人工知能学会研究会資料
SIG-KBS-B403-01

ツイートタイムラインへの階層的クラウドソーシングの適用による住まい探しに関するツイートの抽出

Extraction of Tweets About Housing Property Due to the Application of Deep Crowdsourcing to Tweet Timeline

楡井 泰行[1*]　篠田 孝祐[1]　諏訪 博彦[2]　清田 陽司[3]　栗原 聡[1]
Yasuyuki Nirei[1]　Kousuke Shinoda[1]　Hirohiko Suwa[2]　Yohji Kiyota[3]　Satoshi Kurihara[1]

[1] 電気通信大学　大学院情報システム学研究科
[1] Graduate School of Information Systems, The University of Electro-Communications
[2] 奈良先端科学技術大学院大学情報科学研究科
[2] Graduate School of Information Science, NARA Institute of Science and Technology
[3] 株式会社ネクスト
[3] NEXT CO.,Ltd.

Abstract: In recent years, people who are looking for a housing property by using the real estate portal site is rapidly increasing. On the other hand, it is difficult to read the background and needs of users from the site of the access log. Many users are writing to social media frequently that sometimes feelings and experiences. Therefore, by analyzing the write social media user using real estate portal site, we considered more appropriate recommendation of housing property can be achieved. In this study, the use of crowdsourcing from writing information of social media in order to make an estimate of the housing property search phase. As a result, it was confirmed that discover the user's housing property background and needs.

1　はじめに

現在，不動産情報や就職情報などを探す際には，大量の情報を集約して提供しているポータルサイトがよく利用される．ポータルサイトでは，さまざまな条件を用いて情報を絞り込んで探す検索機能が提供されているものの，生活上のニーズや価値観などを検索条件だけで表現することは不可能である．ポータルサイト上の膨大なアクセスログを用いた情報推薦によってこの問題に対処する試みがなされているものの，アクセスログだけで利用者のニーズを読み取ることは難しい．今後，さらに多くの利用者が活用したいと思うようなサービスを作り上げるためには，利用者のニーズを読み取るために，利用者の深い意図や生活状況といった情報を抽出できる新たなデータが必要である．

本研究では，利用者の深い意図や生活状況などの情報を得るために，ソーシャルメディアに着目する．理由としては，Twitter などのソーシャルメディアにおいて人々は，その時々の気持ちや生活状況などの情報を書き込んでいるためである．その中には，住宅物件に関する書き込みも存在している．ソーシャルメディアを分析することで利用者の深い意図や生活状況などを抽出できると考える．住宅物件に関する利用者の意図や状態を抽出するためには，住宅物件に関する書き込みを抽出する必要がある．住宅物件に関する書き込みを抽出する方法として，手作業で判別していく方法があるが，作業コストが膨大にかかってしまうことが問題となる．

この問題の解決のために，近年ではデータ判別作業を低コストで行うことが可能な，クラウドソーシングというシステムを用いる試みが注目を集めている．クラウドソーシングとはインターネットを通じて不特定多数の人に対して業務を委託することである．例として，Yahoo!クラウドソーシング[1]，Amazon Mechanical Turk[2] などをはじめ，多くのクラウドソーシングサービスが存在する．クラウドソーシングの特徴として，人手で作業を行うことが挙げられる．そのため，計算機で判断が困難なデータに対して評価を行うことができる．

本研究では，ソーシャルメディア上から住宅物件探

*連絡先：電気通信大学情報システム学研究科社会知能情報専攻
182-8585 東京都調布市調布ヶ丘 1-5-1 東ニ号館 4 階
E-mail: ay1351014@si.is.uec.ac.jp

[1] http://crowdsourcing.yahoo.co.jp/
[2] http://aws.amazon.com/jp/mturk/

索に関する可能性が高いデータを抽出し，階層的にクラウドソーシングを適用することで住宅物件探索に関するツイートを抽出する．まず，2 節では関連研究を紹介する．次に，3 節で提案手法の全体像を述べる．4 節で住宅物件探索ツイートの判別について述べ，5 節で 4 節で判別したデータを用いて住宅物件探索フェーズに属するツイートの抽出について述べる．6 節，7 節では 4 節，5 節で述べた方法の結果について述べる．最後に，8 節で考察，9 節で今後の課題について述べる．

2　関連研究

本節では，ソーシャルメディアからの情報収集に関する研究とクラウドソーシングに関する研究について述べる．

まず，ソーシャルメディアからの情報収集に関する研究について述べる．迫村らは，ツイッター情報からテキストの特徴量とグラフの特徴量を抽出することで，ツイッターの話題，その大きさや広がりと，経済動向との関連性を明らかにした [1]．若井らは，Twitter からテレビで放送されている映画について，ツイートの感情を Twitter 特有表現も考慮に入れて時系列に抽出することで，感情の変化を分析した [2]．また，ソーシャルメディアを実世界を観測するためのセンサとして活用を行っている研究も存在する．長野は，Twitter を実世界を観測するためのセンサとして見ることで，Twitter における鉄道運行情報に関係するツイートのみ抽出し，タイムラインを作成するアプリを開発・評価を行った [3]．榊らは，Twitter から人物の目撃情報を SVM，パターンマッチングを組み合わせることで情報抽出を行った結果，検出精度が 8 割程度となった [4]．本研究でも同様に，ソーシャルセンサとして Twitter を用いることで，住宅物件探索を行っているユーザを抽出し，分析を行うことを考えている．

次に，クラウドソーシングに関する研究において，新しいクラウドソーシングサービスの形を提案しているものについて述べる．Senjuti らは「SmartCrowd」という新しいクラウドソーシングサービスの中で，ワーカに長期的にタスクを実行してもらうことで，ワーカのタスク処理の精度を上げる枠組みの提案を行っている [5]．また，西らは，ワーカ同士が知り合いであることを仮定し，タスクを引き受けた能力の低いワーカは知り合いの能力の高いワーカにタスクを委託することで，高い報酬を期待できるように設定する．そのため，能力の高いワーカが多くのタスクを処理することで，従来のクラウドソーシングよりも品質の高い成果物を得られる新しいクラウドソーシングの形として，REC というサービスを提案している [6]．本研究では，新しいクラウドソーシングのサービスの提案ではなく，データの判別作業を行う際にインタラクティブにクラウドソーシングを利用する枠組みを提案する．また，クラウドソーシング行う上で，成果物の品質維持，または向上が求められている．沼田らは，クラウドソーシングのワーカが依頼内容を熟読しているか判定し，熟読しているワーカを選定することで成果物の品質が向上することを示した [7]．清水らは，不真面目に回答する人がいるため，設問との区別がつかないダミー問題の設定が必要であると述べている [8]．本研究では，クラウドソーシングを用いるに際に，チェック設問と多数決による判別を行うことによって成果物の品質を維持する．

このように，ソーシャルメディアやクラウドソーシングに着目した研究は数多く存在する．本研究では，Twitter から抽出したデータを，クラウドソーシングを活用することで住宅物件探索フェーズを推定する．

3　提案手法の全体像

本研究における階層的にクラウドソーシングを適用する提案手法の全体像を図 1 に示す．

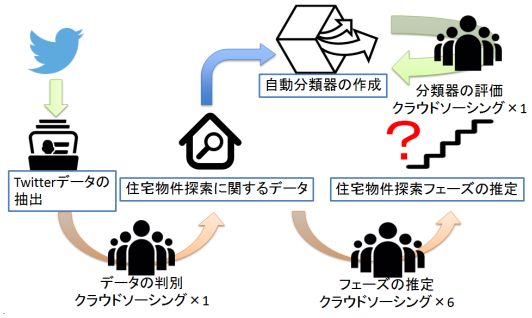

図 1: 提案手法の全体像

まず，Twitter から抽出したデータに対して，クラウドソーシングを適用することで住宅物件探索に関するデータを判別する．次に，住宅物件探索に関するデータを用いることで，自動分類器の作成を行い，評価の段階でクラウドソーシングを用いて行う．また，住宅物件探索に関するデータから住宅物件探索フェーズに当てはまるツイートを推定するために，クラウドソーシングを用いて行う．本研究では，階層的にクラウドソーシングを利用することで，Twitter データを利用した住宅物件探索に関する自動分類器の作成と，住宅物件探索フェーズに当てはまるツイートを抽出する．なお，本稿では，Twitter データに対して階層的にクラウドソーシングを適用することで，住宅物件探索フェーズに当てはまるツイートの抽出する手法について述べる．

4　住宅物件探索ツイートの判別

本節では，住宅物件探索に関するツイートの判別の手順について述べる．また，本研究で利用するクラウド

ソーシングは，データの簡単な分類に特徴を持つ Yahoo! クラウドソーシングとする．

4.1　タスクの設定

タスクの設定は，対象とするアカウントのツイート全体に対して「http」の文字列を含む割合を計算し，一定以上である場合はノイズとして除去する．次に，住宅物件探索に関連が高い「礼金，内見，家賃」の 3 つの単語を利用したキーワードマッチングを行い，ツイートを抽出する．抽出されたツイートの前後 2 つずつ抽出することで，合計 5 つのツイートを 1 件のタスクとして使用し，ワーカに判別してもらう．

4.2　成果物の信頼性の確保

2 節で述べた通り，クラウドソーシングを行う場合には，信頼性を確保することが重要となる．本研究では，2 種類の方法を用いることで信頼性を確保する．

1 つ目の方法は，予め答えが判明しているチェック設問を設定し，ワーカに判別してもらう．そしてワーカのチェック設問に対する正解率を求め，正解率が一定以上のワーカの判別のみを利用する．

2 つ目の方法は，1 つのタスクに対して，3 人のワーカに判別してもらうことで，多数決によるタスクの判別を行う．それによって，タスクの判別に対する尤もらしさが向上する．

この 2 種類の方法を用いることによって，クラウドソーシングの成果物の信頼性を確保する．

4.3　チェック設問の設定

まず,「住まいを探している」,「住まいを探していない」の 2 種類のチェック設問をそれぞれ 10 件ずつ用意する.「住まいを探している」が正解となるチェック設問は，キーワードマッチングで選択されたデータに対して，手作業で「住まいを探している」と判別したデータを使用する.「住まいを探していない」が正解となるチェック設問は，キーワードマッチングで選択されなかったデータに対して，手作業で「住まいを探していない」と判別したデータを使用する．このチェック設問を 1 件と，4.1 節で設定したタスク 4 件を 1 セットとして，ユーザに判別を行ってもらう．

4.4　質問文の設定

住宅物件探索を行っているツイートの判別作業においては，住まい探しの情報を集めている段階や身近な人に相談しているツイートなど，住宅物件探索に関係するツイートは全て集める．そのため，ワーカが判別するための指標となるような具体例を掲載する．また，クラウドソーシングの選択肢は,「住まいを探している」,「住まいを探していない」,「わからない」の 3 種類を用意する.「わからない」を用意した理由は，判別できない場合に,「住まいを探している」と「住まいを探していない」のどちらかを選択させた場合にノイズが入ってしまうためである．

5　住宅物件探索フェーズ推定

本節では，住宅物件宅探索フェーズ推定を行うために階層的にクラウドソーシングを適用する手順を述べる．

5.1　複数回住宅物件探索に関してツイートを行うアカウントの確認

4 節による判別を行うと,「住まいを探している」と判別されたツイートを抽出することができる．本節では,「住まいを探している」と判別されたツイートを行ったアカウントに対して，複数回住宅物件探索についてツイートしているかを確認する．方法としては，まず「住まいを探している」と判別されたツイートを行ったアカウントのタイムラインを用意する．次に，住宅物件探索に関連が高いと「礼金，家賃，内見」の 3 つの単語を用いて全てのツイートに対してキーワードマッチングを行い，ヒットしたツイートを基準点とする．基準点となったツイートから前後 6 カ月（合計 1 年）の時間幅内のツイートデータを抽出し，手作業で住宅物件に関するツイートがあるか判別していく．この時，時間幅内に基準点となるツイートが複数存在する場合には，ツイートされた時間が新しいものを基準点とし，時間が古いものを基準点から除外する．

5.2　住宅物件探索フェーズの定義

本節では，5.1 節から複数回住宅物件探索に関してツイートしていると判別されたアカウントが行ったそれぞれのツイートに対して，住宅物件探索フェーズ毎にタグづけを行う．本論文では住宅物件探索フェーズを図 2 と定義する．

図 2: 住宅物件探索フェーズ

本研究では，住宅物件探索フェーズは主に 4 段階に分ける．例えば，ユーザは勤務地が変更された，今住んでいる場所から遠くの大学に合格した等，住宅物件探索を行う動機が存在するフェーズがある．次に，新しく住む地域における住宅物件の情報を，住宅物件情報誌やネットにおけるポータルサイト等を利用して収集するというフェーズが存在する．収集した情報の中で，良いと感じた物件を見つけた場合には，実際に不動産と接触を行い部屋を内見するというフェーズを得て，

満足した場合には住宅物件を契約するというフェーズへと移行する．本研究では，ツイートが図 2 のどの住宅物件探索フェーズに当てはまるか推定を行う．

5.3　タスクの設定

5.1 節において，複数回住宅物件探索に関してツイートしていることが確認されたアカントのツイートが，それぞれ図 2 のどのフェーズに当てはまるか，またはそのどれにも当てはまらないかをクラウドソーシングを用いることで判別する．そのために，まず複数回住宅物件探索に関してツイートしているアカウントのタイムラインを一定のまとまりに切り分け，判別を行った後に，ツイート単位でどのフェーズに当てはまるかを判別していく．方法としては，まずクラウドソーシングに提出するタスクとして，複数回住宅物件に関してツイートしていることが確認されたアカウントを用意する．次に，5.1 節で述べた基準点から前後 1 カ月（合計 2 カ月）を時間幅とする．そして，10 ツイート単位でタイムラインを切り分けることで今回のタスクに設定する．

5.4　タスク単位における住宅物件探索フェーズ推定の手順

本研究では，住宅物件探索フェーズに当てはまるツイートを推定するために合計で 5 ステップのクラウドソーシングを用いる．まず住宅物件探索フェーズのどれかに当てはまるか判別を行う．この時，選択肢として,「住まいを探したいと考えている」,「住まいを探したいと考えていない」の二つを用意する．このクラウドソーシングをステップ 1 とする．

次に，クラウドソーシングによって「住まいを探したいと考えている」と判別されたタスクに対して,「住宅物件探索を行う動機の存在」に当てはまるか，それとも他のフェーズに当てはまるのかを判別してもらう．この時の選択肢は「住まい情報を集めた」,「住まい情報を集めていない」の二つを用意する．このクラウドソーシングをステップ 2 とする．ステップ 2 で「住まい情報を集めていない」と判別されたタスクに対しては，住まい探し意図は存在するが住まい情報を集めてはいないと判別することができる．そのため,「住まい情報を集めていない」と判別されたタスクは「住宅物件探索を行う動機の存在」に当てはまると判別することができる．

「住まい情報を集めた」と判別されたタスクを用いて，今度は「住宅物件情報の収集」に当てはまるか，それとも「不動産との接触・内見」や,「住宅物件の契約」に当てはまるのか判別していく．この時の選択肢は「不動産会社とコンタクトをとった/物件を見学した」,「どちらもやっていない」の二つを用意する．このクラウドソーシングをステップ 3 とする．ステップ 3 で「どちらもやっていない」と判別された場合,「住宅物件情報の収集」に当てはまると判別することができる．

また,「不動産会社とコンタクトをとった/物件を見学した」と判別されたデータに対して「不動産との接触・内見」と「住宅物件の契約」のどちらに当てはまるか判別を行う．この時の選択肢は「物件の契約を決めた」,「物件の契約は決めていない」の二つとする．このクラウドソーシングをステップ 4 とする．ステップ 4 で「物件の契約を決めた」と判別されたタスクは「住宅物件の契約」に,「物件の契約は決めていない」と判別されたデータに対しては「不動産との接触・内見」に当てはまると判別することができる．

この合計 4 ステップのクラウドソーシングを実行することで，それぞれの住宅物件探索フェーズに当てはまるタスクが判別される．また，チェック設問の設定はそれぞれのステップ毎の選択肢に合わせて，それぞれ 10 件ずつ作成する．

5.5　ツイート単位における住宅物件探索フェーズ推定の手順

5.4 節を実行すると住宅物件探索フェーズに当てはまるツイート群が結果として手に入れることができる．本研究では，ツイートがどの住宅物件探索フェーズに属しているか推定し，ユーザのツイートがこのツイート群をツイートごとに分解することで，どのツイートが住宅物件探索フェーズに関係しているかをクラウドソーシングを用いて判別する．このクラウドソーシングをステップ 5 とする．この時．住宅物件探索フェーズに当てはまるツイートが複数存在した場合には，ツイートされた時間が新しいツイートを選択する様に指示する．理由としては，複数存在する場合にワーカの判別が分散してしまうことが考えられるためである．また，チェック設問はそれぞれの住宅物件探索フェーズについて 3 つずつ作成する．

6　住宅物件探索の判別結果

本節では，4 節における住宅物件探索に関するツイートの判別結果を述べる．

6.1　信頼性の評価

まず，クラウドソーシングに参加したワーカに対して，チェック設問を用いることで信頼性を評価した．この時，タスクの総数は 2400 件，一人のワーカが行うことが可能なセット数として，最大 5 セットを設定した．その結果，全てのワーカ数は 396 人となった．また，正解率が 80% 以上であるワーカは 328 人であった．本研究では，これらのワーカの判別結果をデータとして用いた．

6.2 データの分類結果

6.1 節によって抽出されたワーカが判別したデータに対して，多数決によって「住まいを探している」,「住まいを探していない」,「分からない」と判別されたタスクの総数をまとめたものが表 1 である．今回のクラウドソーシングでは，1 つのタスクにつき，3 人のワーカが判別を行っているが，信頼性を確保するためにチェック設問の正解率に閾値を決定した．そのため，3 人が判別を行っていないタスクも存在する．ここから，タスクに対して 2 人以上のワーカが判別を行っていない場合には，多数決による判別を行うことができないため除外した．

表 1: 多数決によるそれぞれの判別のタスク数

住まいを探している	286
住まいを探していない	1555
わからない	40
タスクの合計	1881

表 1 から「住まいを探している」と判別されたタスクが 286 件となり，全体の約 15% となった．また,「住まいを探していない」と判別されたタスクは 1555 件となり，全体の約 83% となった．

7 住宅物件探索フェーズ毎によるタグ付け

本節では 5 節における住宅物件探索フェーズに当てはまるツイートの推定結果について述べる．

7.1 信頼性の評価

ステップ 1 においては多数決を 2 人で行った後，判定が一致しなかったタスクをもう一度 2 人のワーカで判別を行うためにクラウドソーシングを行った．また 2 回目のクラウドソーシングでも判別が 1 対 1，または 2 対 2 に分かれた場合には自身で判別を行った．そのため，ステップ 1 の 1 回目のクラウドソーシングをステップ 1-1，ステップ 1 の 2 回目のクラウドソーシングをステップ 1-2 とする．またステップ 2 とステップ 3 においては多数決は 3 人で行い，ステップ 4，ステップ 5 に関しては多数決を 5 人で行った．この時，使用するワーカの閾値は全て 80% 以上と設定し，1 人のワーカが行うことが可能なセット数は全て最大 5 セットに設定した．

まず信頼性の評価を行うために，それぞれのステップにおいてタスクを実行したワーカ数と正解率が 80% 以上のワーカ数を表 2 に示す．

本研究では，表 2 における正解率が 80% 以上であるワーカの判別を用いてタスクの判別を行っていく．

表 2: それぞれのステップにおけるワーカの正解率

ステップ	全体のワーカ数	正解率が 80% 以上のワーカ数
1-1	290	237
1-2	40	34
2	42	34
3	34	31
4	38	26
5	54	52

7.2 ツイートの分類結果

7.1 節におけるそれぞれのステップにおいて閾値以上のワーカの判別を利用して多数決を行った結果を表 3 に示す．この時，全てのステップにおいて多数決を行ったワーカが 1 人以下の場合にはタスクを除去した．図 3 から,「住宅物件情報の収集」に属するタスクが 57 件と最も多かった．次に属するタスクが多かったフェーズは「不動産との接触・内見」であり，47 件となった．

表 3: 多数決によるタスクの判別結果

フェーズ	多数決結果
住宅物件探索を行う動機の存在	32
住宅物件情報の収集	51
不動産との接触・内見	47
住宅物件の契約	14
合計	144

7.3 住宅物件探索フェーズに属するツイートの推定結果

7.2 節から，住宅物件探索フェーズに属するタスクを推定できたため，ステップ 5 によってタスク内のどのツイートが住宅物件探索フェーズに属するか推定を行った．表 4 に多数決の内訳を示す．ここで，4 : 1 は多数決によって 5 人中 4 人のワーカが同じツイートを選択し，5 人中 1 人のワーカが別のツイートを選択したこと表わす．また，3 : 1 : 1 は多数決によって 5 人中 3 人のワーカが同じツイートを選択し，5 人中 2 人のワーカがそれぞれ別のツイートを選択したことを表す．表 4 から，全てのワーカの判別が一致したタスクは 64 件となった．

表 4: 多数決による内訳

多数決の内訳	タスク数
5	47
4	17
4:1	30
3:2	20
3:1	6
3:1:1	10
2:1:1	2
2:1:1:1	2
2:1:1	5
2:2	5
合計	144

8 考察

表 4 から，多数決を行った際にワーカの判別が分かれているタスクが半数以上存在した．ワーカの判別が分かれる理由として，次のことが考えられる．まず，例として図 3 を示す．

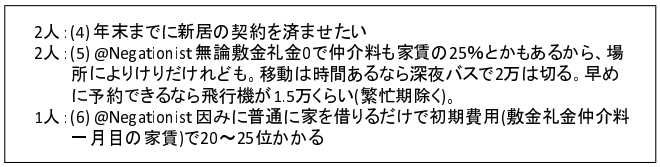

図 3: 例 1 : 意見が分かれたタスクについて

図 3 において左側に記述している人数は多数決におけるワーカの人数となる．図 3 は「住宅物件情報の収集」フェーズに属するツイートの特定を行うためのタスクであるツイート (4) の内容を見ると新居を決定したい旨を書きこんでいる．そのため，「住宅物件探索を行う動機の存在」フェーズに属する内容のツイートではなく，タスクの作業としては選択するべきではないしかし，2 人のワーカがこのツイートを選択している．ここから，設問文を読んでいないワーカが存在するのではないかと考えられる．

次に，図 4 に意見が分かれたタスクを示す．

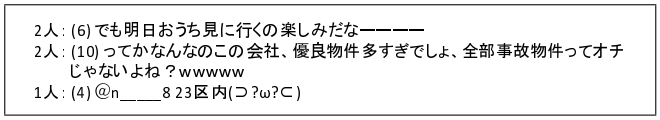

図 4: 例 2 : 意見が分かれたタスクについて

図 4 におけるツイート (6) の内容は，住宅物件探索に関するツイートであり，内見を行うことを示すと推測できる．しかし，友達や親せきの家に行くことも考えることができる．このように内容が曖昧であるため，ツイート (6) を選択したワーカは 2 人であった．また，ツイート (4) における内容は他のツイートとの関係性を見ると，23 区内で住宅物件を探しているのではないかと推測することもできる．そのため，ツイート (4) を選択するワーカが 1 人存在した．このように，ツイートの内容や文脈から読み取れる内容が複数ある場合に，ワーカの意見が分かれることが分かった．

9 おわりに

本研究では，住宅物件探索フェーズに当てはまるツイートを推定するために，階層的にクラウドソーシングを利用する枠組みを提案した．結果として，住宅物件探索フェーズに当てはまるツイートを推定することができた．

今後は，抽出した住宅物件探索に関するツイートを用いて，それぞれの住宅物件探索フェーズに属しているツイートから利用者の背景やニーズを理解することで，より適切な住宅物件の推薦などのサービスの向上を目指す．

参考文献

[1] 迫村光秋，和泉潔，セーヨーサンティ，Twitter のテキストとネットワークの解析による経済動向分析，第 10 回金融情報学研究会，pp.22-27，2013

[2] 若井祐樹, 山本湧輝, 熊本忠彦, 灘本明代，映画の実況ツイートにおける時系列毎の感情抽出手法の提案, 第 12 回日本データベース学会年次大会, 2014

[3] 長野伸一，ソーシャルセンサからの情報抽出技術，東芝レビュー，Vol.69，No.7，pp19-22，2014

[4] 榊剛史，松尾豊，ソーシャルメディアからの人物目撃情報抽出システムの試作，人工知能学会全国大会 2011 論文集，pp1-4，2011

[5] Senjuti Basu Roy, Ionna Lykkourentzou, Saravanan Thirumuruganathan Sihem Amer-Yahia, Gautam Das, Crowds, not Drones: Modeling Human Factors in Interactive Crowdsourcing, DBCrowd 2013: First VLDB Workshop on Databases and Crowdsourcing, pp.39-42, 2013

[6] 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，JSAI オーガナイズドセッション，4pp，2013

[7] 沼田剛明，比嘉邦彦，クラウドワーカー選定方法の提案と効果検証，第 16 回テレワーク学会研究発表大会，pp73-78，2014

[8] 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 回年次大会発表論文集，pp.678-681，2014