様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　５　月　３１　日現在

機関番号：　１４４０１
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2017
課題番号：　２６３３０１３７

研究課題名（和文）マイクロブログからのユーザ適応型実世界観測情報検索システムの構築

研究課題名（英文）Development of Personalized Real-World Observation Retrieval System from Microblog

研究代表者

新田　直子（NITTA, NAOKO）

大阪大学・工学研究科　・准教授

研究者番号：００３７９１３２

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：本研究では、実世界において観測される多様な対象に関するリアルタイムな観測情報をマイクロブログから自動的に収集した上で、個々のユーザの現在地・関心に応じて適切な情報を提示するシステムを構築することを目的とする。これを実現するため、マイクロブログへの不特定多数のユーザの投稿から実世界の多様な観測対象を抽出する手法、及び、抽出された観測対象のうち、各ユーザの関心に応じたものに対する観測情報を抽出する手法を開発した。

研究成果の概要（英文）：The goal of this research is to develop a system for retrieving relevant real-time real-world observations from Microblog according to each user's current position and interests. The developed system iteratively collects various targets of interest observed by Microblog users, and then extracts their observations by considering their semantic relevancy to user's interests.

研究分野：　情報学　知覚情報処理

キーワード：　実世界センシング　ソーシャルセンサ　マイクロブログ　ユーザ適応　テキスト検索

１．研究開始当初の背景

GPS が搭載されたスマートフォンやタブレットなどのモバイル端末のユーザが急増すると共に、多くのユーザが、実世界の各地で自身が観測した情報を、観測時間、場所の情報と共にソーシャルネットワーキングサービス（SNS）に投稿するようになった。人間は自らの五感を用いて多様な情報を観測するのみでなく、その意味を解釈する機能も持つため、人間をセンサとみなし、SNS への投稿をこれらのセンサによる観測情報として利用することは、センサ設置のコストを抑えた上で、多様な実世界観測情報を収集できるという利点がある。特に、SNS の中でも Twitter に代表されるマイクロブログには、主に短いテキスト形式で表した自身の状況や雑記などが投稿され、その手軽さからリアルタイム性の高い情報が多く投稿される。このため、地震やゲリラ豪雨、感染症の動向など、多くのユーザから投稿が期待される対象に関し、予め設定した各対象に関する投稿に表れると考えられる単語などを用いて、マイクロブログからリアルタイムな情報を抽出する研究が進められていた。しかしこれらの研究による抽出対象は、社会的に関心が高く、多数のユーザに影響を与えるような対象に関する情報に限定され、マイクロブログに含まれる観測情報の多様性が十分に活用されていなかった。

２．研究の目的

本研究では、より個人特化型の位置情報サービスの実現を目指し、対象を限定せずに、実世界において観測される多様な対象に関するリアルタイムな観測情報をマイクロブログから自動的に収集した上で、個々のユーザの現在地・関心に応じて適切な情報を提示するシステムを構築することを目的とする。これを実現するため、マイクロブログへの不特定多数のユーザの投稿から実世界の多様な観測対象を逐次的に抽出する手法、及び、抽出された観測対象のうち、各ユーザの関心に応じたものに対する観測情報を抽出する手法を開発する。以降では、マイクロブログの各ユーザからの、実世界における観測情報に相当する投稿を実世界観測情報と呼ぶ。

３．研究の方法

それぞれの要素技術に対する具体的な検討内容は以下の通りである。

(1) 多様な観測対象の抽出

人々は実世界の特定の位置において何らかの対象を観測する。マイクロブログに対し、複数のユーザから同じ対象に関する観測情報が投稿された場合、これらの投稿には観測対象を表す単語など、同じ単語が用いられることが多いと予想される。よって、特定の位置においてのみ複数のユーザにより用いられる単語は、実世界の該当する位置に存在する観測対象を表し、これらの単語を含む投稿は、該当する対象に関する観測情報と考えられる。つまり、観測対象を表す単語が、実世界観測対象の抽出において重要な役割を果たす。

マイクロブログへの投稿には、投稿位置の緯度・経度を表すジオタグと呼ばれる位置情報を付与できる。よってジオタグ付き投稿を大量に収集し、地理空間において局所性の高い単語を抽出することにより、多様な観測対象を表す単語を抽出可能と考えられる。本研究ではこのような単語をローカル語と呼ぶ。

ここで、マイクロブログにおいては、観測対象の人気度によって観測情報の投稿頻度が異なること、また、観測対象には常にその位置で観測される定常的なものと、一時的に観測される非定常的なものが存在するといった特性から、これらの観測対象を表すローカル語を精度よく抽出するために適切な時区間のジオタグ付き投稿を収集する必要がある。そこで本研究では、このような各単語の使用特性の違いを考慮した上で、常に最新のローカル語をできるだけ多く抽出する方法について検討する。

(2) ユーザの関心に応じた実世界観測情報の抽出

(1)により抽出されたローカル語により表される観測対象のうち、特定のユーザの関心に応じたものに関する他のユーザの観測情報を提示することを考える。抽出対象となる観測情報はテキスト形式の投稿であるため、特定のユーザの関心を表すクエリもテキストで与えられるものとする。ここで、ユーザの関心に応じた投稿に、クエリと同じ単語が含まれるとは限らない。また、クエリと同じ単語が含まれる投稿も、必ずしも関連するとは限らない。ユーザの関心に応じた実世界観測情報を精度よく抽出するためには、クエリと投稿の意味的な関連性を評価する必要がある。そこで、クエリや投稿に用いられる複数の単語間の意味的な関連性を考慮しながら、ユーザの関心に応じた実世界観測情報を抽出する手法について検討する。具体的な検討事項については以下の通りである。

① ローカル語に関連する実世界観測情報の抽出

(1)により、実世界において観測される様々な対象を表すローカル語が抽出され、各ローカル語が示す位置周辺からのローカル語を含む投稿が各対象に関する観測情報として抽出できる。しかし、各対象に関連する観測情報には必ずしもローカル語が含まれるとは限らない。例えば、特定の空港名がローカル語として抽出されたとき、飛行機の遅延やセキュリティチェックの行列などに関する観測情報のすべてに空港名が含まれるとは限らない。そこで、ユーザが特定のローカル語を選択した際、関連する投稿がより多く提示されるよう、ローカル語を含まないが意味的に関連する投稿を、各ローカル語が表す対象の観測情報として抽出する方法につ

いて検討する。

② ユーザの関心に応じた観測対象の選択

(1)により、レストラン、空港、道路、学校、スタジアム、コンサートホール、祭りなど多様な観測対象が逐次的に抽出される。これらのうち、例えばスポーツの好きなユーザには周辺のスタジアムに関する観測情報を提示するなど、ユーザの関心に応じた観測対象の情報を提示することが望ましい。そこで各ユーザの関心に対するローカル語の意味的な関連度に基づき、適切な観測対象を選択する方法について検討する。

4．研究成果

(1)多様な観測対象の抽出[発表③⑤⑦⑨]

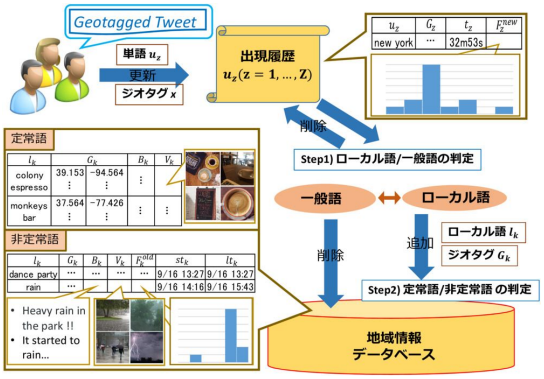

図１：実世界観測情報の抽出手法の概要

本研究で提案した実世界観測情報の抽出手法を図１に示す。提案手法は、実世界において観測された対象を表すローカル語を、観測された位置、観測情報と共に蓄積する地域情報データベースを構築することを目的とする。提案手法のアイデアは主に以下の３点である。まず、単語ごとに出現履歴を蓄積し、ジオタグ付き投稿を受信するごとに、投稿に含まれる単語の出現履歴を更新すると共に、地理空間的局所性を解析する。つまり、各単語に対して、ローカル語と判定されるまで出現履歴を蓄積することにより、観測情報の投稿頻度の異なる対象に対してそれぞれ適切な時区間で局所性を判定することができる。次に、局所性の高い単語をローカル語と判定すると共に、局所性が低い単語の判定も行う。これは、一時的に観測される非定常的な観測対象を精度よく抽出するためである。非定常的な観測対象を表す単語の出現履歴を常に蓄積すると、過去の出現履歴の影響で、一時的な局所性の判定が困難となり得る。そこで、局所性がある程度低くなった時点で一度出現履歴を削除することにより、常に最近の出現履歴のみを蓄積する。最後に、各単語の出現履歴は、あらかじめ実世界を分割し得られた領域ごとの使用頻度とする。これにより、出現履歴の更新時の処理は、使用された領域における使用頻度を 1 回増やすのみとなる。また、全領域中の最大使用頻度、及び使用された領域数のみに基づき局所性の解析が可能となり、リアルタイムな解析が実現できる。

さらに、過去にローカル語と判定された単語について、再度ローカル語と判定された際は位置情報の更新、局所性が低く一般語と判定された際は過去の一時的な観測対象として削除するなどの処理も行うことにより、常に最新の観測対象に関する情報を保持する地域情報データベースが構築される。このように逐次的に更新されるデータベース中の各ローカル語及びその位置情報に基づき、ローカル語が表す多様な対象の観測情報となる投稿が収集できる。

アメリカから発信された 1 ヶ月分の Twitter へのジオタグ付き投稿約 650 万件から、使用頻度が少ない場所名や、特産品など場所名以外の地域特有のものを表す語、最新のイベントを表す語などを含め、1 日平均約 1,300 語のローカル語が抽出された。30 日後までに抽出された約 40,000 語のうち約 50%は、既存の地理情報データベース GeoNames（http://www.geonames.org/）に含まれない単語であることを確認した。さらに、ローカル語の抽出に用いなかった同期間の投稿の位置推定を行った結果、GeoNames に含まれる単語と位置情報を用いた場合、誤差 5km 以内の投稿が 13%、50km 以内の投稿が 19%であったのに対し、提案手法により抽出したローカル語と位置情報を用いた場合、誤差 5km 以内の投稿が 53%、50km 以内の投稿が 75%となった。GeoNames には場所名以外としても使用されるような単語が多く含まれる一方、提案手法では局所性が高いものに限定されるが、既存の地理情報データベースに含まれない観測対象を表す語を多数抽出できる。表１に各データベース(DB)に含まれる単語の例を示す。位置推定の実験により、提案手法により得られるローカル語が、ジオタグが付与されていなくても、該当する対象に関する観測情報と考えられる投稿の収集に利用できることを確認した。

表１：各データベースに含まれるローカル語の例

DB	単語
両方	abel stadium, babilon station, cliff beach, laguardia airport, national aquarium
GeoNames のみ	friend, clinton, accident, cool, rice
提案手法のみ	adega lounge, bacchanal wine, bearded pig, biscuit paint wall, luisiana comic con

(2) ユーザの関心に応じた実世界観測情報の抽出

① ローカル語に関連する実世界観測情報の抽出[論文①②発表②④⑥⑧]

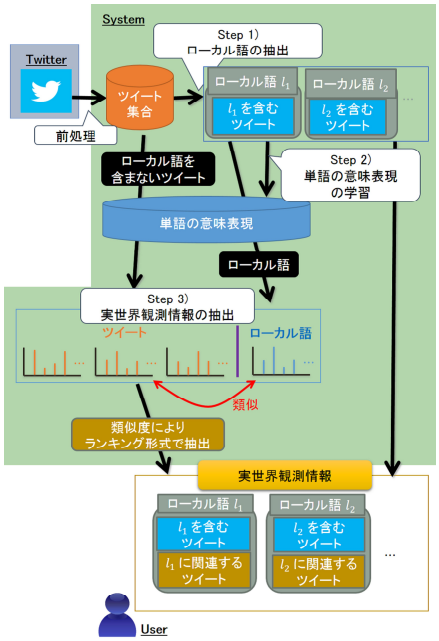

図 2：ローカル語に関連する
実世界観測情報の抽出手法の概要

図 2 に提案手法の概要を示す。1)において抽出されたローカル語のうち、ユーザが関心を持ったものを選択した際、ローカル語が表す対象に関連する観測情報を提示することを目的とする。これを実現するため、各ローカル語が示す位置周辺からの投稿に対し、ローカル語に対する意味的な関連度を評価する必要がある。ここで、各ローカル語が表す対象の観測情報は、ローカル語の有無によらず、その対象を特徴付ける単語を多く含むものと考えられる。例えば、空港における観測情報には空港名の有無によらず、“departure”、“boarding”、“gate”など空港から連想される単語が多く含まれる。しかし、1)において逐次抽出されるローカル語それぞれに対して、連想される単語を予め決定することは現実的ではない。

そこでまず、一定期間に抽出されたすべてのローカル語に対し、各ローカル語が示す位置周辺からのローカル語を含む投稿を、実世界の多様な対象の観測情報の事例として収集する。この中で、例えば、様々な空港からの投稿において共通して“departure”、“boarding”といった単語が用いられるなど、意味的に類似した対象に関する観測情報として、同じ単語が類似した文脈で用いられることが多い。これを踏まえ、多くの文書から、類似した文脈で用いられる単語間の距離が近くなるよう、単語のベクトル表現を学習する手法を、収集した観測情報の事例集合に適用する。これにより、意味的に類似した対象の観測情報によく用いられる単語は類似したベクトル表現を持つこととなる。

学習した単語ベクトルの加算平均などにより、複数の単語から構成されるテキストのベクトル表現を得ることができる。よって、各ローカル語に対して、ローカル語を含む最近の投稿、及び周辺から投稿されたローカル語を含まない投稿それぞれに含まれる単語に基づきベクトル表現を算出し、これらのベクトルの距離に基づき、ローカル語と周辺の投稿の意味的な関連度を評価する。

(1)で用いた 1 か月分のジオタグ付き投稿のうち、前半 25 日間分から抽出されたローカル語に対し、各ローカル語が示す位置周辺からのローカル語を含む投稿を観測情報の事例として用い、単語の意味表現を学習した。表 2 に、後半 5 日間の投稿を用いて、ローカル語に対し関連度の高い周辺の投稿を抽出した例を示す。空港名である“sfo”に対して、空港における観測情報、フットボール場名である“alumni stadium”に対して、その日開催された試合の観測情報、ライブイベント名“desert trip”に対して、出演ミュージションに関する観測情報など、それぞれローカル語は含まないが、意味的に関連する投稿が抽出されることが確認された。

表 2：ローカル語に対して抽出された
観測情報の例

ローカル語	観測情報
sfo	Arrived, at what I still think is America ’ s nicest airport terminal.
	And the “Mother of the Year” award goes to the new mom in seat 2A.
	Omg there ’ s a Yoga room in this terminal and my flight is delayed.
alumni stadium	So many #Clemson fans! (@ Boston College in Boston, MA)
	Attending Clemson vs Boston College tonight!!!!
desert trip	Watching the Rolling Stones sound check.
	Like a Rolling Stone is a great song.

② ユーザの関心に応じた観測対象の選択 [発表①]

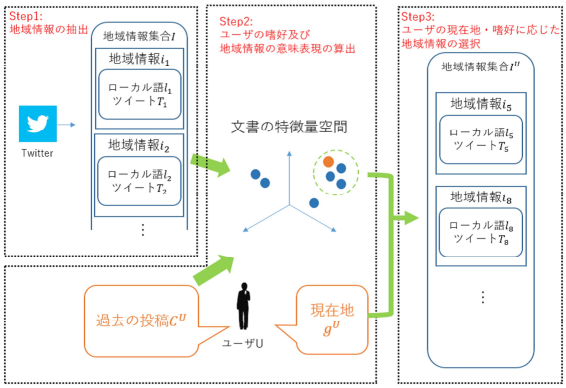

図 3：ユーザの関心に応じた
観測対象の選択手法の概要

図 3 に提案手法の概要図を示す。①において、(1)で抽出される各ローカル語が表す対象に対し、意味に基づくベクトル表現を得ることができる。一方、各ユーザは自身が興味を持った対象に関する観測情報を投稿すると考えられるため、過去の投稿にユーザの関心が表れると予想される。そこで、ユーザの過去の投稿に含まれる単語に基づき、①で学習した単語ベクトルを用いて、同様にユーザの関心のベクトル表現を得る。これにより、ユーザと抽出されたローカル語はそれぞれ位置情報と意味を表すベクトル表現を持つため、これらの距離に基づき、個々のユーザの現在地・関心に応じた観測対象を選択する。

①と同様に、1 か月分のジオタグ付き投稿のうち、後半 5 日間において、ローカル語を用いた投稿を行ったユーザに対し、それまでの投稿を用いてユーザの関心のベクトル表現を算出し、実際に観測情報を投稿した対象が正しく上位に選択されるか検証した。一例を表 3 に示す。ユーザの投稿及びローカル語が表す対象の観測情報に共通した単語が含まれなくとも、意味的な関連度が正しく算出されることが確認された。

表 3：ユーザの関心に応じた観測対象
“bbva compass stadium”が正しく選択された例

ユーザの過去の投稿	“bbva compass stadium”の過去の観測情報
NWSL Regular Season: September 25, 2016: Houston Dash vs Seattle Reign at BBVA	Got to see a Dynamo win with the bffs!!
At the @HoustonDynamo game with @paulwallbaby	Gonna see this game good! Arriba los Tigres!!

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 2 件）

① 新田直子，吉武真人，中村和晃，馬場口登：“マイクロブログからの関連実世界観測情報の抽出”，日本データベース学会和文論文誌，査読有，Vol.16-J，Article No.22，8 pages，2018.

② 新田直子，角谷直人，馬場口登：“単語間の関係性の経時変化を考慮したマイクロブログからの実世界観測情報の抽出”，日本データベース学会論文誌，査読有，Vol.13-J，No.1，pp.7-12，2014.

〔学会発表〕（計 9 件）

① 坂本宏祐，Lim Jeongwoo，新田直子，中村和晃，馬場口登：“マイクロブログを用いたリアルタイム地域情報の推薦”，第 10 回データ工学と情報マネジメントに関するフォーラム（DEIM Forum 2018），D1-2，7 pages，March 2018，福井県あらわ市.

② M. Yoshitake，N. Nitta，K. Nakamura，N. Babaguchi：“Extracting Real-World Observations from Microblog，” IEEE International Conference on Multimedia Big Data，pp.232-237，April 2017，California，USA.

③ T. Kamimura，N. Nitta，K. Nakamura，N. Babaguchi：“On-line Geospatial Term Extraction from Streaming Geotagged Tweets，” IEEE International Conference on Multimedia Big Data，pp.322-329，April 2017，California，USA.

④ 吉武真人，新田直子，中村和晃，馬場口登：“マイクロブログからの関連実世界観測情報の抽出”，第 9 回データ工学と情報マネジメントに関するフォーラム（DEIM Forum 2017），D3-5，8 pages，March 2017，岐阜県高山市，学生プレゼンテーション賞受賞.

⑤ 上村卓也，新田直子，中村和晃，馬場口登：“マイクロブログからのリアルタイム地域情報抽出”，第 9 回データ工学と情報マネジメントに関するフォーラム（DEIM Forum 2017），C7-1，March 2017，岐阜県高山市，学生プレゼンテーション賞受賞.

⑥ M. Yoshitake，N. Nitta，N. Babaguchi：“Real-World Observation Extraction from Microblog based on Word Associative Relations，” IEEE International Conference on Multimedia Big Data，pp. 450-455，April 2016，Taipei，Taiwan.

⑦ T. Kamimura，N. Nitta，N. Babaguchi：“Real-Time Local Word Database Construction from Twitter，” IEEE International Conference on Social Computing and Networking，pp. 299-306，December 2015，Chengdu，China.

⑧ 吉武真人，新田直子，馬場口登：“ユーザの関心に応じたマイクロブログからの実世界観測情報の抽出”，ARG Web インテリジェンスとインタラクション研究会，No.6，pp.25-30，June 2015，大阪府豊中市.

⑨ 上村卓也，新田直子，馬場口登：“時空間的出現特性の違いを考慮した位置を示す語の抽出によるツイートの発信位置推定”，DEIM Forum 2015，C4-2，8 pages，March 2015，福島県郡山市，学生プレゼンテーション賞受賞.

６．研究組織

(1)研究代表者

新田　直子（NITTA，Naoko）

大阪大学・大学院工学研究科・准教授

研究者番号：00379132