人工知能学会研究会資料
SIG-KBS-B401-08

位置情報を考慮した非日常ツイートの抽出の試み

A Fundamental Attempt to Extract Unusual Geo-Tagged Tweets

鈴木陽介　尾崎知伸 *
Yosuke Suzuki　Tomonobu Ozaki

日本大学 文理学部
College of Humanities and Sciences, Nihon University

Abstract: Twitter has been recognized as a popular communication tool. In this paper, we report a fundamental attempt to extract personal tweets on unusual and uncommon events. Because unusual tweets represent exceptional situations and/or strongly desired ones, by using such tweets, we can expect to build a sophisticated user profile used in further information services such as recommendation systems. Among several aspects on uncommonness of tweets, we focus on contents in tweets as well as locations where the tweets are posted, and prepare several evaluation criteria of unusualness of tweets. A preliminary experiment using small real datasets is conducted to assess the feasibility of the proposed framework.

1　はじめに

Twitter[1] とは，最大 140 文字の記事（ツイート・つぶやき）を投稿・閲覧するコミュニケーションサービスである．その簡易さから，リアルタイム性の高い情報交換ツールとして幅広く活用されるとともに，実世界におけるソーシャルセンサとしての役割も期待されている [1].

多くの利用者にとって，Twitter は，有力な情報収集ツールとしての側面を持つ一方，自身の身の回りで起きた出来事やそれに関する感想などを投稿する，ある種の日記やライフログとしての役割も担っている．本研究では，各ユーザが投稿したツイート群から，旅行中のツイートや通常とは違う行動，普段は遭遇しない出来事に対するツイートなど，非日常的なイベントに対するツイート，すなわち非日常ツイートを抽出することを考える．

非日常ツイートを抽出する一つの目的として，高精細なユーザプロファイルの作成とそれを利用した推薦の実現があげられる．近年，膨大な商品群から利用者の嗜好に合わせた商品を提案する推薦システム [2, 3] が注目を集めているが，推薦を実現するための基礎データとしてユーザプロファイルを用いる場合も少なくない．ツイート群を用いてユーザプロファイルを構築する際，ある意味で例外的なツイートである非日常ツイートを除外することで，より高精細なプロファイルの構築が期待できる．一方，旅行中のツイートなど，利用者が強く望むことで実現された非日常的な出来事に対するツイートには，より強く利用者の嗜好が現れるとも考えられ，その様な非日常ツイートを積極的に利用することで，ジャンルに特化したプロファイルの獲得も期待できる．さらに，日常的なツイートと非日常ツイートとのギャップを考慮することで，意外性のある推薦につながる可能性があると考えられる．

一言に，非日常的なイベントと言っても，様々な状況が考えられる．大きくは，(1) 観光旅行や特別な食事，大きな買い物など，本人が望むことで引き起こされるイベントと，(2) 事件や事故，急な病気など，本人が望まないイベントに分けることができる．また非日常的なイベントの検出には，場所や時間，内容的な非日常性が重要な役割を果たす．例えば，多くの利用者にとって，観光旅行は非日常であることに疑いはないが，同じ旅行でも出張（仕事のための旅行）はどうであろうか．行動範囲の点では，観光旅行も出張も普段の主たる行動範囲から離れていることが予想されるので，非日常的と考えることもできる．その一方で，出張は，(移動そのものはともかく）仕事という面で普段の行動と大きな差はなく，必ずしも非日常とは言い切れない．特に，日常的に出張などの移動の多い仕事に従事している利用者にとっては，場所の違いは，非日常性を決定するのに必ずしも十分な情報ではない．また旅行とは逆に，ツイートの投稿場所自体は日常的な行動範囲内であったとしても，事件や事故などのイベントに対するツイートは，その内容や頻度から考えても，非日常的であると考えられる．さらに，投稿場所や内

*連絡先：日本大学 文理学部 情報科学科
〒 156-8550 東京都世田谷区桜上水 3-25-40
tozaki@chs.nihon-u.ac.jp

[1] http://twitter.com

容が他のツイートと大差がなくとも，普段の投稿時間とは大幅に異なる時間に行われたツイートには，何らかの非日常的な意味が内包されていると推測される．

以上を簡単に整理すると，望む・望まれざるにかかわらず，ツイートの非日常性には少なくとも (1) 場所的な観点からの非日常性，(2) 時間的な観点からの非日常性，(3) 内容的な観点からの非日常性が存在し，それらを複合的に評価することで，実際のツイートの非日常性が決まると考えられる．

これらのことを背景に，本研究では，非日常ツイートを抽出する初期的な試みとして，場所及び内容的な観点からの非日常性に着目し．(1) 主たる行動範囲とは離れた場所から投稿された，(2) 普段の投稿とは異なる内容を含むツイートを非日常ツイートとして抽出することを試みる．

以下に本論文の構成を示す．2 章で，非日常ツイート抽出の枠組みを示すとともに，投稿場所と投稿内容を考慮したツイートに対する種々の非日常性評価関数を提案する．次いで 3 章で，実ツイートデータを用いた評価実験とその考察を行う．最後に 4 章で，まとめと今後の課題を述べる．

2 非日常ツイートの抽出

2.1 提案手法の概要

本節では，投稿位置と投稿内容に基づく非日常ツイートの抽出手法を提案する．提案手法では，以下の手順に従い，非日常ツイートの抽出を行う（図 1 参照）．

(1) ツイートの獲得： 利用者 u に対し，(一定期間の) 全ツイートの集合 $T^u = \{t_1^u, t_2^u, \cdots, t_{|T^u|}^u\}$ を獲得する．ここで $t_i^u \in T^u$ は，u により投稿されたツイートを表す．

(2) 投稿位置情報の獲得： 各ツイート $t_i^u \in T^u$ に対し，その投稿位置（緯度経度）情報 $g(t_i^u)$ を獲得する．ジオタグが付与されている場合はその情報を利用するが，ジオタグが付与されていない場合は，投稿位置推定アルゴリズム（[4, 5] など）を利用する．

(3) テキスト情報の獲得： 各ツイート $t_i^u \in T^u$ の本文に形態素解析を適用し，本文に含まれる名詞，形容詞，副詞の集合 $c(t_i^u)$ を獲得する．

(4) 非日常性評価： 非日常性評価関数 f を利用し，投稿位置 $g(t_i^u)$ と投稿内容 $c(t_i^u)$ から，各ツイートの非日常性 $f(t_i^u)$ を算出する．なお，具体的な評価関数に関しては後述する．

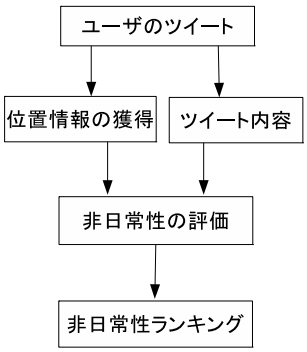

図 1: 提案手法の概要

(5) 非日常ツイートの抽出： 高い非日常性 $f(t_i^u)$ を持つツイート t_i^u を，利用者 u に対する非日常ツイートとして抽出する．具体的には，評価値 f によるランキング

$$rank_f(t_i^u) = \left|\{\, t_j^u \in T^u \,|\, f(t_j^u) > f(t_i^u) \,\}\right| + 1$$

の上位 k 件 $\{\, t_i^u \in T^u \,|\, rank_f(t_i^u) \leq k \,\}$ を，非日常ツイートの集合として獲得する．

2.2 非日常性の評価関数

本節では，まず，投稿位置及び投稿内容のそれぞれの観点に従った非日常性評価関数を導入する．次いでそれらを組み合わせることで，両者を考慮した非日常性の評価を実現する．

2.2.1 投稿位置に関する非日常性の評価

利用者が，日常的な生活の中でツイートを行っていると仮定すると，投稿数が多いエリアは，その利用者にとって日常的な生活圏である可能性が高い．この場合，日常的にツイートが行われる位置から離れた場所から投稿されたツイートは非日常的であると考えられる．この考えに従い，利用者 u によるツイート t_i^u の投稿位置に関する非日常性 $f_g(t_i^u)$ を，t_i^u と他のツイートの投稿位置との距離の総和と定義する．以下に，形式的な定義を示す．

$$f_g(t_i^u) = \sum_{t_j^u \in T^u, i \neq j} dist_g(\, g(t_i^u), g(t_j^u) \,)$$

ここで $dist_g(\, g(t_i^u), g(t_j^u) \,)$ は，t_i^u と t_j^u の投稿位置間の距離を表し，緯度経度情報から計算される．

2.2.2 投稿内容に関する非日常性の評価

あるツイートに対して，同じような内容のツイートが繰り返し投稿されている場合，そのツイートの対象となった状況やイベントは，普段よく起きている日常的なものであると考えられる．逆に言えば，あるツイートに対して，同じような内容の投稿がない場合，そのツイートは非日常的であると考えられる．この考えに基づき，ツイート t_i^u の投稿内容に関する非日常性 $f_c(t_i^u)$ を，t_i^u と他のツイートのテキスト間距離の総和と定義する．以下に，形式的な定義を示す．

$$f_c(t_i^u) = \sum_{t_j^u \in T^u, i \neq j} dist_c(\, c(t_i^u),\, c(t_j^u)\,)$$

ここで $dist_c(\, c(t_i^u),\, c(t_j^u)\,)$ は，ツイート本文間の Jaccard 距離であり，t_i^u，t_j^u のそれぞれに含まれる名詞，形容詞，副詞の集合 $c(t_i^u)$ と $c(t_j^u)$ を用いて，

$$dist_c(\, c(t_i^u),\, c(t_j^u)\,) = 1 - \frac{\left|c(t_i^u) \cap c(t_j^u)\right|}{\left|c(t_i^u) \cup c(t_j^u)\right|}$$

と定義される．

2.2.3 投稿位置と投稿内容の併用による非日常性の評価

先述した “出張” の例のように，投稿位置と投稿内容の両方が非日常的であることを要請する場合も考えられる．このことに対応するため，投稿位置及び投稿内容のそれぞれで非日常性のランキングを考え，それらを統合することで，総合的な非日常性を評価することを考える．具体的には，各ランキングの逆数を取り，その積を非日常性と定義する．以下に，投稿位置及び投稿内容を併用したツイート t_i^u の非日常性 $f_{gc}(t_i^u)$ の形式的な定義を示す．

$$f_{gc}(t_i^u) = \frac{1}{rank_{f_g}(t_i^u) \times rank_{f_c}(t_i^u)}$$

2.2.4 投稿位置によるフィルタリングを伴う投稿内容に関する非日常性の評価

投稿位置と投稿内容の両方が非日常的であることを要請する場合の亜種として，ツイートの投稿位置が，日常的な投稿位置より一定距離以上離れていることを前提とすることを考える．またその上で，投稿内容が非日常的であることを要請する．言い換えれば，投稿位置により非日常ツイートを絞り込み，その上で，投稿内容によりランキングを行うということである．この考えに従ったツイート t_i^u の非日常性 $f_g^c(t_i^u)$ を以下のように定義する．

$$f_g^c(t_i^u) = \begin{cases} \min\limits_{t_j^u \in U_u} dist_c(t_j^u, t_i^u) & dist_g(t_\mu^u, t_i^u) \geq d_2 \\ -\infty & \text{otherwise} \end{cases}$$

where

$$U_u = \{\, t_j^u \in T^u \,|\, dist_g(t_\mu^u, t_j^u) \leq d_1 \,\} \text{ and}$$

$$t_\mu^u = \operatorname*{argmin}_{t_j^u \in T^u} f_g(t_j^u)$$

評価関数 f_g^c は，投稿位置に関して最も日常的なツイート t_μ^u から，閾値 d_2 以上離れている位置から投稿されたツイート t_i^u のみを非日常ツイートと認識し評価値を与える．また，非日常性を評価する際，t_μ^u と投稿位置の近い（具体的には閾値 d_1 以下の）ツイート群 U_u に含まれるツイート f_j^u との内容的な距離 $dist_c(t_j^u, t_i^u)$ の最小値を採用している．

2.3 コーパスを利用した非日常性の評価

前節では，利用者 u 自身が投稿したツイートを利用した非日常性評価のための関数を提案した．これに対し本節では，「非日常」とその類義語である「非現実」や「普通でない」といった語を含むツイートを非日常性を含むツイートとして収集し，コーパスとして準備することで，ツイートの非日常性を評価することを考える．以下に，コーパス C を利用した非日常性の評価関数 f_C^k の形式的な定義を示す．

$$f_C^k(t_i^u) = \frac{1}{|T_C^k(t_i^u)|} \sum_{t_j \in T_C^k(t_i^u)} 1 - dist_c(t_j, t_i^u)$$

where

$$T_C^k(t_i^u) = \{t_j \in C \,|\, rank_c(t_i^u, t_j, C) \leq k\} \;\; and$$

$$rank_c(t_i^u, t_j, C) = |\{t_k \in C \,|\, dist_c(t_k, t_i^u) < dist_c(t_j, t_i^u)\}| + 1$$

この評価関数 $f_C^k(t_i^u)$ は，コーパス C 中に含まれるツイート t_j のうち，t_i^u と投稿内容の近いもの上位 k 件を考え，その類似性（$= 1 -$投稿内容間の Jaccard 距離）の平均値を採用している．

3 評価実験と考察

3.1 データの準備

提案手法を評価するため，Twitter API[2] の Java ラッパである Twitter4J[3] を利用し，日常的にジオタグ付き

[2] https://dev.twitter.com/docs
[3] http://twitter4j.org

ツイートを行っている 3 名の利用者（利用者 A〜C）を対象に，2014 年 1 月 1 日から 2014 年 6 月 23 日の期間のツイートを収集した．収集されたツイート数はそれぞれ，利用者 A : 327 ツイート，利用者 B : 1717 ツイート，利用者 C : 2116 ツイートである．また，投稿本文に対する形態素解析には，Mecab[4] を利用した．

実験では，各利用者毎に，評価関数 f_g^c における投稿位置間の距離に関する二つの閾値 d_1 と d_2 を設定した．具体的には，各ツイート t_i^u と t_μ^u との距離 $dist_g(t_\mu^u, t_i^u)$ の平均を μ，標準偏差を σ とし，$d_2 = \mu+2\sigma$，$d_1 = \mu+\sigma$ を採用している．一方，約 1300 のツイートを含む非日常コーパス C を利用した評価関数 f_C^k に対しては，$k = 3$ を採用している．

3.2 評価実験 1 : 評価関数の比較

提案した評価関数間の関連性を確認するため，各評価関数で得られるランキングに対してケンドールの順位相関を求め，考察を行った．

n 個の要素を含むツイート群 T に対する 2 つのランキング $rank_a$ と $rank_b$ のケンドールの順位相関は，

$$kendall(rank_a, rank_b) = \frac{4 \times P(rank_a, rank_b)}{n \times (n-1)} - 1$$

と定義される．ここで $P(rank_a, rank_b)$ は，2 つのランキング間で順位関係が一致するツイート対 $t_i, t_j \in T$ の数を表す．順位相関は，$-1 \sim +1$ の値を取り，値が大きいほど正の相関が，小さいほど負の相関があることを表す．また値 0 は無相関を表す．

実験結果を表 1 に示す．なお，投稿位置によるフィルタリングを伴う評価関数 f_g^c に関しては，フィルタリングされずにランキングの対象となったツイート，すなわち $dist_g(t_\mu, t_i) \geq d_2$ を満たすツイート t_i のみを対象に，投稿位置による評価関数 f_g との相関のみを計算した．

実験結果より，投稿場所と投稿内容の双方を考慮した非日常性 f_{gc} は，それぞれに基づく非日常性である f_g と f_c と強い正の相関を持つことが分かる．しかしこの結果は，評価関数 f_{gc} の構成に起因するものであり，当然の結果であると考えられる．

一方，利用者 B と利用者 C において，投稿位置による非日常性 (f_g) と投稿内容による非日常性 (f_c) が無相関であることが分かる．このことは，投稿場所もしくは投稿内容だけでは，両者を考慮した非日常性を検出することができないことを表しており，両者を組み合わせる意義を支持するものである．同様に，投稿位置による非日常性 f_g と投稿位置によるフィルタリングを伴う投稿内容による非日常性 f_g^c の間にも，大きな相関

[4] https://code.google.com/p/mecab/

表 2: 各評価関数に基づく非日常ツイートのランキング

	f_g	f_c	f_{gc}	f_g^c	f_C^k
t_1^A	5	77	6	8	295
t_2^A	36	12	12	–	271
t_3^A	43	68	46	–	289
t_4^A	147	88	125	–	234
t_5^A	231	131	202	–	154
t_6^A	225	18	57	–	217

がないことが分かる．このことも，投稿位置と投稿内容の両者を使うことの意義を示すものであると考えられる．

ところで，コーパスに基づく非日常性である f_C^k との関連性に関しては，利用者 A においては負の相関が認められ，また利用者 B においては無相関という結果となった．この結果は，利用者毎に投稿場所と投稿内容を考慮することで，コーパスを利用する場合には得られない結果が導出されることを表していると考えられる．

3.3 評価実験 2 : 非日常ツイートの抽出

提案手法を用い，実際に非日常的なツイートの抽出が可能かを確認するため，利用者 A のツイートを対象に，評価実験を行った．実験では，利用者 A の全ツイートを精査した上で，手作業により 6 件の非日常ツイート ($t_1^A \sim t_6^A$) を抽出し，それらが各評価関数において，どの程度の順位にランキングされるかを確認した．なお，$t_1^A \sim t_3^A$ は，日常的な行動範囲とは離れた場所から投稿された旅行中のツイートである．一方，t_4^A と t_5^A は，日常的な行動範囲の中で投稿されたツイートであるが，日常的ではない仕事に関するツイートである．また t_6^A は，日常的な行動範囲の中で投稿された，非日常的なイベントであるお祭りに関するツイートである．実験結果を表 2 に示す．なお，表中の f_g^c 列における '–' は，投稿位置によりフィルタリングされてしまったことを表す．

実験結果より，いずれの評価関数においても t_1^A 以外は上位にランキングされておらず，投稿位置や内容に加え，更なる観点からの非日常性へのアプローチの必要性が示された．

評価関数間の比較としては，ツイート t_1^A と t_3^A に関しては，投稿内容のみに基づく f_c よりも，投稿内容と投稿位置を考慮した f_{gc} において順位が良くなっていることが分かる．また，$t_4^A \sim t_6^A$ に関しては，元々日常

表 1: 評価関数毎の順位相関

	利用者 A				利用者 B				利用者 C			
	f_c	f_{gc}	f_C^k	f_g^c	f_c	f_{gc}	f_C^k	f_g^c	f_c	f_{gc}	f_C^k	f_g^c
f_g	0.29	0.63	-0.17	-0.12	0.00	0.51	-0.05	0.16	0.07	0.54	0.40	0.04
f_c		0.66	-0.13			0.50	0.05			0.53	0.02	
f_{gc}			-0.18				-0.02				0.25	

の行動範囲内で投稿されたツイートということもあり，投稿位置のみに基づく f_g よりも，f_{gc} において順位の向上が確認できる．しかし，投稿位置 f_g と投稿内容 f_c を個別に考慮した場合のランキングに対して，f_{gc} によるランキングが必ずしも向上しているとは限らず，投稿位置と内容をどのように組み合わせるかという点において課題が残る結果となった．一方，コーパスを利用した場合 (f_C^k) と比較してランキングの大幅な改善が認められ，利用者毎に非日常性を考慮する点や投稿位置を考慮する点は，非日常ツイートの抽出においてそれなりの貢献があると考えている．

4 まとめと今後の課題

本研究では，非日常ツイートを抽出する初期的な試みとして，ツイートの投稿位置と投稿内容に基づく種々の非日常性評価関数を提案した．また，小規模な初期実験ながら，投稿位置情報もしくは投稿内容だけでは抽出することのできない非日常ツイートの抽出可能性を確認した．

本研究は，非日常ツイートの抽出に関して，非常に基礎的な考察を行ったに過ぎず，多くの課題を残している．具体的には，ツイートの投稿時間を考慮したプロファイルの作成 [6] など，異なる観点からの非日常性の導入や，投稿内容のより積極的な利用のための意見分析技術 [7] の適用などがあげられる．また，非日常ツイートを用いたプロファイル構築を行い，高精度かつ意外性のある推薦を実現することも大きな課題である．

参考文献

[1] 榊 剛史，松尾 豊 : ソーシャルセンサとしての Twitter : ソーシャルセンサは物理センサを凌駕するか?，人工知能学会誌，Vol.27, No.1, pp.67–74 (2012)

[2] 土方嘉徳 : 嗜好抽出と情報推薦技術，情報処理，Vol.48, No.9, pp.957-965 (2007)

[3] 神嶌敏弘 : 推薦システムのアルゴリズム (1)，人工知能学会誌，Vol.22, No.6, pp.826–837 (2007)

[4] 杉谷 卓哉，白川 真澄，原 隆浩，西尾 章治郎 : 教師あり機械学習を用いたツイート投稿時のユーザ位置推定手法，情報処理学会研究報告，データベース・システム研究会報告，2013-DBS-158(26), pp.1-8 (2013)

[5] Y. Ikawa, M. Enoki and M. Tatsubori : Location Inference Using Microblog Messages, *proc.* of the 21st International Conference Companion on World Wide Web, pp.687–690 (2012)

[6] 今井 規善，奥 健太，服部 文夫 : 位置情報クラスタリングに基づく地理的ユーザプロファイリング手法，情報処理学会第 75 回全国大会講演論文集，Vol.2013，No.1, pp.651-653 (2013)

[7] 関 洋平 : 意見分析コーパスに関する現状調査，情報処理学会研究報告，情報学基礎研究会報告，2012-IFAT-108(2)，pp. 1–8 (2012)