

地域ユーザに着目した口コミツイート収集手法の提案

長島 里奈[†] 関 洋平^{††} 猪 圭^{†††}

[†] 筑波大学 情報学群 知識情報・図書館学類 〒305-8550 茨城県つくば市春日 1-2

^{††} 筑波大学 図書館情報メディア系 〒305-8550 茨城県つくば市春日 1-2

^{†††} つくば市役所 〒305-8555 茨城県つくば市研究学園 1-1-1

E-mail: [†]s1211530@u.tsukuba.ac.jp, ^{††}yohei@slis.tsukuba.ac.jp

あらまし 本研究では、店舗、自治体、地域住民が消費者の口コミを活用することを支援するために、地域店舗を対象に絞って「口コミツイート」を収集するための手法を提案する。提案手法は「地域ユーザ」を口コミツイート収集の手がかりとして用いる。しかし、居住地をプロフィールに記述するユーザは多くないことから、地域ユーザのカバー率向上のためにフォロー関係を活用する。地域活性化を目的とするイベントの出店店舗を対象とした口コミツイートの収集において、提案手法と、手がかりに地域ユーザを用いないベースラインとの比較実験を行った結果、提案手法による口コミツイート収集結果の F 値の平均は 0.677 となり、ベースラインから 93%の改善が見られた。特に、店舗名が多義性を持ち、かつ、店舗のある地域と結びつけることで店舗名が一意性を持つようになる店舗において、提案手法は有効であった。

キーワード Twitter, 口コミ, 地域活性化, 地域店舗

1. はじめに

1.1 本研究の概要

本研究では、Twitter^(注1)における、地域店舗を対象とした「口コミツイート」の収集手法について提案を行う。

近年、リアルタイム性や投稿の手軽さにより、Twitter 上に消費者の口コミが多数投稿されるようになった。手軽に発信されるため消費者の本音が出やすいという記述内容の特色や、データ量の優位性などの理由により、Twitter 上の口コミは、企業や自治体からも注目を集めている。本研究では、このような Twitter 上の口コミを「口コミツイート」とする。

一方、全国的に有名な、会社、商品、人物などを対象に絞った場合においては、口コミツイートが比較的収集しやすいのに対し、特定の地域においてのみ人気のある、地域店舗を対象に絞った場合は、口コミツイートの収集は難しい。その理由としては、地域店舗名が複数の意味で用いられる場合、地域店舗に関わるツイートだけを集めようとしても、地域店舗に適合しないツイートが混ざってしまうことが挙げられる。

このように、口コミツイートが消費者の反応を知るために有効なものであるにも関わらず、地域店舗を対象とした口コミツイートの活用は容易ではない。そのため、地域店舗を対象に絞った口コミツイートの収集手法を開発することが必要となる。

本研究では、先行研究で用いられている手がかりに、新たに「地域ユーザ」を加えることによって、地域店舗を対象に絞った口コミツイートの収集手法を提案する。これにより、店舗、自治体、地域住民が消費者の口コミを活用することが可能になると考える。

1.2 背景

近年、口コミツイートは、消費者の生の声を聞く手段として、企業や自治体などから注目を集めている。

実際に自治体による地域活性化に口コミツイートが活用された例として、「つくば福来らーめん紀行」^(注2)が挙げられる。このイベントは、ラーメン店で市の名産品をプロモーションするという主旨で行われた。イベントの開催にあわせて、著者らがイベントに出店した 12 店舗についての口コミツイートを収集し、つくば市と協力して、公式サイトへの掲載を行った。実際のページの一部を、図 1^(注2)に示す。



図 1 公式サイトに表示された口コミツイート

口コミツイートを掲載したサイトは、イベントが開催された 2015 年 8 月 1 日から 2015 年 9 月 30 日までの期間で、3,057pv を獲得することができた。しかし、イベント出店店舗の中には、複数の意味で用いられる単語を名称とする店舗が存在し、対象店舗に絞った口コミツイートの収集が困難であった。実際の例を図 2 に示す。この図における店舗名「小五郎」は、アニメ

(注1): <https://twitter.com>

(注2): <http://tsukuba-honmono.jp/noodle/>

「名探偵コナン」のキャラクターである「毛利小五郎」や、明治時代の政治家である「桂小五郎」を示す文脈で用いられることが多く、つくば市の店舗「麺や小五郎」を示す文脈で用いられているのは、2ヶ月間で収集した 21,238 ツイートのうちわずか 35 ツイートのみであった。このようなことは他のイベント出店店舗においても発生している。

以上のことを踏まえ、本研究では、店舗、自治体、地域住民が消費者の口コミを活用することを支援するために、地域店舗を対象に絞った口コミツイートの収集手法を提案する。

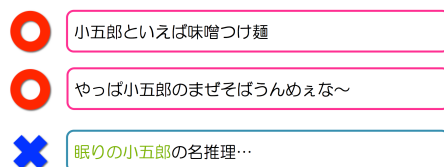


図 2 「小五郎」を含むツイートの例

本論文の構成を以下に示す。2 節は、口コミツイートの収集と、Twitter ユーザの居住地推定に関連する研究を紹介し、本研究の位置づけを述べる。3 節は、本研究の提案手法について述べる。4 節は、地域ユーザをプロフィールに基づいて収集した場合における、口コミツイート収集手法の比較実験について述べる。5 節は、フォロー関係を利用して地域ユーザのカバー率を向上させた場合における、口コミツイート収集手法の比較実験について述べる。6 節は、まとめについて述べる。

2. 関連研究

本研究は、地域ユーザを活用した口コミツイートの収集手法について提案を行う。そのため、2.1 節では、口コミツイートの収集に関する研究を紹介する。また、地域ユーザの推定の際に参考にした、SNS ユーザの居住地推定に関する研究について、2.2 節で述べる。

2.1 口コミツイートを収集する研究

那須川ら [8] は、「日本酒」という表現を用いて、特定地域の飲食店の評判に関わるツイートの収集を行った。彼らは、「日本酒」という表現を用いて収集したツイートの中から、東京都の区の名前や駅名を手がかりに、日本酒の美味しい店について言及されたツイートを検索した。しかし、これらの条件を満たし、かつ店舗名が含まれているツイートの収集件数は極めて少なかったため、特定の表現に偏らず、網羅的にツイートを収集することで収集件数を増やしていきたいと述べている。那須川らの考察を踏まえ、本研究では、地域ユーザに着目してツイートを収集することにより、地域名や商品名を含むという制約を排除し、網羅的な件数増加を狙う。

また、地域に特色のある語を手がかりに、観光情報にまつわるツイートを収集する研究を行った小原ら [5] の研究と、位置情報を利用して、観光にまつわるツイートを収集する研究を行った中嶋ら [7] の研究を踏まえ、地域に特色のある語と位置情報も提案手法に取り入れる。

2.2 SNS ユーザの居住地推定に関する研究

SNS ユーザの居住地推定について、コンテンツベースとグラフベースの 2 種類の研究を紹介する。

2.2.1 コンテンツベースによる居住地推定

Cheng ら [2] は、Twitter の投稿から、ローカルワードと呼ばれる地域に偏りのある単語を抽出し、これを手がかりにユーザの居住地推定を行っている。

日本語のコンテンツベースによる居住地推定の研究として、堂前ら [6] の研究が挙げられる。彼女らは、ツイート集合から生成した地域ラベルを付与したトピックを利用して、ユーザの生活に関わる地域を都道府県レベルで推定している。

コンテンツベース手法の欠点として、地域に偏りのある語を用いて一般的な話題に言及するユーザの居住地を誤って推定してしまう可能性があるため、高い精度の推定が難しい点が挙げられる。また、本研究の地域ユーザ推定対象範囲の市区町村は、特色が現れる語数、推定対象とするユーザ数ともに、都道府県を対象とする先行研究と比較して少ない。このことから、地域に偏りのある語を正しく選ぶことが難しいと考え、本研究ではコンテンツベースによる地域ユーザの推定は行わない。

2.2.2 グラフベースによる居住地推定

Backstrom ら [1] は、居住地が判明している友人を手がかりとして、Facebook^(注3) ユーザの居住地を推定する研究を行った。

Jurgens [3] は、Twitter ユーザのフォロー関係を利用して、ユーザの居住地を推定する研究を行った。彼は、位置情報付きツイートを発信する相互フォローユーザの居住地から、対象ユーザの居住地を推定している。推定を行う際に、有名人や BOT などフォロー関係の広すぎるユーザを除外するために、相互フォローユーザ数に上限を設けている。

本研究では、グラフベースによる SNS ユーザの居住地推定の研究を参考に、地域ユーザの推定を行う。手がかりとして用いるユーザの居住地には、Jurgens [3] の用いた位置情報付きツイートではなく、ユーザプロフィールの居住地についての記述を利用する。その理由としては、両者の使用率の違いが挙げられる。伊川 [4] の調査では、46.5%のユーザが居住地を設定しており、都道府県と市区名の記述のあるユーザに絞ると 15.5%となることが明らかにされている。これに対し、位置情報付きツイートの投稿数は全体のうち 0.46% [4] に過ぎない。以上のことを踏まえて、本研究では、ユーザプロフィールの居住地を参考に地域ユーザを決定し、そのユーザのフォロー関係に基づき、地域ユーザの推定を行う。

3. 提案：地域ユーザの活用による口コミツイートの収集手法

本研究では、地域店舗の店舗名を含むツイート集合から、地域店舗に実際に言及している口コミツイートを絞り込む。先行研究で用いられた地域名と位置情報を手がかりとする場合、収集できるツイート数は限られてしまう。その理由としては、Twitter の匿名性や投稿文字数の制限から、店舗を特定できる

(注3): <https://www.facebook.com>

地域名が含まれる口コミツイートが少ないことと、位置情報付きツイートの絶対数が少ないことが挙げられる。

この課題を解決するために、本研究では、図3に示すように、上記の2つに加えて地域ユーザを、地域店舗を対象に絞った口コミツイートを収集するための手がかりとして用いる。地域ユーザの活用により、地域名や位置情報を含まないツイートの収集が可能になり、口コミツイートを増やすことができるようになると思う。

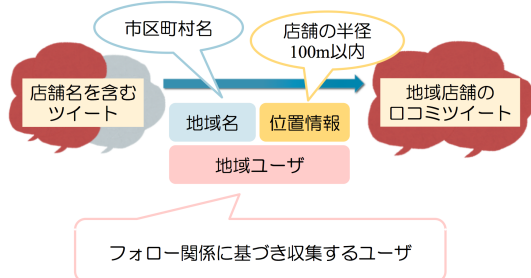


図3 地域ユーザを活用した口コミツイートの収集

しかし、Twitterにおいて、地域ユーザを収集することは容易ではない。Twitterは本来匿名性の高いサービスとなっており、ユーザを地域ごとに集約する公式のサービスは存在しない。また、非公式のサービスとして、Twitterユーザの地域などのプロフィール検索を行うことのできるツイプロ^(注4)が挙げられるが、1つの条件で検索できるユーザの数には上限があり、市区町村を居住地として記載するすべてのユーザを収集することはできない。

そこで、本研究では、2節で述べた先行研究を参考に、図4に示すように、フォロー関係を利用して地域ユーザを推定し、地域ユーザのカバー率を増加させる。具体的には、次に述べる2つの手順をとる。

- (1) ツイプロを利用し、対象とする地域店舗のある市区町村を居住地としてプロフィールに記述するユーザの収集を行う。
- (2) (1)で集めたユーザのフォロワーを収集し、5節で述べる実験に基づき、ある一定の条件を満たしたユーザを地域ユーザとする。

このようにして集めた地域ユーザを手がかりに加えることにより、地域店舗を対象に絞った口コミツイートの収集を行う。

4. 実験：地域ユーザを活用した口コミツイートの収集

4.1 実験の目的

本研究では、先行研究で用いられている、地域に特色のある語と位置情報に、新たに地域ユーザを手がかりとして加えた、口コミツイートの収集手法を提案する。また、地域ユーザの活用が、地域店舗を対象に絞った口コミツイートの収集に有効であるかを検証する。本節の実験では、地域ユーザとして、ツイ

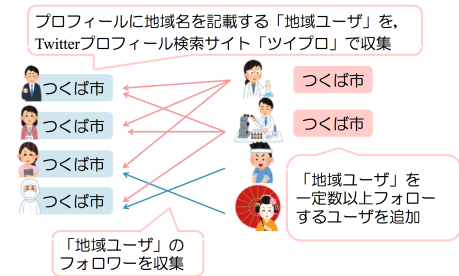


図4 地域ユーザの収集方法

プロで取得したTwitterのプロフィール欄につくば市を居住地として記述しているユーザを設定し、地域名と位置情報のみを手がかりとして用いたベースラインとの比較を行う。

4.2 実験データ

実験では、1節で述べた「つくば福来らーめん紀行」に出店する12店舗を対象に絞った口コミツイートの収集を行う。使用したデータは、2015年8月1日から2015年9月30日までの2ヶ月間、Twitterの非公式ライブラリであるTwitter4j^(注5)を利用し、12店舗それぞれの店舗名をクエリとして収集したツイートである。ただし、ノイズを除去するために、以下のものは取り除いた。

- 「RT」が文頭にくるツイート
- 「I'm at」が文頭にくるツイート
- 「@」が文頭にくるツイート
- IDの末尾に「bot」が含まれるユーザによるツイート

ノイズ除去作業を終えた結果、12店舗あわせて132,316ツイートを収集することができた。実験を行うために、人手により、収集したツイートの内容が対象店舗に関連しているものについて「適合している」というラベル付けを行う。これに該当しない、対象店舗に言及していないツイートには「適合していない」というラベル付けを行う。

適合しているツイートの例を、図5に示す。例1のように、つくば市のラーメン店についてのツイートであることが明記されている場合は、適合していると判断する。また、例2のように、店舗特有のメニュー名（「スタミナ冷やし」）がツイートにあらわれている場合や、例3のように、他のつくば市のラーメン店の店舗名（「三水」）がツイートにあらわれている場合も、適合していると判断する。その他にも、投稿ユーザの普段のツイートを確認した結果、つくば市のラーメン店のツイートであることが明確である場合は、適合していると判断する。

1. やっぱつくばきたら太樹の味噌ラーメンでしょ！
2. 久々にスタミナ冷やしのがむしゃへ
3. 三水より小五郎のが美味いよ

図5 適合しているツイートの例

(注4): <http://twpro.jp>

(注5): <http://twitter4j.org/>

適合していないツイートの例を、図 6 に示す。例 1 のように、店舗名が対象店舗以外を示す語として用いられているツイートの場合、適合していないと判断する。また、例 2 のように、同一名だが明らかに別の地域（「町田」）の店舗について述べているツイートも、適合していないと判断する。例 3 のように、文脈上店舗を表すものとして用いられていない場合も、適合していないと判断する。

1. 店員がEXILEの佐藤大樹に似るとる（笑）
2. 町田大勝軒の豚増しラーメン食べたい
3. 鶏が思った以上に鶏肉しかかった上に骨も多いし食べるの大変やったw

図 6 適合していないツイートの例

データの内訳

クエリを用いて収集を行った、店舗ごとのツイート数と、そのうちの適合するツイート数を表 1 に示す。また、適合するツイートを選択する際に、ツイート数の多さや店舗の特徴などの理由により、例外的にクエリの工夫を行った店舗について、以下に述べる。

- 「大樹」については、「ラーメン」「麺」「らーめん」のいずれかの語を本文中に含むツイートを抜き出し、その中から適合するものを選択した。
- 「大勝軒」については、対象店舗周辺の位置情報を持つものと、「つくば」もしくは「うさぎ」^(注6)と共起するツイートを抜き出し、その中から適合するものを選択した。

表 1 店舗名をクエリとしたツイートの収集結果

クエリ	店舗の正式名称	全ツイート数	適合ツイート数
油虎	油そば 油虎	697	697
おび屋	鶏白湯らーめん おび屋	45	34
活龍	つけめん・らーめん 活龍	738	687
がむしゃ	スタミナラーメン がむしゃ	615	129
喜元門	喜元門	498	498
五衛門	関東風とんこつらーめん 五衛門	461	19
小五郎	麺や 小五郎	21,238	35
たちばな	熟成蔵出し味噌 麺処 たちばな	5,395	18
鶏々	鶏々 TORIDORI	213	195
松辰	麺や 松辰	239	154
大樹	麺屋 大樹	92,756	9
大勝軒	東池袋 大勝軒 うさぎ家	9,421	30

4.3 実験方法

以下で説明する、提案手法、ベースライン、元データによる口コミツイート収集結果について比較を行う。

提案手法

地域ユーザ、地域名、位置情報を用いて、店舗名を含むツイート集合から、地域店舗を対象に絞った口コミツイートの収集を行う手法を提案手法とする。地域名を含むツイートは「つくば」を含むツイートとし、位置情報を含むツイートは、対象店舗の半径 100 メートル以内で発信されたツイートとする。また、本節の実験では、地域ユーザによるツイートは、Twitter のプロフィール検索サイトであるツイプロ^(注1)を用いて取得した、プロフィールの居住地を「つくば市」に設定するユーザ 4,548 人が発信するツイートとする。

ベースライン

提案手法から地域ユーザを除き、地域名と位置情報のみを用いて、店舗名を含むツイート集合から、地域店舗を対象に絞った口コミツイートの収集を行う手法をベースラインとする。

元データ

店舗名をクエリとして収集したツイート集合を元データとする。

評価尺度

実験結果の評価尺度としては、精度、再現率、F 値を用いる。具体的な計算方法を式 (1) ~ 式 (3) に示す。

精度 = $\frac{\text{手法で収集したツイート内の適合ツイート数}}{\text{手法で収集したツイートの総数}}$ (1)

再現率 = $\frac{\text{手法で収集したツイート内の適合ツイート数}}{\text{元データ内の適合ツイートの総数}}$ (2)

F 値 = $\frac{2}{\frac{1}{\text{精度}} + \frac{1}{\text{再現率}}}$ (3)

4.4 実験結果

提案手法を、ベースライン、元データと比較した実験結果を、表 2 に示す。提案手法をベースラインと比較した結果、F 値の改善が確認できた（有意水準 1%，両側検定で有意差あり）。

また、地域ユーザを手がかりに加えたことによって、図 7 に示すような、地域名^(注7)を含まないツイートも収集できるようになった。

- がむしゃ並んでる^^;
- 鶏々は火金が定休日になったんだ...!
- 松辰の期間限定、秋刀魚煮干しラーメン。とても美味しくて満足

図 7 地域ユーザの活用により収集できるようになったツイート

(注6): 店舗の正式名称に含まれる「うさぎ家」より

(注7): 本実験では「つくば」とする

表 2 提案手法、ベースライン、元データの比較

店舗名	提案手法			ベースライン			元データ
	F 値	精度	再現率	F 値	精度	再現率	精度
油虎	0.366	1.000	0.223	0.312	1.000	0.185	1.000
おび屋	0.522	1.000	0.353	0.300	1.000	0.176	0.756
活龍	0.259	1.000	0.148	0.218	1.000	0.122	0.931
がむしゃ	0.602	0.982	0.434	0.506	0.978	0.341	0.210
喜元門	0.516	1.000	0.347	0.359	1.000	0.219	1.000
五衛門	0.733	1.000	0.579	0.538	1.000	0.368	0.041
小五郎	0.408	0.714	0.286	0.200	0.800	0.114	0.002
たちばな	0.286	1.000	0.167	0.286	1.000	0.167	0.003
鶏々	0.354	1.000	0.215	0.274	1.000	0.159	0.915
松辰	0.613	1.000	0.442	0.308	1.000	0.182	0.644
大樹	0.083	0.043	0.889	0.090	0.048	0.889	0.000
大勝軒	0.772	0.815	0.733	0.815	0.917	0.733	0.003
平均	0.459*	0.880	0.401	0.351	0.895	0.305	0.459

* ベースラインの F 値との比較において、t-検定で有意差あり (有意水準 1%)

4.5 考 察

ベースラインと比較して F 値が改善した店舗

「油虎」「喜元門」「たちばな」を除く、12 店舗中 9 店舗がこれにあたる。これらの店舗においては、ベースラインに対し、同程度の精度を保ったまま、適合するツイートが増加して再現率が上がったため、F 値の改善が見られる。

(1) 元データと比較して精度が改善した店舗

「おび屋」「活龍」「がむしゃ」「五衛門」「鶏々」「松辰」、小五郎」の 7 店舗がこれにあたる。

元データの精度が特に低い「五衛門」(0.041)と「小五郎」(0.002)は、店舗名が人名やキャラクター名としてよく使用される。不適合としたツイートには、名探偵コナンのキャラクターの「毛利小五郎」や、ルパン三世のキャラクターである「石川五ェ門」の誤字などが含まれる。このような、店舗名が人名としても用いられる店舗においては、地域店舗を対象に絞ったツイートの収集は難しいが、提案手法により、元データからこれらの不適合なツイートを取り除くことができ、精度が改善することを確認した。

その他の 5 店舗も、店舗名が多義性を持つ。不適合としたツイートには、「おび屋」に対する他の店舗の店舗名である「くろおび屋」や、「がむしゃ」に対する関ジャニの曲名の「がむしゃら行進曲」などが含まれる。提案手法により、元データからこれらの不適合なツイートをほぼ取り除くことができ、精度が改善した。

以上のことを踏まえると、提案手法は、店舗名が多義性を持ち、かつ、店舗のある地域と結びつけることで店舗名が一意性を持つようになる店舗において有効であるといえる。

(2) 元データと比較して精度が変化しなかった店舗

「油虎」「喜元門」がこれにあたる。この 2 店舗は、店舗名が極めて特徴的であるため、元データのすべてが適合するツイートとなった。したがって、提案手法を適用しなくても、地域店舗を対象に絞った口コミツイートの収集が可能であるといえる。

ベースラインと比較して F 値が変化しなかった店舗

「たちばな」がこれにあたる。ベースラインに対し、ツイートの収集結果の変化がなかったため、F 値も変化しなかった。こうした店舗については、提案手法の改善により、地域ユーザの数を増加させることで、F 値の改善が見込めるか検証する。

ベースラインと比較して F 値が悪化した店舗

「大樹」「大勝軒」がこれにあたる。「大樹」については、極めて使用頻度の高い人名であるため、店舗のある地域においても、店舗名としてではなく、人名として使用されることの方が多いと考える。また、「大勝軒」については、全国的に有名であり、かつ、同じ名前だが味の系統が違う店舗が全国各地に多数存在する。そのため、つくば市ユーザによるツイートの中にも、「神田大勝軒」「柏の大勝軒」など、「別の地域名 + 大勝軒」という形で不適合なツイートが現れている。

以上のことを踏まえると、これらの店舗においては、地域ユーザの発信するツイートを対象としても、店舗名が対象とする地域店舗を示すものとして一意にはならないといえる。したがって、ツイート収集件数の増加に伴い精度が落ち、ベースラインと比較して F 値が悪化したと考える。

全体の総括

提案手法とベースラインを比較すると、F 値が 31%改善し、0.459 となったが、この値には改善の余地があると考えられる。全体的に適合するツイート数があまり伸びず、再現率の改善が小さいことが原因である。適合ツイート数が伸びなかった要因としては、地域ユーザのカバー率不足が挙げられる。プロフィール欄に居住地を記述しているユーザが全体の 15.5% [4] のみであることから、プロフィールに基づくユーザの収集だけでは、地域ユーザのカバー率は不十分である。5 節では、本節で収集した地域ユーザのフォロー関係を用いて地域ユーザのカバー率を改善させることで、適合するツイート数が増加し、口コミツイートを収集する際の F 値が改善するか検証する。

5. 実験：フォロー関係に基づく地域ユーザのカバー率の向上

5.1 実験の目的

本節では、フォロー関係を利用して地域ユーザのカバー率を向上させることが、口コミツイートの収集に有効か検証を行う。

5.2 フォロー関係に基づく地域ユーザのカバー率の向上 (1 段階目)

本節の実験は、プロフィールに基づき収集した地域ユーザ (4 節参照) のフォロー関係を利用した、地域ユーザのカバー率の向上を図る。

5.2.1 実験方法

使用した実験データ、評価尺度、ベースラインと元データについては、4 節と同様である。提案手法については、地域ユーザの収集方法のみを変更する。

フォロー関係を利用した地域ユーザの推定を行うにあたっては、先行研究 [3] を踏まえ、有名人や BOT などのノイズとなりにやすいアカウントを取り除くために、フォロー数やフォロワー数の上限を設定する。また、地域ユーザのフォロー人数が多け

れば、フォローしているユーザの居住地が同一である可能性は高まる [1] [3] . これらを踏まえて、3 つの基準値を設ける .

- (1) フォロワー総数の上限
- (2) フォロワー総数の上限
- (3) プロフィールに基づき収集した地域ユーザのフォロー人数の下限

これらの基準値をそれぞれ変化させて、口コミツイートの収集結果の F 値が最も高くなる値を、最適な値として採用する .

(1) フォロワー総数の上限

プロフィールに基づき収集した地域ユーザのフォロー人数が 5 人以上、フォロワー総数が 3,000 人以下のユーザにおいて、フォロワー総数が 6,000 人以下、5,000 人以下、4,000 人以下、3,000 人以下、2,000 人以下それぞれの場合の口コミツイートの収集結果を比較する .

(2) フォロワー総数の上限

プロフィールに基づき収集した地域ユーザのフォロー人数が 5 人以上、フォロワー総数が 4,000 人以下のユーザにおいて、フォロワー総数が 6,000 人以下、5,000 人以下、4,000 人以下、3,000 人以下、2,000 人以下それぞれの場合の口コミツイートの収集結果を比較する .

(3) プロフィールに基づき収集した地域ユーザのフォロー人数の下限

フォロワー総数が 4,000 人以下、フォロワー総数が 3,000 人以下のユーザにおいて、4 節で集めた地域ユーザのフォロー人数が 4 人以上、5 人以上、6 人以上、7 人以上、8 人以上それぞれの場合の口コミツイートの収集結果を比較する .

5.2.2 実験結果

フォロワー総数を变化させた実験の結果を、図 8、図 9 に示す . フォロワー総数の上限が 4,000 人以下より増えると、再現率がほとんど変化せず、4,000 人以下と 5,000 人以下の間を境目に精度が大きく落ちている . F 値が最も高い条件を選択すると、フォロワー総数が 4,000 人以下の場合となった .

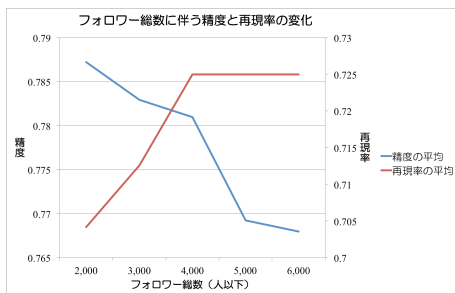


図 8 フォロワー総数に伴う精度と再現率の変化

フォロワー総数の上限を变化させた実験の結果を、図 10、図 11 に示す . フォロワー総数が 3,000 人以下より増えると、再現率がほとんど変化せず、精度は 3,000 人以下と 4,000 人以下の間でわずかに落ちていく . F 値が最も高い条件を選択すると、

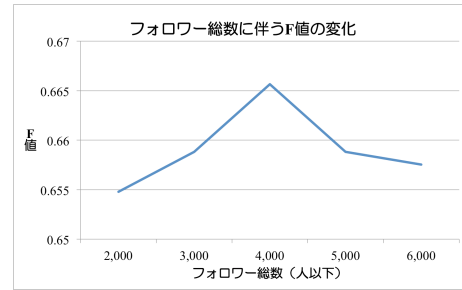


図 9 フォロワー総数に伴う F 値の変化

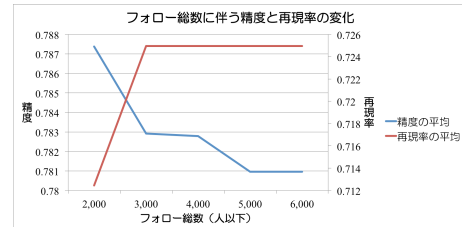


図 10 フォロワー総数に伴う精度と再現率の変化

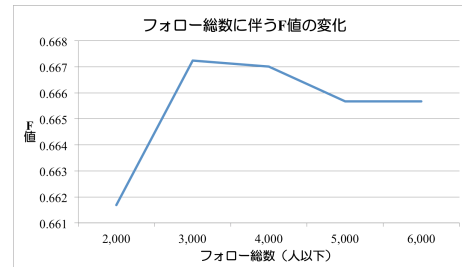


図 11 フォロワー総数に伴う F 値の変化

フォロワー総数が 3,000 人以下の場合となった .

プロフィールに基づき収集した地域ユーザのフォロー人数の下限を变化させた実験の結果を、図 12、図 13 に示す . プロフィールに基づき収集した地域ユーザのフォロー人数が増えるにつれて再現率が落ち、精度は 5 人以上と 6 人以上のときに高い値となっている . F 値が最も高い条件を選択すると、プロフィールに基づき収集した地域ユーザのフォロー人数が 5 人以上の場合となった .

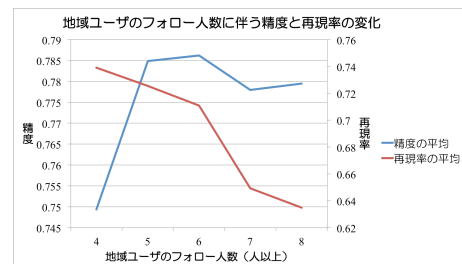


図 12 プロフィールに基づき収集した地域ユーザのフォロー人数に伴う精度と再現率の変化

1 段階目のまとめ

実験の結果、プロフィールに基づき収集した地域ユーザを利用した、1 段階目の地域ユーザのカバー率の向上における基準値を、以下のように決定した .

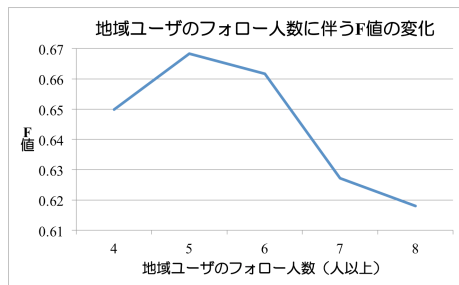


図 13 プロフィールに基づき収集した地域ユーザのフォロー人数に伴う F 値の変化

- (1) フォロワー総数の上限：4,000 人以下
- (2) フォロー総数の上限：3,000 人以下
- (3) プロフィールに基づき収集した地域ユーザのフォロー人数の下限：5 人以上

カバー率を向上させた結果、地域ユーザ数は 27,589 人、提案手法に基づいて収集を行った口コミツイトの F 値は 0.668 となった。

5.3 フォロー関係に基づく地域ユーザのカバー率の向上 (2 段階目)

本節では、1 段階目の結果から得られた地域ユーザに対するフォロー関係を利用した地域ユーザの推定に基づき、地域ユーザのカバー率の向上を図る。

5.3.1 実験方法

使用した実験データ、評価尺度、ベースラインについては、4 節と同様である。提案手法については、地域ユーザの収集方法のみを変更する。BOT や有名人を取り除くための基準である、フォロワー総数の上限、フォロー総数の上限については、1 段階目のカバー率の向上で定めた基準値をそのまま利用してユーザの収集を行う。しかし、地域との結びつきの強さを計る基準値である、地域ユーザのフォロー人数の下限については、1 段階目で収集したユーザにより、地域ユーザのカバー率が向上していることを踏まえ、値を調整する必要がある。したがって、フォロワー総数 4,000 人以下、フォロー総数 3,000 人以下のユーザにおいて、1 段階目の結果から得られた地域ユーザのフォロー人数が 5 人以上、6 人以上、7 人以上、8 人以上、9 人以上、10 人以上、11 人以上それぞれの場合の口コミツイトの収集結果の比較を行い、F 値が最良となる、最適な値を決定する。

5.3.2 実験結果

1 段階目の結果から得られた地域ユーザのフォロー人数のみを変化させ、カバー率を向上させた地域ユーザを利用して、口コミツイトの収集を行った結果を、図 14、図 15 に示す。1 段階目の結果から得られた地域ユーザのフォロー人数が増えるにつれて再現率が落ち、9 人以上と 10 人以上の間を境目に精度が大きく改善している。F 値が最も高い条件を選択すると、1 段階目の結果から得られた地域ユーザのフォロー人数が 10 人以上の場合となった。

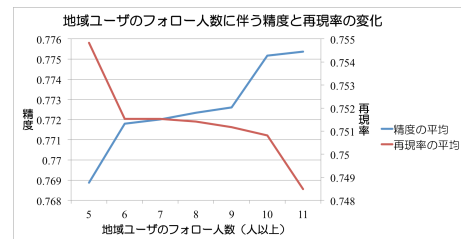


図 14 1 段階目の結果から得られた地域ユーザのフォロー人数に伴う精度と再現率の変化

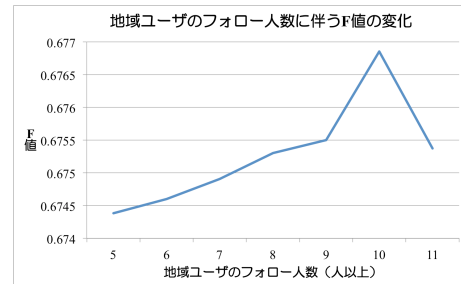


図 15 1 段階目の結果から得られた地域ユーザのフォロー人数に伴う F 値の変化

地域ユーザの収集方法において、プロフィールのみに基づいた場合、1 段階目のカバー率向上を行った場合、2 段階目のカバー率向上を行った場合を比較した結果を表 3 に示す。プロフィールのみに基づいた場合と比較すると、1 段階目、2 段階目ともに大きくユーザ数と F 値の改善が見られる。また、2 段階で地域ユーザのカバー率を向上させたことにより、1 段階目と比較して精度は落ちたが再現率が増加し、わずかであるが F 値も改善している。この結果から、2 段階目のカバー率の向上が、口コミツイト収集に関して、有効であることが明らかとなった。

表 3 地域ユーザの収集方法を変化させた口コミツイトの収集結果

地域ユーザの収集方法	ユーザ数	F 値	精度	再現率
プロフィールのみを利用	4,548 人	0.459	0.880	0.401
カバー率向上 (1 段階目)	27,589 人	0.668	0.785	0.726
カバー率向上 (2 段階目)	31,666 人	0.677	0.775	0.751

また、地域ユーザのカバー率を向上した場合と、プロフィールに基づき地域ユーザを収集した場合 (4 節) との、店舗別の口コミツイト収集の結果をまとめて掲載したものを、表 4 に示す。

実験の結果、地域ユーザのカバー率の向上に伴い、プロフィールに基づき地域ユーザを収集した場合と比較して、F 値についての改善が確認できた (有意水準 5%, 両側検定で有意差あり)。

5.4 考察

F 値が改善した店舗

「大樹」「大勝軒」を除く 12 店舗中 10 店舗がこれにあたる。地域ユーザのカバー率を向上させたことによって、適合するツイトが増え、再現率が大きくあがったことが F 値が改善した要因と考える。

表 4 地域ユーザのカバー率を向上した場合とプロフィールに基づき地域ユーザを収集した場合との口コミツイート収集の比較

店舗名	フォロー関係でカバー率を向上			プロフィールのみを利用		
	F 値	精度	再現率	F 値	精度	再現率
油虎	0.822	1.000	0.697	0.366	1.000	0.223
おび屋	0.692	1.000	0.529	0.522	1.000	0.353
活龍	0.606	1.000	0.435	0.259	1.000	0.148
がむしゃ	0.802	0.989	0.674	0.602	0.982	0.434
喜元門	0.782	1.000	0.643	0.516	1.000	0.347
五衛門	0.865	0.889	0.842	0.733	1.000	0.579
小五郎	0.733	0.600	0.943	0.408	0.714	0.286
たちばな	0.632	0.600	0.667	0.286	1.000	0.167
鶏々	0.870	1.000	0.769	0.354	1.000	0.215
松辰	0.934	1.000	0.877	0.613	1.000	0.442
大樹	0.073	0.038	1.000	0.083	0.043	0.889
大勝軒	0.311	0.187	0.933	0.772	0.815	0.733
平均	0.677*	0.775	0.751	0.459	0.880	0.401

* ベースラインの F 値との比較において、t-検定で有意差あり (有意水準 5%)

F 値が悪化した店舗

「大樹」「大勝軒」がこれにあたる。4 節で述べた通り、これらの店舗においては、地域ユーザの発信するツイートを対象としても、店舗名が対象とする地域店舗を示すものとして一意にはならないといえる。したがって、地域ユーザのカバー率を向上させて、地域ユーザが発信するツイートの収集件数が増えるのに伴い、対象店舗に適合しないツイートも増加して精度が落ち、ベースラインと比較して F 値が悪化したと考える。

6. おわりに

本研究では、店舗、自治体、地域住民が消費者の口コミを活用することを支援するために、地域店舗を対象に絞った「口コミツイート」の収集を行う手法について検証した。地域店舗名は複数の意味で用いられることがあり、このような場合、地域店舗に適合しないツイートが混ざってしまう。本研究では、この問題の解決を目指し、地域ユーザを活用することにより、地域店舗を対象に絞った口コミツイートを収集する手法について提案した。また、地域活性化を目的とするイベントに出店した店舗について、提案手法とベースラインとの比較実験を行い、地域ユーザを活用することの有効性について検証した。

Twitter のプロフィール欄の記述内容から収集した地域ユーザを活用して口コミツイートを収集した結果、ベースラインから F 値が 31%改善することを確認できた。

また、プロフィールに基づき収集した地域ユーザのフォロー関係を利用して、地域ユーザのカバー率を向上させた。フォロー総数の下限、フォロー総数の下限、地域ユーザのフォロー数の上限について、それぞれの基準値を求め、カバー率を向上させた地域ユーザを活用して口コミツイートを収集した結果、地域ユーザを活用しないベースラインからは 93%、プロフィールのみに基づき地域ユーザを収集した場合には 47%、F 値が改善することを確認できた。

地域ユーザの活用により、地域名を含まないツイートも収

集できるようになり、フォロー関係を利用して地域ユーザのカバー率を向上させることにより、ツイート件数を増加させることができるようになった。

提案手法は、店舗名が多義性を持ち、かつ、店舗のある地域と結びつけることで店舗名が一意を持つようになる店舗において有効であるといえる。一方、地域のツイートを対象としても店舗名が一意にならない店舗では、提案手法を用いることの効果はあまり得られなかった。

謝 辞

本研究の一部は、筑波大学研究基盤支援プログラム (B タイプ)、科学研究費補助金基盤研究 B (課題番号 25280110)、萌芽研究 (課題番号 25540159) の助成を受けて遂行された。

文 献

- [1] L. Backstrom, E. Sun, and C. Marlow. That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships. In Proceedings of the 19th International Conference on World Wide Web (WWW 2010), pp.61-70, New York, USA, 2010.
- [2] Z. Cheng, J. Caverlee, and K. Lee. You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM 2010), pp.759-768, Toronto, Canada, 2010.
- [3] D. Jurgens. That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships. In Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM 2013), pp.273-282, Boston, USA, 2013.
- [4] 伊川洋平. ソーシャルメディア位置情報分析を行う前の Tips. ARG 第 7 回 Web インテリジェンスとインタラクション研究会報告, 2015. 特別セッション「不揃いなデータ達の分析を行う前の Tips」. <http://www.slideshare.net/Yoheikawa/tips-55961281> (accessed 2015-12-23).
- [5] 小原基季, 森田和宏, 泓田正雄, 青江順一. Twitter 本文を用いた観光情報抽出及び分析システムの構築. 人工知能学会第 29 回全国大会, 4M1-4, 2015.
- [6] 堂前友貴, 関洋平. 半教師ありトピックモデルにより選択した地域特徴語を用いた Twitter ユーザの生活に関わる地域の推定. 情報処理学会論文誌 データベース, Vol.7, No.3, pp.1-13, 2014.
- [7] 中嶋勇人, 新妻弘崇, 太田学. 位置情報付きツイートを利用した観光ルート推薦. 情報処理学会研究報告 データベースシステム (DBS), Vol.2013-DBS-158, No.28, pp.1-6, 2013.
- [8] 那須川哲哉, 吉田一星, 西山莉紗, 吉川克正, 伊川洋平, 大野正樹, 金山博, 鈴木祥子, 村上明子. 大量のつぶやきから日本酒の美味しい店を発掘する. 言語処理学会第 21 回年次大会発表論文集, pp.820-823, 2015.