

# クラウドソーシングにおける マイクロタスクの質問文の改善手法の提案

## A Method for Improving Microtask-based Crowdsourcing

丹治 寛佳\*<sup>1</sup>    清水 伸幸\*<sup>2</sup>    森嶋 厚行\*<sup>3</sup>    北川 博之\*<sup>4</sup>  
Hiro Yoshi Tanji    Nobuyuki Shimizu    Atsuyuki Morishima    Hiroyuki Kitagawa

\*<sup>1</sup>筑波大学 システム情報工学研究科

Graduate School of Systems and Information Engineering, University of Tsukuba

\*<sup>2</sup>ヤフー株式会社 Yahoo!JAPAN 研究所

Yahoo! JAPAN Research, Yahoo Japan Corporation

\*<sup>3</sup>筑波大学 知的コミュニティ基盤研究センター/図書館情報メディア系

Research Center for Knowledge Communities, Faculty of Library, Information and Media Science, University of Tsukuba

\*<sup>4</sup>筑波大学 システム情報系

Faculty of Engineering, Information and Systems, University of Tsukuba

It is not easy to appropriately design microtasks. The answers to the question described in a microtask are often beyond the expectation of requesters when the questions are not clear to workers. This paper proposes a method to apply crowdsourcing to improve questions in microtasks.

## 1. はじめに

近年、群衆の知や力を利用して、計算機だけでは処理が困難な問題を解決するクラウドソーシングが注目を集めている。それに伴い、クラウドソーシングのための基礎機能を提供するクラウドソーシングプラットフォームが登場してきた。その代表的なものとして、**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**がある。その例として、Amazon Mechanical Turk[1] や、Yahoo!クラウドソーシング [11] などがある。

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は次のように利用される。まず、問題解決を望む人(リクエスタ)がクラウドソーシングプラットフォームが持つタスクプールにマイクロタスク(以下**タスク**)を登録する。次に、登録されたタスクを、不特定多数の人々(ワーカー)が処理する。

一般的に、マイクロタスクには質問文とデータが表示される。例えば、図1のタスクは、「この写真は不適切ですか。」という質問文と画像データから構成されている。しかし、クラウドソーシングにおいて、適切なタスクの質問文を作成するのは必ずしも容易ではない。なぜなら、タスクの質問文の多くは自然文で表現されるため、リクエスタの意図がワーカーに伝わらない可能性があるからである。例えば、図1のタスクの場合、リクエスタが「不適切」の意味を「ぼやけて見づらい」という意図で表現したとしても、「暴力的な表現」や「性的な表現」とワーカーが解釈し、意図しないタスク結果になる可能性がある。

このように、質問文が不適切な場合、タスクの目的がワーカーに伝わらずタスク結果の品質が下がることがある [7]。また、この問題は、ワーカーの作業品質を推定するためにあらかじめ答えを用意した質問文(ゴールドクエスチョン)の効果にも影響を及ぼす。Yahoo!クラウドソーシングの内部調査によると、適切なワーカーでも間違ってしまうような質問文のゴールドクエスチョンは必ずしも少なくない。事例は [10] にある。

連絡先: 丹治 寛佳, 筑波大学 システム情報工学研究科,  
〒 305-8550 茨城県つくば市春日 1-2, 029-859-1559,  
hiro Yoshi.tanji.2012b@mmlab.info

そこで、本論文ではマイクロタスクの質問文の改善手法の提案を行う。提案手法の概要は次の通りである。(1) 質問文、データ、答えの組を用意する。(2) 質問文を改善するための情報を得る**改善タスク**をクラウドソーシングする。これは、答えとデータから、元の質問文を推測させるタスクである。(3) 改善タスクの結果を、質問文の改善に利用する。結果の利用方法としては、直接、質問文の改善に利用する場合と、他の質問文の改善に利用可能な**知識ベース**を構築し、間接的に、他の質問文の改善に利用する場合がある。

**関連研究.** 我々の知る限り、質問文そのものの改善に関する研究は存在しないが、次のような関連する研究が行われてきた。

**(1) よりよいマイクロタスク設計・品質向上を目的とした研究.** これらの多くは、タスクフロー設計 [6][8] や、タスクのインセンティブ構造 [3][9]、タスク割当 [4] などの議論を行っている。これらの手法と本論文の提案手法を組み合わせることで、タスク結果のさらなる品質向上が可能であると考えられる。

**(2) 質問応答システムに関する研究.** [2] では、自然言語処理(NLP)とオントロジーを用いて、質問文を機械が処理可能な表現に変換することで、データベースから回答を取得する。本研究では、これとは異なり、質問文を、対象が人であることは変えず、より良い表現に変換するものである。

## 2. 提案手法

提案手法では、まず、答えとデータから質問文を推測させるタスク(**改善タスク**)のクラウドソーシングを行う。その結果を用いて、リクエスタのタスクの質問文の改善と、改善のためのデータベース(**知識ベース**)の作成を行う。以下では、まず用語などの説明を行い、その後、提案手法を詳細に説明する。

### 2.1 用語と形式化

**マイクロタスクテンプレートとマイクロタスク.** 一般に、マイクロタスクは、同種のタスクで共通するテンプレート(質問文など)および、タスク毎に異なるデータ(画像など)から構成される(図1)。本論文では、マイクロタスク  $t_i$  を  $t_i = \text{template}_j(q_j, A_j)[d_i]$  と表現する。ここで、 $q_j$  はマイクロタス



図 1: マイクロタスクの例

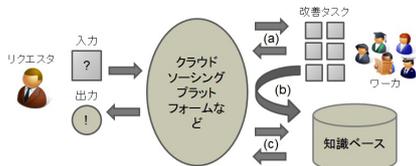


図 2: 提案手法概要

クの質問文,  $A_j$  は, 質問文  $q_j$  に対する回答の候補の集合,  $d_i \in D$  は  $t_i$  毎にワークに提示するデータである.  $template_j(q_j, A_j)$  はテンプレートを表す. 例えば, 図 1 のタスクは,  $template_j('この写真は不適切ですか?', \{ 'はい', 'いいえ', 'わからない' \})[p]$ , ( $p \in Photo$ ) と表現される. 同じテンプレートを持つタスクの集合は, タスクテンプレート  $template_j(q_j, A_j)$  および  $d_i$  のドメイン  $D_j$  を与えることにより表現することができる.

**ゴールドデータ.** クラウドソーシングのデータ品質管理のために, あらかじめ答えがわかっているマイクロタスクを用意することが多い. その答えはゴールドデータと呼ばれる. 本論文では, ゴールドデータを関数  $C_j : D_j \rightarrow A_j$  として表現する.

## 2.2 全体像

提案手法のおおまかな流れを図 2 に示す. 提案手法は 2 種類の支援 (支援 A, 支援 B) を行う.

**入力.** タスクテンプレート  $template_j(q_j, A_j)$ . 支援 A の場合には追加としてデータ集合  $D_j$  およびゴールドデータ  $C_j$  を必要とする.

**出力.** 新しい質問文  $q_j^1 \dots q_j^n$ .

詳細は 2.4 節と 2.6 節で説明するが, 本手法では, 入力を与えられると, 次の 2 種類の支援手法により質問文を生成する.

**支援 A** 質問文改善のためのマイクロタスク (改善タスク) を生成しクラウドソーシングする (図 2(a)). その結果に基づき, 新しい質問文を生成し, 出力する. また, 同時に, 支援 B で使用する知識ベースにデータを格納する (図 2(b)).

**支援 B** 知識ベースに既に格納されているデータを用いて, 新しい質問文を出力する (図 2(c)).

1 章でも説明したが, 本手法の特徴は, 改善タスクとして「答えを見せて質問を推測させる」というタスクを利用することである. これにより, 直接的な質問文改善, およびその後の支援 B による質問文改善のための知識ベースの構築を行う.

## 2.3 改善タスク

本手法では, 入力されたタスクを改善するためのタスクとして, 「改善タスク」(図 3) を生成する. これは, 「答えとデータを見せて質問を推測させる」というマイクロタスクである. 図 3 の例の場合には, 想定される入力としては「この写真に写っているのは人工物ですか?」などが考えられる.

**改善タスクの生成と依頼.** 入力 ( $template_j(q_j, A_j)$ ), データ集合  $D_j$ , およびゴールドデータ  $C_j$ ) が与えられたとき, 改善タスク

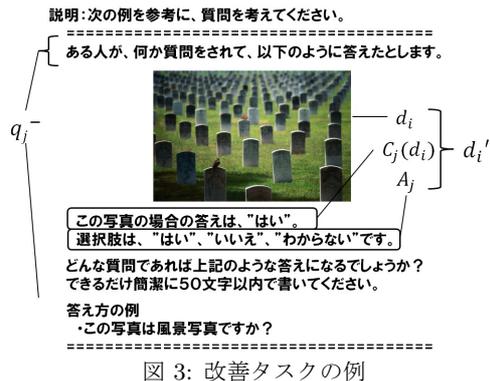


図 3: 改善タスクの例

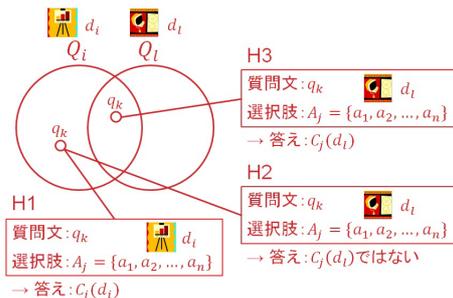


図 4: ヒューリスティクス

の集合  $IT_j$  が生成される. 具体的には, 各改善タスク  $t_i^- \in IT_j$  は  $d_i \in D_j$  毎に生成され,  $t_i^- = template_j^-(q_j^-, A_j^-)[d_i]$  となる. ここで,  $t_i^-$  の各構成要素は次の通りである (図 3).

- $q_j^- =$  「ある人が .. 風景写真ですか?」
- $d_i^- = (A_j, d_i, C_j(d_i))$  ただし  $d_i \in D_j$
- $A_j^-$  は存在しうる質問文の集合

図 2(a) では,  $IT_j$  中のすべての改善タスク  $t_i^-$  をクラウドソーシングする. このとき, 各  $t_i^-$  の処理は, 複数のワークに依頼する.

**改善タスクの結果.** ある  $t_i^-$  を,  $t_i^-$  から得ることのできる質問文を全て得るのに十分に大きな人数のワークに依頼したと仮定する. このとき, 結果として返る質問文の集合を  $Q_i$  と表す. また, 簡単化のため, 同じ意味の質問文は同一の表現であると仮定する.  $t_i^-, t_l^- \in IT_j$  としたとき,  $q_k \in Q_i \cup Q_l$  に関して次のヒューリスティクスを考えることができる.

**H1.**  $q_k \in Q_i$  ならば,  $template_k(q_k, A_j)[d_i]$  の答えは  $C_j(d_i)$  となる (図 4 H1).

**H2.**  $q_k \in Q_i - Q_l$  ならば,  $template_k(q_k, A_j)[d_i]$  の答えは  $C_j(d_i)$  とはならない (図 4 H2).

また, **H1** から, 次のヒューリスティクスを導出することができる. さらに, このヒューリスティクスは改善タスクが  $n$  個の場合に一般化できる.

**H3.**  $q_k \in Q_i \cap Q_l$  ならば,  $d \in \{d_i, d_l\}$  に対して  $template_k(q_k, A_j)[d]$  の答えはそれぞれ  $C_j(d_i)$  と  $C_j(d_l)$  となる (図 4 H3).

本手法では, **H3** を用いて, 改善タスクによる質問文の生成を行い, **H2** を用いて, 今後の質問文改善のための知識ベースの作成 (図 2(b)) を行う.

**自然言語による質問文の積集合・和集合の計算.** 実際には, 各  $Q_i$  に含まれる質問文は自然言語による質問文であるため, 前

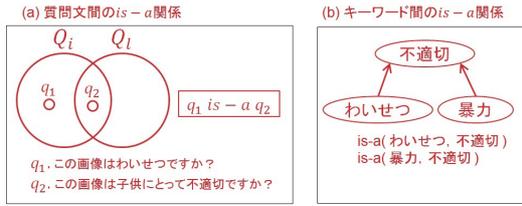


図 5: is-a 関係

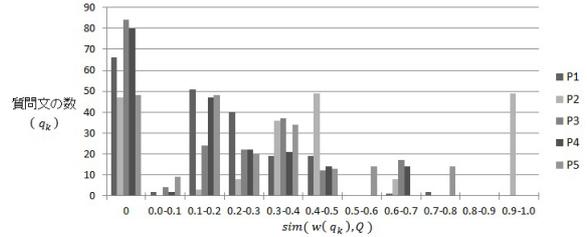


図 6: 質問文  $q_k$  の  $sim(w(q_k), Q)$  に関する度数分布

節までに利用した積集合, 和集合の計算は自明ではない. 本手法では, 次のように計算する.

- 質問文  $q$  に含まれる単語の集合を  $w(q)$  とする.
- $Q_i \equiv \bigcup_{q_k \in Q_i} w(q_k)$  とする.
- $Q_i \cap Q_j$  および  $Q_i \cup Q_j$  は, 通常の積集合・和集合とする.
- $q_k \in Q_i \equiv sim(w(q_k), Q_i) > \theta$  とする.
- $q_k \in Q_i$  にランキングが必要な場合には  $sim(w(q_k), Q_i)$  の値により行う.

## 2.4 支援 A: 改善タスクによる質問文の改善

$q_j$  および  $IT_j (|IT_j| = m)$  が与えられたとき,  $Q = Q_1 \cap Q_2 \cap \dots \cap Q_m$  とする. このとき, H3 より次を行う.

$$q_1^1 \dots q_j^n \in Q \text{ を出力}$$

$q_1^1 \dots q_j^n$  の出力時にはランキングを行う. このランキングは  $sim(w(q_k), Q_i)$  の値の降順とする.

## 2.5 改善タスクによる知識ベースの構築

改善タスクの結果は, 同時に今後のための知識ベースへの構造データの登録にも利用する. この知識ベースは, 支援 B において活用される.

本手法における知識ベースとは, 質問文の間に存在する is-a 関係を示したものである. 質問文  $q_j$  が  $q_j'$  に対して is-a 関係にあるとは, 同じデータ集合  $D_j$  に対して質問をしたときに, 任意の  $a_i \in A_j$  に対して,  $q_j$  の結果が  $a_i$  となる  $D_{ji} \subseteq D_j$  と,  $q_j'$  の結果が  $a_i$  となる  $D_{ji}' \subseteq D_j$  の関係が,  $D_{ji} \subseteq D_{ji}'$  となることである. 例えば,  $q_1$ : 「この画像はわいせつですか?」は  $q_2$ : 「この画像は子供にとって不適切ですか?」に対して is-a 関係にある (図 5(a)). 本手法における知識ベースでは, 改善タスクの結果を用いて, このような質問文間の is-a 関係を保持する.

具体的な登録方法は次の通りである.  $q_j$  および  $IT_j (|IT_j| = m)$  が与えられたとき,  $Q = Q_1 \cap Q_2 \cap \dots \cap Q_m$ ,  $Q_i' = Q_i - \{Q_1 \cup \dots \cup Q_{i-1} \cup Q_{i+1} \cup \dots \cup Q_m\}$  とする. このとき, H2 より, 全ての  $Q_i'$  に対して  $is-a(Q_i', Q)$  を知識ベースに登録する.

## 2.6 支援 B: 知識ベースを用いた質問文の改善

知識ベースを利用すれば, 入力としてゴールドデータを用意できない場合でも, 質問文の改善をすることが可能である. 入力として  $q_j$  が与えられたとき, 知識ベースを  $KB$  とする. このとき, 次を行う.

$$\begin{cases} q \in Q \text{ を出力} & (\exists is-a(Q_i', Q) \in KB \wedge q_j \in Q_i') \\ q \in Q_i' \text{ を出力} & (\exists is-a(Q_i', Q) \in KB \wedge q_j \in Q) \end{cases}$$

これにより, 問い合わせ  $q_j$  と is-a 関係で結ばれた質問文が出力される. 例えば,  $q_j =$  この画像はわいせつですか? が与

えられたとき, より特化した質問として, 「この画像は子供にとって不適切ですか?」といった質問文が出力される.

## 2.7 拡張

**拡張 1: 支援 B の拡張.** 支援 B において, 必ずしも与えられた  $q_j$  に常にマッチする質問文が知識ベースに存在するとは限らない. そこで, 各集合  $Q_i$  の代表的なキーワードを抽出する関数  $K(Q)$  を用いて, キーワードの関係を表す構造データを作成する. 例えば, 「不適切な画像はどれですか」という質問文を含む  $Q_1$ , 「わいせつな画像はどれですか」という質問文を含む  $Q_2$ , 「暴力の画像はどれですか」という質問文を含む  $Q_3$  が存在し,  $is-a(Q_2, Q_1), is-a(Q_3, Q_1) \in KB$  とする. このとき,  $K(Q_1) =$  不適切,  $K(Q_2) =$  わいせつ,  $K(Q_3) =$  暴力 とすると, キーワードの関係は図 5(b) となる.

このデータを利用すれば,  $q_j$  に完全マッチする質問が知識ベースに存在せずとも, これらのキーワードにマッチした場合に, そのキーワードの置き換えを候補として提示可能になる.

**拡張 2: 複数のデータを用いた改善タスクへの拡張.** これまで, 改善タスクで提示する情報  $d_i'$  は  $(A_j, d_i, C_j(d_i))$  ただし  $d_i \in D_j$  であった. 本拡張では, 入力 ( $template_j(q_j, A_j)$ , データ集合  $D_j$ , およびゴールドデータ  $C_j$ ) が与えられたとき, 提示する情報を, 次のように拡張することを考える.

**(2a) 同じ選択肢 (正例) のデータをすべて表示.** 改善タスク  $t_k^- \in IT_j$  を  $a_k \in A_j$  毎に生成する. 各改善タスク  $t_k^-$  は  $t_k^- = template_j^-(q_j^-, A_j^-)[d_k^-]$  とする. ここで,  $t_k^-$  の各構成要素は次の通りである.

- $q_j^- =$  「ある人が .. 風景写真ですか?」
- $d_k^- = (A_j, \{d_i | C(d_i) = a_k\}, a_k)$
- $A_j^-$  は存在しうる質問文の集合

**(2b) ある選択肢になる全てのデータに加えてそうでないデータ (負例) を全て表示.** 改善タスク  $t_k^- \in IT_j$  を  $a_k \in A_j$  毎に生成する. 各改善タスク  $t_k^-$  は  $t_k^- = template_j^-(q_j^-, A_j^-)[d_k^-]$  とする. ここで,  $t_k^-$  の各構成要素は次の通りである.

- $q_j^- =$  「ある人が .. 風景写真ですか?」
- $d_k^- = (A_j, \{d_i | d_i \in D_j, C(d_i) = a_k\}, \{d_i | d_i \in D_j, C(d_i) \neq a_k\}, a_k)$
- $A_j^-$  は存在しうる質問文の集合

上記拡張では表示データが一般に多くなるため, 実用上は, サンプリングなどで削減し表示する工夫等が必要である. また, これらの変種 (正例 N 個, 負例 M 個) 等も同様に定義できる.

## 3. 実験

本節では, 2 章で述べた改善タスクに関する実験の結果を報告する. 実験は, Yahoo!クラウドソーシングを用いて行った.

表 1: 元の質問文と改善タスクにより得られた質問文

元の質問文	得られた質問文 (上位)	得られた質問文 (下位)
$P_1$ : <>内の語句は、調理の方法に関連する語句ですか？	これは、料理の調理法ですか？ これは料理ですか？ これは、調理法ですか？	肉や魚がメインのメニューですか？ 居酒屋でよく食べるものは何ですか？ 美味しい和食は？
$P_2$ : <>内の語句は、「飲み物やお酒のジャンル名 (例: 麦茶、ワイン など)」ですか？	これは、飲み物ですか？ これは飲み物の名前ですか？ これは飲み物のメニューですか？	お酒ですか？ カフェにあるメニューですか？ それは、液体ですか？
$P_3$ : <>内の語句は、「料理のジャンル名 (例: 日本料理、イタリアン など)」ですか？	これは料理の種類ですか？ これは、料理のジャンルですか？ これは料理名ですか？	それは中国で飲食できるものですか？ 食事処ではない名称は？ 変わったもので食べたいものは？
$P_4$ : <>内の語句は、飲食店のメニューや料理名に関連する語句ですか？	これは、料理ですか？ これは食べ物ですか？ これは料理の名前ですか？	それは主食ですか？ 炭水化物ですか？ 大人でも子供でも好きなメニューですか？
$P_5$ : <>内の語句は、レストラン、飲食店の名前ですか？	これは飲食店の名前ですか？ これは飲食店の名称ですか？ これは飲食店名ですか？	それは食べ物ですか？ ラーメン屋焼肉屋牛丼屋ですか？ 今日はお肉が食べたいですか？

### 3.1 実験方法

**入力データと改善タスク.** 本実験では、5つの質問文  $P_j (1 \leq j \leq 5)$  (表 1 - 元の質問文)、質問文共通の回答の候補  $A_j = \{\text{“はい”}, \text{“いいえ”}\}$ 、質問文共通のデータ  $D_j = \{d_1, \dots, d_{25}\}$ 、各  $P_j$  のためのゴールドデータ  $C_j$  を用意し、それらを用いて実験を行った。全ての  $P_j (1 \leq j \leq 5)$  に関して  $\{d_i | C_j(d_i) = \text{“はい”}, d_i \in D_j\} = 5$  である。本実験では、2.7 節の拡張 2b (の変種) に従い、“はい”の選択肢のデータ (正例) を同じカテゴリから 3 つ、同様に、“いいえ”の選択肢のデータ (負例) として正例ではないカテゴリを 1 つ選び、サンプルを 3 つ表示した。ここでのカテゴリとは、どの質問文  $P_j$  のゴールドデータ  $C_j$  で“はい”となるかの分類である。また、 $\text{sim}(w(q_k), Q)$  の計算には、Jaccard 係数[5]を使用した。

**タスク数とワーカーへのタスク割当.** 今回の実験では、改善タスクは (正例のサンプル 3 つの組合せ)  $\times$  (負例のサンプル 3 つの組合せ) 毎に生成した。ただし、改善タスクの結果は負例よりも正例に大きく関係すると考えられるため、正例は全ての可能な組合せをとるが、負例のサンプル組合せは 1 つとし、正例ではないカテゴリ 1 つの中から無作為に 3 つ選んだ。したがって、本実験で生成した改善タスク数は元の質問文  $P_j$  毎に  ${}_5C_3 \times 1 = 10$  個である。また、本実験では、出来る限り多くの質問文を得るため、1 つの改善タスクに対して 20 人のワーカーを割り当て、同じワーカーに同じタスクを割り当てることはしなかった。すなわち、1 つの改善タスクから得られる質問文の数は 20 個である。したがって、各  $P_j$  毎に  $10 \times 20 = 200$  個の改善タスクをワーカーに問合わせた。

### 3.2 実験結果と考察

元の質問文各  $P_j$  に対して、実験によって得られた質問文  $q_k$  の、 $\text{sim}(w(q_k), Q)$  に関する度数分布を図 6 に示す。また、各  $P_j$  とそれに対して得られた質問文のうちランキング上位・下位のもの 3 つを表 1 に示す。表 1 から分かるように、Jaccard 係数が高い質問文は、元の質問文の意味と同一の意味であるものが多かった。また、得られた質問文のうち Jaccard 係数が 0 であったものは全体の 3 割程度であったが、その中には元の質問文の意味と同一と考えられるものもいくつか存在した。

実験結果より、元の質問文の意味と同一となる質問文が改善タスクの結果から得られることが確認された。しかし、得られた質問文のうち、Jaccard 係数が高い質問文であっても元の質問文と意味が異なるものや、逆に、Jaccard 係数が低い質問文であっても元の質問文の意味と同一となるものが存在した。したがって、質問文集合の計算方法や、 $\text{sim}(w(q_k), Q)$  にさらなる工夫が必要である。

## 4. まとめと今後の課題

本論文では、マイクロタスクの質問文の改善を目的として、タスクの答えとデータから元の質問文を推測させるという「改善タスク」をクラウドソーシングする手法の提案を行った。また、実験により、改善タスクが質問文の改善に必要な情報をある程度得ることができていることが確認された。

今後の課題としては、知識ベースを利用したタスク改善支援の実験があげられる。また、ゴールドデータがない場合のタスク改善手法の検討も今後の課題である。

謝辞。中川雅史氏、山本学氏をはじめとするヤフー株式会社 Yahoo!クラウドソーシング PJ メンバーのご協力に感謝申し上げます。

## 参考文献

- [1] Amazon Mechanical Turk, <https://www.mturk.com/>.
- [2] Stephen Beale, Benoit Lavoie, Marjorie McShane, Sergei Nirenburg, Tanya Korelsky. Question answering using ontological semantics. TextMean '04, 41-48, 2004.
- [3] Ailbhe Finnerty, Pavel Kucherbaev, Stefano Tranquillini, Gregorio Convertino. Keep it simple: reward and task design in crowdsourcing. SIGCHI '13, Article No.14, 2013.
- [4] Jinyang Gao, Xuan Liu, Beng Chin Ooi, Haixun Wang, Gang Chen. "An online cost sensitive decision-making method in crowdsourcing systems". SIGMOD '13, 217-228, 2013.
- [5] Jaccard index, [http://en.wikipedia.org/wiki/Jaccard\\_index](http://en.wikipedia.org/wiki/Jaccard_index)
- [6] Anand Kulkarni, Matthew Can, Bjorn Hartmann. Collaboratively crowdsourcing workflows with turkomatic. CSCW '12, 1003-1012, 2012.
- [7] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, John Horton. The future of crowd work. CSCW '13, 1301-1318, 2013.
- [8] A. Marcus, E. Wu, D. Karger, S. Madden, R. Miller. "Human-powered sorts and joins". Proceedings of the VLDB Endowment, Volume. 5, No. 1, 13-24, 2011.
- [9] Adish Singla, Andreas Krause. "Truthful Incentives in Crowdsourcing Tasks using Regret Minimization Mechanisms". WWW '13, 1167-1178, 2013.
- [10] 清水 伸幸, 山下 達雄, 塚本 浩司, 颯々野 学. "クラウドソーシングにおける成果物の品質維持のためのダミー問題出題手法の検討". 言語処理学会第 20 回年次大会, 2014.
- [11] Yahoo!クラウドソーシング BETA, <http://crowdsourcing.yahoo.co.jp/>.