが言える。しかし、記事要約タスクにおいて、各手法あまり差が見られず、ほとんど有意差は見られなかった。これは作成者による能力の差があまりないことが原因だと考えられる。

6. 結論

本研究では、作成と評価の二段階過程において、一対比較を用いて成果物の品質推定を行う手法を提案した。ベースライン手法との比較を行い、一対比較を用いた品質推定法においても、二段階過程の作成者能力の導入が精度向上につながることを示した。また、HITS は絶対評価における品質推定だけでなく、相対評価の品質推定にも応用できることを示した。

参考文献

[Baba 13] Baba, Y. and Kashima, H.: Statistical quality estimation for general crowdsourcing tasks, in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2013)

[Bradley 52] Bradley, R. A. and Terry, M. E.: The rank analysis of incomplete block designs: I. the method of paired comparisons, *Biometrika*, Vol. 39, No. 3/4 (1952)

[Chen 13] Chen, X., Bennett, P. N., Collins-Thompson, K., and Horvitz, E.: Pairwise ranking aggregation in a crowdsourced setting, in *Proceedings of the 6th ACM international conference on Web Search and Data Mining* (2013)

[Dawid 79] Dawid, A. P. and Skene, A. M.: Maximum likelihood estimation of observer error-rates using the EM algorithm, *Applied Statistics*, Vol. 28, No. 1 (1979)

[Ipeirotis 10] Ipeirotis, P. G.: Analyzing the Amazon Mechanical Turk marketplace, *ACM XRDS*, Vol. 4, No. 2 (2010)

[Kleinberg 99] Kleinberg, J. M.: Authoritative sources in a hyperlinked environment, *Journal of the ACM*, Vol. 46, No. 5 (1999)

[Welinder 10] Welinder, P., Branson, S., Belongie, S., and Perona, P.: The multidimensional wisdom of crowds, in *Advances in Neural Information Processing Systems* (2010)

[Whitehill 09] Whitehill, J., Wu, fan T., Bergsma, J., Movellan, J. R., and Ruvolo, P. L.: Whose vote should count more: Optimal integration of labels from labelers of unknown expertise, in *Advances in Neural Information Processing Systems* (2009)