Vol.2014-ICS-176 No.6
2014/7/22

ソーシャルメディアからの賃貸物件探索アカウント抽出

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

概要：現在，多くの人々が不動産ポータルサイトを利用して賃貸物件を探している．しかし，利用者の深い意図や具体的な行動は，不動産ポータルサイトのログだけからは読み取ることは難しい．一方，ソーシャルメディアにおいて人々は，自分の気持ちや体験を書き込んでいる．本研究では，利用者の感情を含めた賃貸物件探索パターンを抽出する前準備として，ソーシャルメディアのツイートデータから賃貸物件を探索しているアカウントを抽出することを試みる．また，抽出したアカウントに対してクラウドソーシングを利用することで，学習用コーパスを作成する．

1. はじめに

現在，賃貸物件の探索において，多くの人々が不動産ポータルを利用している．不動産ポータルサイトでは，利用者がどのような物件にアクセスしているかのログデータが蓄積されている．今までは，不動産ポータルはログデータを活用することで利用者のニーズを理解し，サービスの向上へと努めてきた．ここからさらに，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，利用者の潜在的な要望を知る必要がある．しかし，潜在的な要望を知ることは不動産ポータルサイトのログだけからは読み取ることは難しい．

一方，今日では，企業がアクティブサポートを行い，消費者の潜在的な期待に応えることで，消費者の購買意欲，満足度を向上させる動きがある．アクティブサポートとは，企業が消費者の感情を読み取ることで，タイミング良く応対することである．例をあげると，都内に住む A さんが Twitter 上でニッセンの商品が欲しいとつぶやいたことに対して，ニッセンから反応があった．この反応があったために A さんはニッセンの商品を購入する動機となった [1]．また，ソフトバンクモバイルサポートに使用されている Twitter アカウントのカスタマーサービス担当@SBCare[*1] がある．これは，カスタマーサポートの窓口の一つでもあり，Twitter 上でソフトバンクに関する不満や困ったことをつぶやくと，つぶやいたアカウントに対して接触を行い，解決案を提示する．

アクティブサポートにおいて消費者の感情とは，消費者のパターンや接触するタイミング，また消費者の接し方を知る上で重要な要素となる．そのために，企業は消費者が感情を表現している，Twitter をはじめとしたソーシャルメディアを利用している．

ここで，Twitter では，ニッセンの商品に対して自分の気持ちを書き込んでいる様に，賃貸物件に関するつぶやきをしているアカウントも存在している．しかし，賃貸物件以外に関してつぶやいているアカウントも存在している．そのため，賃貸物件に関してつぶやいているアカウントだけを抽出する必要がある．

そこで，本研究では，利用者の感情を含めた賃貸物件探索パターンを抽出することを試みる．そのためには，感情を含めた賃貸物件探索を行っているデータを収集するために，ソーシャルメディア上から賃貸物件探索に関するアカウントを抽出する分類器を作成する必要がある．また，分類器を作成する際に，学習コーパスを作成するために，クラウドソーシングを利用する．本稿では，クラウドソーシングに利用するデータを準備するための抽出法を示す．2 章では関連研究を紹介し，3 章では賃貸物件に関してつぶやいているアカウントの抽出法を説明し，4 章で抽出結果を示す．5 章で抽出法に関して考察を行い，6 章で今後の課題について記述する．

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

[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] 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] https://twitter.com/SBCare

連性を明らかにした [2]．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 [3]．荒井らは，Blog や Twitter に書かれた日常的な疑問を抽出するために，テキスト自動分類を用いることで Web 上から疑問記事を抽出し，疑問に対して回答を呼び掛ける Web サイトを構築した．この研究では，分類器を作成する際に，キーワードマッチングを行い，疑問記事を抽出するために手作業で分別している [4]．本研究では，ソーシャルメディアからキーワードマッチングを用いてノイズを除去することで必要な情報を抽出し，学習用コーパスを作成するためのデータセットとする．また，パターン抽出を行う際に，感情の分析を行うことを考えている．

またデータを抽出する際に，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 [*2]，Lancers[*3] 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，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ため，研究でも利用されている．

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]．本研究では，システムの一部としてではなく，学習用コーパスを作成するためにクラウドソーシングを利用する．

テキストマイニングを行う際に，手作業で分類するには膨大な時間とコストがかかるため，様々な分類器を用いた研究が行われている．廣田らは，方言にも頑健な言語処理システムを構築するための方言コーパスを収集するために，SVM を用いて方言コーパスの収集システムを構築した [7]．黒澤らは，ユーザの嗜好に基づいたページの自動作成のために，ファッションアイテムに関する紹介分の自動分類をナイーブベイズ分類器を用いて行った [8]．本研究で，将来的に賃貸物件探索に関するアカウントを抽出する際には，手作業ではなくこれらの研究と同様に，分類器を用いる．

[*2] http://crowdsourcing.yahoo.co.jp/
[*3] http://www.lancers.jp/

3. アカウントの抽出方法

3.1 抽出方法の全体像

本研究では，ソーシャルメディアである Twitter から賃貸物件に関してつぶやいているアカウントの抽出法を示す．図 1 に抽出法の流れを示し，次にそれぞれの工程の簡単な説明を行う．

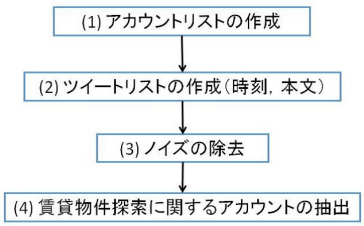

図 1 抽出の流れ

(1) 賃貸物件探索に関してつぶやいている可能性が高いと考えられる Twitter のアカウントを抽出する．
(2) (1) で抽出したアカウントのつぶやきを取得する．
(3) (1) で抽出したアカウントに対して，賃貸物件に関してつぶやいていない可能性が高いアカウントの除去を行う．
(4) (3) で作成したアカウントリストに対してキーワードマッチングを行うことで，賃貸物件に関してつぶやいている可能性が高いアカウントを抽出する．

3.2 アカウントリストの作成

対象とする Twitter アカウントは，不動産ポータルの公式 Twitter アカウントであるホームズくんのフォロワーとする．理由として，不動産ポータルの公式 Twitter アカウントをフォローしているアカウントは，現在もしくは過去に賃貸物件探索を行っていると考えたためである．そのため，アカウントのユーザ ID を抽出するために Twitter API を用いることにする．Twitter API とは，Twitter のある機能を簡略化することでプログラムを書きやすくするためのものである．

3.3 ツイートリストの作成

3.2 節で抽出したアカウントが賃貸物件探索に関するアカウントかどうかを判別するためには，アカウントのつぶやきから判断する必要がある．そのため，ユーザ ID を持つアカウントの最新のつぶやきから最大 3,200 件分まで過去にさかのぼり，つぶやきを取得する．最大 3,200 件である理由は現状, TwitterAPI の最大取得数が最新のつぶやきから 3,200 件までしか取得できないためである．本研究で用いるデータセットは，上記の手順で取得した 2915 人分の 2,981,983 ツイートを対象とする．

3.4 ノイズの除去

3.2 節で抽出したアカウントには業者のアカウントが多分に存在している．

表 1 アカウントの数

合計アカウント数	100
業者でないアカウント数	74
業者でないアカウントの割合	74%

表 1 は実際に手作業で判断した結果である．また，合計アカウント数はデータセットから任意で抽出して，手作業で見た総数である 100 人分である．ここから，業者のアカウントが 30% 近く存在していることが分かる．業者のアカウントは賃貸物件探索についてつぶやいていないため，抽出候補から除外する必要がある．また，業者のアカウントのつぶやきには，「http」という文字列が数多く含まれている．そのため，本研究では，業者のアカウントは一般アカウントよりもつぶやき中に含まれる「http」という文字列が多いと仮定する．ここから，一つのアカウントのつぶやきの中で，「http」という文字列の割合が閾値を超えた場合に，業者のアカウントであると判別する．

割合の算出は次の様に行った．

$$\text{閾値} = \frac{\text{「http」の文字列を含むつぶやき数}}{\text{アカウントの全てのつぶやき数}}$$

3.5 賃貸物件探索に関するアカウントの抽出

本研究では，賃貸物件に関するアカウントを抽出するためにキーワードマッチングを行う．そのために，利用するキーワードを HOME'S[*4] のサイトに存在する不動産用語集 [*5], また，HOME'S における賃貸物件探索に使用する条件を参考に，賃貸物件探索に関すると判断した単語を 10 個選択した．表 2 に選択した単語を示す．

選択した単語を用いてキーワードマッチングによる抽出を行ったつぶやきについて，賃貸物件探索に関するものかどうかを手作業で判別する．このとき，業者のアカウントは手動で判別し除去している．

表 2 選択した単語

礼金	内見	家賃	管理費	ユニットバス
諸費用	物件検索	仲介	間取り	バルコニー

4. 抽出結果

4.1 アカウントリストおよびツイートリストの作成

Twitter API を利用することによって，ホームズくんのフォロワー 38,091 人分のユーザ ID を抽出した．作成したアカウントリストのつぶやきを最大 3,200 件取得することでツイートリストを作成する．

4.2 ノイズの除去

表 1 より，業者でないと判別されたアカウント数が 74% に近いものが最も業者を除去していると考えることができる．本研究では，閾値を，20% ，25% ，30% ，40% ，50% ，60% についてそれぞれアカウントの判別を行った．

表 3 は，対象とデータセットに対して，それぞれの閾値を超えたアカウントを業者であると判別し，下回る場合は業者でないと判別したものである．それぞれの閾値に関して，閾値が 25% の場合，業者でないと判別したアカウントと業者と判別したアカウントの割合が約半分になった．また，閾値を上げることで，業者でないと判別したアカウント数が増えていくことが分かる．表 3 から，業者でないと判別したアカウント数が最も 74% に近い閾値の値は 60% である．

実際に手作業で判別した表 1 のアカウントが，「http」の文字列を含む割合の閾値によって，正しく振り分けられている割合を一致率とする．それぞれの「http」の文字列が含まれている割合の閾値における一致率を表 4 に示す．

[*4] http://www.homes.co.jp/chintai/
[*5] http://www.homes.co.jp/words/k2/

表 3 「http」をツイート中に含む割合によるユーザの割合の変化

閾値	20%	25%	30%	40%	50%	60%
業者でないと判別したアカウントの割合	43.57%	48.95%	54.13%	60.82%	67.14%	73.17%

表 5 閾値ごとにヒットしたつぶやき数

	20%	25%	30%	40%	50%	60%
家賃	113	140	183	221	237	246
礼金	15	15	33	35	37	38
管理費	1	2	18	20	25	25
合計	129	157	234	276	299	309

表 6 閾値ごとの業者以外のつぶやき数

	手作業で業者の判別を行ったつぶやき数	20%	25%	30%	40%	50%	60%
家賃	190	111	137	170	182	190	190
礼金	22	14	14	14	15	22	22
管理費	5	1	2	2	2	5	5
合計	217	126	153	186	205	217	217

表 4　閾値ごとの一致率

閾値	20%	25%	30%	40%	50%	60%
一致率	62%	65%	70%	72%	82%	84%

表 4 から，閾値が大きくなると一致率が高くなる．そのため，アカウントから見てみると，閾値が 60% の場合に，業者のアカウントを除去することができていると考えられる．

次に，表 3 において，業者でないと判別したアカウントに対して「家賃，礼金，管理費」という単語を用いてキーワードマッチングを行う．この結果と，ノイズを除去していない段階で，手作業で業者でないと判別したアカウントに対して「家賃，礼金，管理費」に対して検索を行った結果を表 5 に示す．またそれぞれの閾値に関して，ヒットしたつぶやきの中に業者でないアカウントのつぶやきが存在するか調べるため，手作業で業者を除去したつぶやき数を示したものが表 6 となる．表 6 において手作業で業者の判別を行ったつぶやき数とあるが，この抽出手順は，まずそれぞれの単語に対してキーワードマッチングを行う．次に，一致したつぶやきに対して手作業で業者の判別を行うことで，抽出したものである．

表 5，表 6 を比較すると，閾値が 20% の場合，それぞれの単語において業者のアカウントのつぶやき数の合計が 3 件となり，最も少ないことが分かる．しかし，手作業で業者の判別を行ったつぶやき数と比較すると，91 件の業者でないアカウントのつぶやきを取得できていないことが分かる．また，閾値が，60% の場合，それぞれの単語における業者のアカウントのつぶやき数の合計が最も多い 92 件となるが，業者でないアカウントのつぶやきは全て取得できていることが分かる．ここから，閾値が小さいほど業者のアカウントのつぶやきを除去することが可能であり，閾値が大きいほど取得できる業者でないアカウントのつぶやきが多くなる．

表 5，表 6 からそれぞれの精度，再現率，F 値を算出したものが表 7，表 8，表 9 となる．

表 7　閾値ごとの精度

	20%	25%	30%	40%	50%	60%
家賃	0.98	0.98	0.93	0.82	0.80	0.77
礼金	0.93	0.93	0.42	0.43	0.59	0.57
管理費	1.00	1.00	0.11	0.10	0.20	0.20

表 8　閾値ごとの再現率

	20%	25%	30%	40%	50%	60%
家賃	0.58	0.72	0.89	0.96	1.00	1.00
礼金	0.64	0.64	0.64	0.68	1.00	1.00
管理費	0.20	0.40	0.40	0.40	1.00	1.00

表 9　閾値ごとの F 値

	20%	25%	30%	40%	50%	60%
家賃	0.73	0.83	0.91	0.88	0.89	0.87
礼金	0.76	0.76	0.51	0.53	0.74	0.72
管理費	0.33	0.57	0.17	0.16	0.33	0.33

表 8 から，閾値が小さいほど再現率が低く，閾値が大きいほど再現率が高いことが分かる．また表 7 については，閾値が小さいほど精度が大きくなっている．しかし，単語「礼金」について比較すると，最も精度が小さいのが閾値が 30% の場合であり，次に精度が小さいのが閾値が 40% の場合である．単語「管理費」に関しては閾値が大きいほど精度も大きくなる．表 9 から，単語「家賃」に関して見てみる．このとき，最も F 値が大きいものは順番に閾値が 30% ，50% ，40% となる．次に単語「礼金」の場合を見ると，F 値が最も大きいものが閾値が 20% ，25% の場合であり，続いて 50% となる．単語「管理費」の場合は，閾値が 25% のときに最も大きくなり，20% ，50% ，60% が続いて F 値が大きい場合となる．

4.3　賃貸物件探索に関するアカウントの抽出

3.5 節で説明した方法を行った結果が表 10 となる．この時，賃貸物件に関するアカウントの割合は，

$$\frac{\text{賃貸物件に関するアカウント数}}{\text{単語を含むつぶやきを行ったアカウント数}}$$

にて計算する．キーワードマッチングを行った結果，それ

表 10　キーワードマッチングの結果

	単語を含むつぶやき数	賃貸物件探索に関するつぶやき数	単語を含むつぶやきを行ったアカウント数	賃貸物件探索に関するアカウント数	賃貸物件探索に関するアカウントの割合
礼金	22	15	19	14	74%
内見	21	13	15	10	67%
家賃	190	83	119	62	52%
管理費	5	2	4	2	50%
ユニットバス	12	5	12	5	42%
諸費用	4	1	3	1	33%
物件検索	3	1	3	1	33%
仲介	37	5	26	5	19%
間取り	33	6	26	5	19%
バルコニー	13	1	9	1	11%

ぞれの単語を含むつぶやき数，アカウント数に関して最も賃貸物件探索に関するものを抽出できた単語は「家賃」となった．次に，賃貸物件に関するアカウントの割合が最も高かった単語は「礼金」となった．また賃貸物件探索に関するアカウントの割合が 50% より高くなった単語は，「礼金，内見，家賃」の 3 つとなった．この結果から，「礼金，内見，家賃」の 3 つの単語を用いて，キーワードマッチングによるアカウント抽出を行うと，50% より高い割合で賃貸物件探索に関するアカウントを取得することができると考えられる．そのため，本研究では，「礼金，内見，家賃」の 3 つの単語を用いることで，賃貸物件探索に関するアカウントを抽出する．

5. 考察

ノイズの除去について，アカウントを対象としてそれぞれの閾値を見た場合には，閾値が大きいと良い結果が得られるだろうと判断することができる．しかし，つぶやきを対象としてそれぞれの閾値を見た場合には，必ずしも閾値が大きいほうが良くなるとは言えない結果となっている．対象をアカウント，つぶやきで変化させた場合に結果が変化する理由は，業者のつぶやきの内容にあると考える．業者には賃貸物件を紹介するアカウントが存在し，つぶやきの中で今回キーワードマッチングで利用した単語を多く使用している．そのため，賃貸物件を紹介する業者のアカウントが除去できなかった場合に，業者のアカウントは少ないが，つぶやきの中には業者のつぶやきが多数存在してしまうことが考えられる．

表 10 から，手作業で業者のアカウントを全て取り除いたとしても，賃貸物件探索に関するもの以外のアカウントも存在している．これは，業者でないアカウントの中に賃貸物件を探索していないが，今回キーワードマッチングで利用した単語を用いてつぶやいている人が存在するためと考えられる．また「船内見学」のような単語もヒットしてしまうため，まったく意図していないアカウントを抽出してしまうこともある．ここから，業者でないアカウントの中から，賃貸物件探索に関するアカウントのみを抽出するには，キーワードマッチングのみではなく，キーワードが入っているつぶやき，もしくはその前後のつぶやきの文脈が重要となってくる．そのためには，実際に一つ一つのつぶやきに対して，手作業で文脈を読み取る必要がある．

本研究では，今後クラウドソーシングを利用することで，賃貸物件探索に関するアカウントかどうかを判別することを考えている．

6. おわりに

本稿では，Twitter 上で賃貸物件探索を行っているアカウントの抽出法を示した．今後の課題として，本研究で抽出した賃貸物件探索に関するアカウントと判別したものに対して，クラウドソーシングを利用することによって，学習コーパスを作成する．作成した学習コーパスを分類器に学習させることで，今後は自動で賃貸物件探索に関するアカウントを抽出することができるようにする．また，自動で賃貸物件探索を行っていると判断したアカウントに対して，テキストマイニングを行うことによって，賃貸物件を探索する動機から，完了するまでのつぶやきを抽出する．抽出されたつぶやきに対して，アカウントの感情を抽出する．これに，不動産ポータルサイトのログデータを加えてパターンマイニングを行うことで，利用者の感情を含めた賃貸物件探索パターンの抽出を行う．

参考文献

[1] 西雄大，「アクティブサポート」で販促 ソーシャルで薦め時を探る，NIKKEI COMPUTER 2012.8.16，pp.68-73，2012
[2] 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
[3] 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
[4] 荒井俊介，辻慶太，Blog・Twitter に書かれた疑問を収集・提供する Web サイトの構築，情報知識学会誌，Vol.23，No.1，2013
[5] 財前涼，森嶋厚行，クラウドソーシングを用いた情報信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] 廣田壮一郎，笹野遼平，高村大也，奥村学，方言コーパス収集システムの構築，第 27 回人工知能学会全国大会，2013
[8] 黒澤義明，小川湧眞，竹沢寿幸，ファッションアイテム紹介分の構成要素自動分類，言語処理学会第 20 回年次大会発表論文集，pp.278-281,2014