Vol.2014-ICS-177 No.2
2014/12/15

クラウドソーシングデータを用いた住宅物件探索アカウントの自動分類

楡井 泰行[1,a)]　篠田 孝祐[1]　諏訪 博彦[2]　清田 陽司[3]　栗原 聡[1,b)]

概要：近年，不動産ポータルサイトを利用して住宅物件を探すケースが急増している．一方，利用者の深い意図や，生活状況といった情報をサイトへのアクセスログから読み取ることは難しい．これに対して，我々はソーシャルメディアに対しては，その時々の気持ちや体験を当たり前のように書き込んでいる．これにより，不動産ポータルサイトアクセスユーザのソーシャルメディアでの書き込み情報を分析することで，より適切な住宅物件の推薦といったサービスが可能になると考えられる．そこで，本研究では，ソーシャルメディアの書き込み情報から住宅物件を探しているツイートを抽出する方法として，クラウドソーシングを利用する枠組みを提案する．

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．ここから，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．しかし，住宅物件以外に関する書き込みも存在しているため，住宅物件に関する書き込みだけ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欠点となる．

そのため，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正確に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抽出した住宅物件探索に関する可能性が高いデータを抽出し，クラウドソーシングに適用することで判別を行うことを試みる．また，クラウドソーシングによって判別されたデータを利用することで，自動で住宅物件探索にかんするツイートを抽出する分類器の作成・評価を試みる．まず，2 章では関連研究を紹介する．次に，3 章でクラウドソーシングを行うための前準備，4 章でクラウドソーシングを利用して判別したデータを用いた自動分類器の作成・評価について説明し，5 章，6 章でその結果を示す．最後に，7 章で考察，8 章で今後の課題について記述する．

[1] 電気通信大学大学院情報システム学研究科
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
NARA Institute of Science and Technology
[3] 株式会社ネクスト
Next Co.,Ltd.
[a)] nirei@ni.is.uec.ac.jp
[b)] kuri@ni.is.uec.ac.jp

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

2. 関連研究

近年では，ソーシャルメディアから情報を抽出，分析を行う研究は数多く存在する．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 [3]．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 [4]．本研究でも同様に，ソーシャルメディアから住宅物件を探しているアカウントを抽出し，分析を行うことを考えている．

また，クラウドソーシングに対する研究も数多く存在している．財前らは，ある命題を入力することで，命題の根拠となる情報源を提示する際に，情報源を計算機を用いて自動的に発見することは困難であるため，クラウドソーシングを用いて根拠となる情報源を検索することができるシステムを提案した [5]．山本らは，医療分野で利用されている用例をマイクロブログを用いて正しさを評価し，クラウドソーシングを用いて，用例を多言語に対応したより正確な用例対訳を作成していくシステムを提案した [6]．ただし，クラウドソーシングを行う際には，不真面目なワーカも存在するため，成果物に対する信頼性の確保が重要となる．これについて清水らが検討を行っている．内容としては，クラウドソーシングからフィードバックを得ることができなければダミー問題が適切かどうか判断することができない．しかし，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は必要であると述べている [7]. 本研究ではクラウドソーシングを行う際に，チェック設問を設定することで不真面目なワーカの回答を除去する方法を採用する．本研究では，クラウドソーシングを用いることによって，ソーシャルメディアの書き込み情報から効率的に住宅物件を探しているアカウントを抽出することを目指す．

2 値分類を行う自動分類器を作成する際には，SVM が数多く研究で利用されている．榊らは，Twitter からユーザの投稿内容，自己紹介文，被リスト名を抽出し，SVM の素性とすることで，高い精度で会社員の 2 値分類を行うことを可能としている [8].

このように学習データを集めて，自動分類器を作成する研究は数多く存在する．本研究では，クラウドソーシングによって抽出した住宅物件を探しているアカウントの情報を用いることによって，自動分類器の作成・評価を行うことを目指す．

3. クラウドソーシングの前準備

本研究では，住宅物件を探しているツイートを効率よく抽出するための方法として，クラウドソーシングを利用することを考えている．今回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を利用する．

3.1 タスクの設定

タスクの設定は，まず業者などのノイズとなるアカウントを全体のツイートにおいて，「http」の文字列がどの程度含まれているかを見ることによって業者の除去を行う．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．

3.2 成果物の信頼性の確保

2 章で述べた通り，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．

3.3 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，3.1 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．

3.4 クラウドソーシングの選択肢

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．

4. SVM を用いた自動分類器の作成・評価

本研究では，住宅物件を探しているツイートを抽出するために自動分類器を作成することを目指す．そのため 3 章においてクラウドソーシングを適用したタスクを学習データとして用いる．

4.1 名詞の接続

自動分類器を作成するために，クラウドソーシングを適用することで判別を行ったタスクに対して最初に形態素解析を行う．本研究では，形態素解析ソフトの一つである MeCab を用いる [1]．ただし，MeCab を用いて形態素解析を行った場合，数字の後に「円」や「分」等の助数詞が来た場合にそれぞれの形態素に分けられてしまうことがある．本研究では，これを防ぐために次のルールを用いて形態素を接続する．

- 接頭詞の後に名詞が来た場合　例：「副」＋「代表」
- 数字の次に名詞が来た場合　例：「50000」＋「円」
- 接尾辞の前に名詞が来た場合　例：「安全」＋「性」

4.2 SVM の素性選択

本研究では，自動分類器を作成するために，機械学習ソフトである WEKA を利用することで，SVM の一種である SMO を利用する [2]．また，SVM の素性として次の様に選択を行う．まず，で形態素解析を行ったデータに対して，頻度分析を行う．頻度分析を行った結果，200 回以上頻出した名詞の単語を SVM の素性として利用する．

4.3 パラメータ調整とカーネルの選択

4.2 を元に作成した自動分類器に対して，精度を向上させるためにパラメータ調整とカーネルの選択を行う．まず，それぞれのカーネルに対して精度が高くなるようにパラメータを調整する．そして，精度が高くなったカーネル同士で精度，再現率，F 値を比較することで，自動分類器に用いるカーネルを選択する．

4.4 自動分類器の評価

4.2 節で述べた通り，素性を選択することで自動分類器を作成した後，自動分類器の評価を行う．本研究では，交差検証を用いて評価を行う．交差検証とは，まず利用する学習データをいくつかに分割し，一部のみで学習を行う．残った学習データに対して分類を行うことで，学習の妥当性の検証を行う方法である．この方法のメリットは，元の学習データを評価専用とする必要がないこと，学習データと評価データに偏りがある場合でも，平均を取ることで影響を小さくすることが挙げられる．また，作成した自動分類器を用いて「住まいを探している」と判別されたデータに対して，クラウドソーシングを適用し，人手で判別を行う．この時，3 章と同様の設定でクラウドソーシングに適用する．クラウドソーシングを適用した判別結果と，交差検証から導き出された精度を比較することで，自動分類器の妥当性を評価する．

5. クラウドソーシングによる結果

5.1 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．図 1 にワーカの正解率を示す．この時，タスクの総数は 2400 件，全てのワーカ数は 396 人とした．また，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．

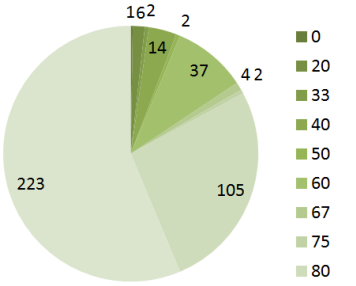

図 1 チェック設問の正解率とワーカの人数

図 1 から，正解率が 100% のワーカが 223 人と最も多く，次に正解率が 80% であるワーカが 105 人と多いことが分かった．本研究では．正解率を 80% 以上であるワーカの判別を利用することを考えている．そのため，合計で 328 人のワーカの判別を使用した．

5.2 データの分類結果

5.1 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，多数決によって判別を行った結果が表 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 から多数決によってワーカ全員が「住まいを探している」と判別したタスク数は 169 件，「住まいを探していない」と判別したタスク数は 1188 件，「わからない」と判別したタスクは 5 件となった．また，多数決によって判別することができなかったタスクが全部で 286 件存在した．

次に，表 1 から多数決によって「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 2 である．

表 2 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

表 2　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

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．

6. 自動分類器の評価実験

6.1 交差検定

4 章の方法で作成した自動分類器を評価した結果が表 3，表 4，表 5，表 6 となる．この時，交差検証のフォールドは 10 と設定した．また，選択したカーネルはそれぞれ normalized poly カーネル，poly カーネル，puk カーネル，RBF カーネルの 4 つを選択した．

表 3　住まいを探していると判別されたデータの精度，再現率，F 値

	normalized poly	poly	puk	RBF
精度	0.707	0.726	0.807	0.764
再現率	0.329	0.287	0.248	0.339
F 値	0.449	0.411	0.38	0.47

表 4　住まいを探していないと判別されたデータの精度，再現率，F 値

	normalized poly	poly	puk	RBF
精度	0.888	0.882	0.877	0.89
再現率	0.975	0.98	0.989	0.981
F 値	0.929	0.928	0.874	0.933

表 5　平均の精度，再現率，F 値

	normalized poly	poly	puk	RBF
精度	0.859	0.858	0.866	0.87
再現率	0.874	0.872	0.874	0.881
F 値	0.854	0.848	0.844	0.861

表 6　全体の精度

	normalized poly	poly	puk	RBF
全体の精度 (%)	87.45	87.23	87.39	88.10

表 3 から「住まいを探している」と判別されたデータに対して，精度が大きいのは puk カーネルであり，次に大きいのは RBF カーネルとなった．再現率，F 値に関しては，最も大きいものは RBF カーネルとなった．また，「住まいを探していない」と判別されたデータにおいて，再現率に関しては puk カーネルが最も大きくなり，次に RBF カーネルとなった．それ以外の精度，F 値に関しては，RBF カーネルが最も大きくなった．「住まいを探している」，「住まいを探していない」と判別された 2 つのデータの平均の精度，再現率，F 値は RBF カーネルが最も大きくなった．全体の精度に関しても，RBF カーネルが最も大きくなり，次に normalized poly カーネルとなった．ここから，RBF カーネルが最も優秀であると考えることができる．今後，本研究では自動分類器のカーネルに RBF カーネルを利用する．

6.2 クラウドソーシングによる評価

6.1 節で評価した自動分類器に対して，4.4 の方法でクラウドソーシングを適用することで評価した．まず．チェック設問に対しての正解率を図 2 に示す．この時，タスクの総数は 214 件，全てのワーカ数は 86 人となった．また，一人のワーカが行うことが可能なセット数として，最大 2 セットを設定した．

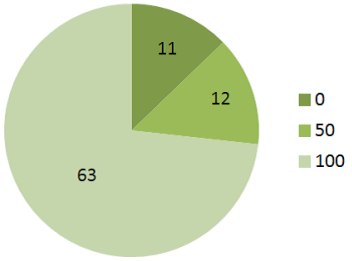

図 2　評価におけるチェック設問の正解率とワーカの人数

表 2 から，正解率が 100% であるワーカ数が 63 人となった．今回は正解率が 100% のみのワーカの判別を使用した．次に正解率が 100% のワーカが行った多数決によるタスクの判別結果が表 7 となる．この時，タスクを判別するワー

表 1　多数決によるタスクの判別結果

住まいを探している	住まいを探していない	わからない	多数決結果
3	0	0	93
0	3	0	705
0	0	3	0
2	0	0	76
0	2	0	483
0	0	2	5
2	1	0	91
2	0	1	26
1	2	0	210
0	2	1	157 2
1	0	2	11
0	1	2	24
1	1	1	70
1	1	0	100
1	0	1	26
0	1	1	90

カが 1 人以下の場合には除去した．表 7 から多数決によってワーカ全員が「住まいを探している」と判別したタスク数は 88 件，「住まいを探していない」と判別したタスク数は 25 件，「わからない」と判別したタスクは 2 件となった．また，多数決によって判別することができなかったタスクが全部で 33 件存在した．表 7 から多数決によって「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 8 である．

表 8 評価における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	110
住まいを探していない	37
わからない	2
タスクの合計	149

表 2 から「住まいを探している」と判別されたタスクが 110 件となり，全体の約 73.8% 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 37 件となり，全体の約 24.8% となった．

7. 考察

表 2 から，「住まいを探している」と判別されたタスクが，「住まいを探していない」と判別されたデータよりも少ないことが分かる．これは，データを抽出する際に，キーワードマッチングを利用していることが考えられる．例を挙げると，単語「内見」を使用してキーワードマッチングを行う場合，「社内見学」などの言葉もヒットしてしまう．そのため，住宅物件に関係ないツイートを抽出することで，「住まいを探していない」と判別されたデータが多いことが考えられる．この場合の様な関係のない単語は，形態素をベースとしてマッチングを行うことで，除去することが可能である．

また，本研究では，業者のアカウントを全体のツイートにおいて，「http」の文字列がどの程度含まれているかを見ることによって業者の除去を行っている．そのため，業者のアカウントであっても，「http」の文字列を含む割合が少ない場合には除去することができない．これが，「住まいを探していない」と判別されたタスクが増えた原因の一つであると考えられる．

表 3 から，自動分類器によって「住まいを探している」と判別されたデータの精度は 0.764 である．また，6.2 節から，クラウドソーシングから「住まいを探している」と判別されたデータは全体の 73.8% となる．ここから，本研究で作成した自動分類器の精度は，交差検証とクラウドソーシングの観点から見て，ほとんど同じ結果を得ることができることが分かる．

8. おわりに

本研究の現状として，ソーシャルメディアの書き込み情報から住宅物件を探しているアカウントを抽出する方法として，クラウドソーシングを利用する枠組みを提案した．内容としては，クラウドソーシングで判別したデータを利用することで作成した自動分類器に対して，クラウドソーシング，交差検証の両方から評価を行った．その結果，「住まいを探している」データに対して，自動分類器の精度が 7 割以上あることが分かった．

今後は住宅物件を探しているユーザに対して，ソーシャルメディアにおける書き込みを分析することで，深い意図や生活状況などを理解し，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] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[2] Weka 3: Data Mining Software in Java，http://www.cs.waikato.ac.nz/ml/weka/
[3] 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
[4] 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
[5] 財前涼，森嶋厚行，クラウドソーシングを用いた情報信

表 7 多数決による評価用タスクの判別結果

住まいを探している	住まいを探していない	わからない	多数決結果
3	0	0	44
0	3	0	5
0	0	3	0
2	0	0	44
0	2	0	20
0	0	2	2
2	1	0	13
2	0	1	9
1	2	0	9
0	2	1	3
1	0	2	0
0	1	2	0
1	1	1	2
1	1	0	23
1	0	1	4
0	1	1	4

憑性判断支援のための情報源検索，情報処理学会，第 74 回全国大会講演論文集（第 1 分冊），3N-1，pp.603-604，2012
[6] 山本里美，福島拓，吉野孝，マイクロブログとクラウドソーシングを用いた用例評価手法および多言語用例対訳作成手法の提案，情報処理学会，ワークショップ 2013 論文集，pp.1-8，2013
[7] 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
[8] 榊剛史，松尾豊，ソーシャルメディアユーザの職業推定手法の提案，日本知能情報ファジィ学会誌，Vol.26，No.4，pp.773-780，2014