

Twitter からのイベント検出および関連情報収集の高精度化 Improvement of Local Event Detection with Related Information from Tweets

中澤 昌美[†] 池田 和史[†] 服部 元[†] 小野 智弘[†]
Masami Nakazawa Kazushi Ikeda Gen Hattori and Chihiro Ono

1. はじめに

近年 SNS(Social Networking Service)やモバイル端末の普及により、端末に内蔵された GPS を利用して位置情報を付与した一般ユーザによるコメント投稿が増加してきた。Foursquare[1]や Localmind[2]のような、位置情報を活用して興味のある場所に関する付加価値情報を得るサービスに対する需要は大きくなっている。

筆者らはこれまでに、位置情報付きツイートからイベント発生地点を検出し、特徴的な建物名(ランドマーク名)を基に位置情報が付与されていないイベント関連ツイートを収集することにより、その場で起きているイベント内容を推定する手法を提案した[3]。この手法により、イベント発見と関連情報の自動収集が可能となったが、必ずしも関連しない情報が多く含まれるという課題があった。そこで、本稿では、イベント地点、イベント名、イベント関連情報をそれぞれ高精度に推定・収集する手法を提案し、実データにより有効性を検証する。

2. 課題と対処方針

本稿では、特定エリアにおいて共通の話題で盛り上がっている出来事(お祭り、花火、ライブなど)を「イベント」、その場所を「イベント地点」と定義する。筆者らのこれまでの手法[3]では、①実際にはイベントが発生していない地点が数多く検出される点、②ランドマーク名の推定精度が低い点、③イベントと無関係な関連ツイートが収集される点が課題であった。課題①について、従来手法[3]では、大型ショッピング施設や駅のような、多数の人が集まっているが、共通な話題が無く人々が様々な行動をとる場所がイベント候補地として検出されてしまう。そこで本提案手法では、このような地点を削除するため、該地点に特徴的な単語(=共通の話題)を含むツイートを投稿した人数が多い地点をイベントが発生している地点とする。課題②について、従来手法では、イベント地点と検出した地点の経緯度を基にローカルサーチ API でランドマークを決定していたため、店舗名がランドマーク名と推定されることが多く、その推定精度は約 20%と低い。そこで本提案手法では、推定されたイベント地点および周辺のランドマーク名を基にツイートを収集し、収集したツイートの類似度により、ランドマーク名を高精度に推定する。課題③について、従来手法ではイベントと無関係なツイートも含まれてしまう。特に@付きのツイートは、現地の人と離れた人同士のやり取りが多く、イベント内容と無関係な場合が多い。そこで本提案手法では、収集したツイートから@付きのツイートを削除するとともに、該地点に特徴的に表れる単語を共通の話題として抽出し、それらの単語を含むツイートのみを抽出する。



図 1. 提案手法の処理フロー

3. 提案手法

先に述べた対処方針に従い、イベント地点を高精度に検出し、さらにそのイベントに関連するツイートを効率的に収集する手法を提案する。本手法は、(Step1) イベント候補地検出、(Step2) イベント地点絞り込み、(Step3) ランドマーク名推定、(Step4) イベント関連ツイート抽出の 4 つのステップを実行する(図 1)。上記課題①を Step2、課題②を Step3、課題③を Step4 において解決する。以降では、提案手法の各ステップの詳細を示す。

3.1 Step1 : イベント候補地検出

位置情報付きツイートを収集し、その密集地点をクラスタリングしてイベント候補地を検出する。従来手法[3]Step1 と同様にクラスタリングには DBSCAN[4]を用いる。

3.2 Step2 : イベント地点の絞り込み

Step1 で検出したイベント候補地から、実際にイベントが発生していない地点のみを抽出する。各イベント候補地について以下の手順で判定する。

①イベント候補地ツイートから TF-IDF により特徴的なキーワードを抽出し、キーワードリストを作成する。TF はイベント中その地点から投稿されたツイートの単語の出現回数、DF は新聞コーパスに基づいて算出する。②位置登録サービスのチェックイン時によく利用される、駅名、地名、ランドマーク名、「タッチ」をキーワードリストから除く。③キーワードリスト内の単語を含むつぶやきを投稿した人数を数える④一定割合以上の人がつぶやいたキーワードがある場合、推定イベント地点として採用する。

3.3 Step3 : ランドマーク名の推定

Step2 で特定したイベント地点のランドマーク名を推定する。事前に、ランドマーク DB を作成する。Wikipedia において表 1 のカテゴリ名で登録されている名称と経緯度を抽出し、ランドマーク名 DB に登録することで作成する。ただし Wikipedia 上に経緯度の記載がない場合、GoogleMaps のジオコーディング API を利用して経緯度を登録する。以下に、ランドマークの選択手法を説明する。①Step2 で検出した推定イベント地点から約 200m 以内に存在するランド

[†] KDDI 研究所 KDDI R&D Laboratories Inc.

表 1. ランドマーク名 DB 作成に利用した Wikipedia のカテゴリ

建築物, 博物館, 美術館, 科学館, 図書館, ホール概要, 体育館, スタジアム情報ボックス, 動物園, 植物園, 水族館, 行政施設, 商業施設, 超高層ビル, テーマパーク, 大学, 駅情報, 公園, 野球場情報

表 2. 各キーワードをつぶやいた投稿者数 (東京ドームの例)

		キーワードリスト		
		嵐	クリア ファイル	ラクーア
推定イベント地点		4	1	1
周辺 ランド マーク 名	後樂園駅	3	0	0
	野球体育 博物館	0	0	0
	文京区役所	0	0	0
	東京ドーム	129	1	7

マーク名を周辺ランドマーク名として DB から抽出する。

②周辺ランドマーク名をキーとして Step1 と同時時間帯のツイートを収集する。③Step2 で作成したキーワードリストの単語をつぶやいたユニーク投稿者数をカウントする。④推定イベント地点と同じ単語をつぶやいたユニーク投稿者数が最も多い地点をランドマーク名と推定する。例えば、東京ドームでライブが開催された際、キーワードリストの単語を含むツイート投稿者数は表 2 のようになった。推定イベント地点と同じ単語を含むツイート投稿者が最も多い東京ドームを当該地点のランドマーク名と推定する。

3.4 Step4 : イベント関連ツイート抽出

推定イベント地点に関連するツイートを収集するため、Step3②で収集した、推定ランドマーク名を含むツイートから@付きツイートを削除し、キーワードリストに含まれる単語を 1 語以上含むツイートを抽出する。

4. 実験

提案手法の有効性を検証するため、評価実験を行った。2012 年 6 月 16 日 12:30~14:30 に投稿されたツイートを対象とし、提案手法により関東圏のイベントを検出した。DBSCAN により検出した 39 のイベント候補地に対し、人手によるラベリングにより、実際にイベントが発生した 5 地点と、イベントが発生していなかった 34 地点に分類した。また、5 地点について、正解ツイートリストを手で作成した。Step2 における一定割合は実験的に 4 割と設定した。結果を以下に述べる。

4.1 イベント地点の絞り込み結果

イベント候補地と、絞り込みを行った推定イベント地点を図 2 に示す。今回新たに追加した Step2 で絞り込んだ 10 地点に実際にイベントが発生した 5 地点すべてが含まれることを確認した。従来手法より高精度に推定イベント地点が検出できたと言える。

4.2 ランドマーク名推定結果

正解イベント地点 5 地点におけるランドマーク名推定精度を確認した。従来手法では、5 地点中正解は 1 地点であるが、提案手法では正解が 4 地点となり、ランドマーク名の推定が高度化できた。不正解の 1 件は、「東京スカイツリー」を「東京ソラマチ」と推定した。近距離内に関連施設が存在する場合、キーワードによるランドマーク名の推定が困難であった。このような場合の対処法を新たに考え



図 2. イベント候補地と推定イベント地点

表 3. 追加収集したツイートの評価

ランドマーク名	従来手法		提案手法	
	収集ツイート数	正解数とその割合	収集ツイート数	正解数とその割合
西武ドーム	743	660 (88.8%)	489	446 (91.2%)
千葉マリンスタジアム	44	35 (79.5%)	26	21 (80.8%)
東京ビッグサイト	33	27 (81.8%)	24	23 (95.8%)
東京ディズニーシー	17	13 (76.5%)	15	11 (73.3%)
4 地点合計	837	735 (87.8%)	554	501 (90.4%)

る必要がある。

4.3 推定イベント関連ツイート抽出結果

Step4 で抽出したツイートが推定イベント地点の関連ツイートであるか評価する。正解した 4 地点において従来手法および提案手法で収集したツイート数とその正解数を表 3 に示す。東京ディズニーシーを除く 3 地点において関連ツイート収集が効率化できた。提案手法により削除された正解データには短文の感想ツイートが多い。これはキーワードリストの単語を含まないためであるが、推定イベント地点の様子を表す重要なツイートであるため、削除されない仕組みを加える必要がある。

5. おわりに

本稿では、特定のエリアで多数の共通話題のツイートが投稿されているイベント地点を検出し、ランドマーク名を推定、関連ツイートを効率良く収集する手法を提案した。評価実験により、提案手法は従来手法より非イベント地点の検出を減らし、ランドマーク推定を高精度化し、関連ツイートを効率的に収集できるようになった。今後は、ランドマーク名が存在しない推定イベント地点の関連ツイート収集を行う手法を検討する。

参考文献

- [1] Foursquare : <https://foursquare.com/>
- [2] localmind : <http://www.localmind.com/>
- [3] 中澤昌美, 池田和史, 服部元, 小野智弘, "位置情報付きツイートからのイベント検出手法の提案", 情報処理学会全国大会講演論文集, 74th, 1, pp.1503-1504, 2012
- [4] Martin Ester, Hans-peter Kriegel, Jörg S, Xiaowei Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise", In Proceedings of 2nd International Conference on KDD, 1996.