2Q-5

マイクロブログの解析による活動範囲の推定とその応用

出良　隆明†　　中村　健二‡　　小柳　滋†
†立命館大学情報理工学部　　‡大阪経済大学情報社会学部

1　はじめに

スマートフォンやタブレット端末の普及により Web と現実世界とのつながりが密になっている．Facebook などの SNS や GoogleMAP などの地図サービス，ブログなどの日記投稿サービスなどがある．その中でも気軽に投稿できるマイクロブログに注目が集まっている．マイクロブログとは，チャットとブログの中間的なサービスであり，その多くは比較的短いテキストを投稿できる．　マイクロブログにはユーザの行動や多くの話題語が含まれている．これらの情報を解析することで，現実世界の状況を把握する研究 [1] や，マイクロブログからユーザの興味を推定し情報を推薦を行う研究 [2] が実施されている．

一般的に飲食店などの地物情報を推薦する研究では，ユーザの現在地に近く，ユーザの興味に合った場所を推薦する．しかし，興味に沿った現在地に近いものを推薦した場合，ユーザの状況を考慮できておらず，一様の推薦がなされるという課題がある．ユーザが出張や旅行で地方に出かけた場合には，その地方の特色が出た場所が推薦された方がユーザの満足度は高いと考えられる．そのため，本研究ではマイクロブログユーザの活動範囲を推定する手法を提案する．このことで，普段の活動範囲内にいる場合と活動範囲外にいる場合とで推薦内容の切り替えを支援する．

2　関連研究

本研究に関連する研究としてジオタグが付加された投稿を分析することで，イベントを発見する研究 [3] やジオタグに着目して投稿位置を推定する研究 [4] がある．　前者の研究 [3] ではイベント発生時のパターンをモデル化することで，地域的なイベントの発生とその影響範囲を推定する．具体的には，ある特定の地域においてイベントが発生すると，多くの人が一斉に集まり，その後離れていくという人の行動パターンを定義し，このパターンを用いてイベントを発見している．

後者の研究 [4] では位置情報の候補地をスコアリングすることで複数の候補地から位置情報を決定するという手法である．しかし、この研究ではユーザ単位の評価を行っておらず本研究では利用できない．

Estimating Range of Activity by Analysing Microblog
†Takaaki IDERA　‡Kenji NAKAMURA　†Shigeru OYANAGI
†Colleg of Information Science and Engennering Ritsumeikan University
‡Faculty of Imformation Technology and Social Science Osaka University of Economics

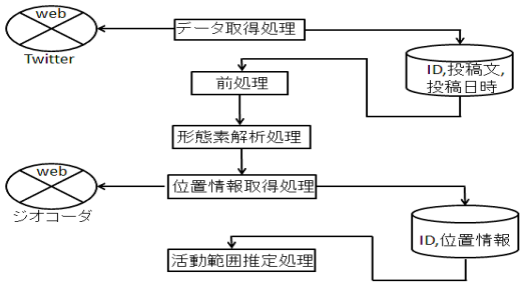

図 1: 処理フロー

3　提案手法

3.1　研究概要

本研究は処理フロー (図 1) に示す 5 つの処理から構成される．なお本研究ではマイクロブログの中でも広く利用されている「Twitter」[*1]を用いる．データ取得処理では Twitter からユーザ ID，投稿文，投稿日時を取得する．前処理ではハッシュタグ，URL，RT，@などをノイズとして除去する．形態素解析処理では前処理で加工した投稿文を MeCab [*2] を用いて解析する．位置情報取得処理では形態素解析結果を基にジオワードを生成し Google Geocoding API [*3] を用いて位置情報を取得する．ジオワードとは地名や固有名詞などの地物情報を含む単語のことでジオコーダによって位置情報を取得できる単語である．活動範囲推定処理では 1 人のユーザから取得した複数の位置情報をクラスタリングを行い活動範囲を推定する．

3.2　投稿文章の前処理

Twitter には「なう」,「うぃる」,「わず」といったように行動や場所などに「いつ行うか」といったスラング的な情報が付加されている．また「＃○○」といったハッシュタグ,「RT　○○」や「@○○」といった他のユーザの投稿のリツイートや特定のユーザに対する返信といった情報も付加されいる場合が多い．これらの情報を除去する事によって形態素解析の精度向上を図る．

3.3　位置情報取得

ジオワードは複数の単語に分解されていることが多い．このため，形態素解析された結果から地名や人姓に

[*1] https://dev.twitter.com
[*2] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[*3] http://code.google.com/intl/ja/apis/maps

表 1: ジオワード

投稿文章	加工なし	加工あり
南草津、初めて降りた。	草津	南草津
地下鉄で移動して来たー京都駅なう！	京都	ー京都駅

表 2: 投稿回数別ユーザ人数

回数	1	2〜5	6〜10	11〜20	21〜30	30〜
人数	147754	21052	932	325	75	34
割合	86.83 %	12.37 %	0.55 %	0.19 %	0.04 %	0.02 %

注目し，複数の単語を組み合わせてジオワードを抽出する．単語の組み合わせはジオワードがどのように分解されやすいかを予備実験により確認した．その結果を基にジオワードになりやすい組み合わせを生成した．

1. [名詞]+[名詞 (地名 or 人姓)]+[名詞]
2. [名詞]+[格助詞 (に)]

抽出したジオワードをジオエンコーダにかけて一意に位置が確定する場合もあるが，多くの場合は複数の候補地が出力される．そのため，その投稿の場所よりも前にいた位置を基準として，最も距離が小さい場所を候補地として抽出する．

3.4 活動範囲推定

取得した複数の位置情報を用いて活動範囲の推定を行う．この際に普段の活動範囲と旅行など非日常的活動範囲を判定するために k-means 法を用いて，クラスタリングする．しかし，クラスタリングする際に初期値によって偏ってしまうので時系列的にノードを追加していき閾値以下の距離にほかのノードがない場合は新しいクラスタとして定義する．これにより初期値によって結果が左右される問題点を解決する．

4 評価実験と考察

4.1 ジオワード生成

作成した組み合わせ表を用いた場合と地名のみをジオワードとした場合の結果を表 1 に示す．ランダムに選択した同じ投稿文章を対象とし，ジオワードは何が抽出されたかを示した．ジオワードの正規化によって正しく検出できるものもあれば文章によっては悪化する場合もあった．今後の課題として地名や人姓のみに着目するのではなく固有名詞にも着目することでより多くのジオワードを検出できるのではないかと考えられる．

4.2 位置情報推定の精度評価実験

評価実験は 2012/12/13〜19 に投稿されたデータを用いた．取得できたつぶやきは 8868899 件そのうち地名などのジオワードが取得できたのは 213205 件であった．複数回ジオワードをつぶやいているユーザがどの程度いるかを表 2 に示す．また，それぞれの場合で活動範囲の推定を行い，ランダムに 20 人を抽出し目視により確認した．その結果を表 3 に示す．目視による確認はユーザ ID のみを抽出しそのユーザの過去の投稿やプロフィールページを参照した．

表 3: 活動範囲推定人数

回数	1	2〜5	6〜10	11〜20	21〜30	30〜
人数	1	3	8	10	18	17
割合	5 %	15 %	40 %	50 %	90 %	85 %

位置情報がユーザあたり 21 個以上あれば高い精度で推定できた．しかし bot などの現実世界との関係が無いユーザに対しての位置情報も取得してしまった．よって今後は一定数以上の位置情報を取得できたユーザのみを解析するといった手法で全体の推定精度の向上を目指す．また，bot の判定処理を追加して bot を解析対象から外す処理が必要である．

5 おわりに

本研究では Twitter の投稿から位置情報を取得しそれに基づきユーザの活動範囲を推定する．一定数以上の位置情報を取得できれば活動範囲の推定は可能であることがわかった．今後は，より多くの位置情報を取得できるようにジオワード生成や解析対象を絞ることによって精度の向上を図りたい．

参考文献

[1] Java, A., Song, X,. Finin,T. and Tseng, B.: Why We Twitter:Understanding Microblogging Usage and Communities, *Proceeding of the 9th WebKDD and 1st SNA-KDD 2007*, ACM, pp.56-65(2007)

[2] 向井友宏, 黒澤義明, 目良和也, 竹澤寿幸:マイクロブログの分析に基づくユーザの嗜好とタイミングを考慮した情報推薦手法の提案，言語処理学会第 17 回年次大会発表論文集, 言語処理学会,pp.452-455(2011)

[3] 藤坂達也, 李 龍, 角谷和俊:実空間マイクロブログ分析による地域イベントの影響範囲推定, 第 2 回データ工学と情報マネジメントに関するフォーラム (DEIM2010), 電子情報通信学会 (2010)

[4] 河野愛樹, 中村健二, 小柳滋:マイクロブログから抽出した地物情報と投稿間隔を考慮した位置情報推定, 情報処理学会全国大会講演論文集, 情報処理学会,Vol.73,No.1,pp.785-786,(2011)