場所と時間を考慮したタイムリーなイベント提示のための SNS を用いた情報抽出

工藤　瑠璃子† 丸 千尋† 榎 美紀†† 中尾　彰宏††† 山本　周††† 山口　実靖†††† 小口 正人†
†お茶の水女子大学　††IBM Research - Tokyo　†††東京大学　††††工学院大学

1. はじめに

近年, 様々な SNS が人々の間に普及し, 一種の情報伝達手段として一翼を担っている. SNS にはローカルイベントや地域特有の情報など, 特定の場所にいる人にとって有益な情報が含まれている. しかし, 有名な観光スポットなどの情報はガイドブックや WEB サイトから取得できるが, それらの媒体に載っていないようなローカルな情報や今まさに開催されているイベントを取得するのは, 膨大な情報の中から自力で探し出さなければならず, 非常に困難である.

また,2020 年の東京オリンピック開催が確定した 2013 年以降, 訪日外国人は急増しており [1], インバウンドへの情報配信手段を充実させるため,IoT デバイスの開発も進んでいる. 東京大学では 2015 年にネットワーク仮想化に対応したアクセスポイントを用いて, 様々な情報配信をするサービスを行う実証実験を実施している [2]. この実証実験では,Wi-Fi の無線ビーコンに情報をのせユーザのスマートフォンに情報を配信する BeaconCast を実装したアクセスポイントを, 空港と都内を結ぶリムジンバスに設置しており, インターネットにアクセスできない場合のメッセージ配信を実現している. これは, 通信契約をしていないインバウンドへの情報配信手段として有用である.

本研究では, このような IoT デバイスを基盤と考え, 旅行者などの時間とともに移動していく人に有用な情報を SNS の代表である Twitter から抽出し, インバウンド対応のタイムリーな情報提示手法を提案する.

2. 提案システム

観光者などに有用な情報をタイムリーにインバウンド対応で提示するために本研究で提案するシステムの概要を図 1 に示す.

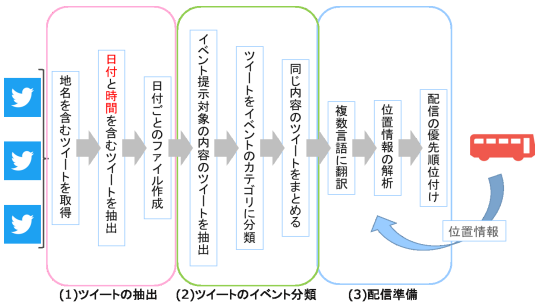

図 1: 提案システムの概要

(1) ツイートの抽出

1. Twitter API のキーワード検索で地名をキーワードに設定し, 地名に紐付いたツイートを取得
2. 1. で取得したツイートの本文に日時を含むツイートを抽出
   取得したツイートの本文に日付と時間が記載されているかを正規表現を使用して判定する. 日付と時間が記載されているツイートを抽出し, 日付ごとのファイルにまとめる.

(2) ツイートのイベント分類

1. イベント提示対象の内容のツイートを抽出
2. 抽出されたツイートをイベントのカテゴリごとに分類

(3) 配信準備

1. 提供する情報を複数言語に翻訳
   インバウンドに対応するため, ツイートを多言語に翻訳する必要がある. しかし, ツイートには不完全な日本語も多く, ツイート本文を原文のまま翻訳すると正しく翻訳されない場合が多い. そのため, ツイート本文中の情報を予め整理し, 必要な情報のみを提供する. 整理した情報を情報通信研究機構が開発した自動翻訳エンジン「みんなの自動翻訳@ TexTra」[3] を使用して翻訳する. 英語翻訳の例を図 2,3 に示す. 情報を整理する前処理を行わなかった結果が図 2 で, 前処理を施した結果が図 3 である.

〈紙わざ大賞26 〉 12/13(火)~17(土)10:00~19:00(最終日18:00まで)東京交通会館 東京都千代田区有楽町2-10-1私制作の紙わざ大賞 歴代入選作品今年26回〈南天雄鶏像〉25回〈飛翔一白鷹〉24回〈ヴェ… https://t.co/zhblqq79ZH

<Paper's Awards 26> 12/13 (Tue) 17 (soil) 10: paper of 00 19:00 (end 18:00) 2-10-1 I make Tokyo Traffic Hall Yurakucho, Chiyoda-ku, Tokyo Prize in the past winners of this year 26 <Nanten rooster image> 25 times without flying-Shirataka> 24 <ヴェ… https://t.co/zhblqq79ZH

図 2: Result of translation(full text).

イベント名: 〈紙わざ大賞26 〉	Event Name: <Paper Techniques Grand Prix 26>
日付: 12/13	Date: 13/12
時間: 10:00	Time: 10 00
場所: 千代田区	Location: Chiyoda Ward
カテゴリ: 展覧会	Category: exhibition

図 3: Result of translation.

これより, 図 2 のように原文のままでは正しく翻訳できない場合でも, 前処理を施した図 3 では必要な情報を簡潔に正しく提示できることが示された.

Information extraction using SNS for timely event presentation considering location and time
† Ruriko Kudo, Chihiro Maru, Masato Oguchi
†† Miki Enoki
††† Akihiro Nakao, Shu Yamamoto
†††† Saneatsu Yamaguchi
Ochanomizu University (†)
IBM Research - Tokyo(††)
Tokyo University (†††)
Kogakuin University(††††)

2. ユーザと提供するイベントの位置情報を取得
   本システムでは空港と都内を結ぶリムジンバスの乗客をユーザに想定しているため, ユーザの位置情報や目的地が取得可能である.
3. 提供する情報を並び替え
   場所と日付と時間の条件に適合する情報を, ユーザにとって有益な順で提示するために, 情報の順位付けを行う. 2. で取得した情報をもとに, ユーザの位置とイベント開催地の 2 地点の距離を算出し, 距離の近い順位提示する.

3. ツイートのイベント分類評価

場所と時間を含むツイートには観光者などに有用でないツイートも多く存在する. そこで, 有用でないツイートを排除するためにツイートの分類を行う. また, 観光者などに有用なツイートには, 様々なジャンルが混在しており, ユーザの嗜好にあった情報提供ができない. そのため, 抽出した有用なツイートをカテゴリごとに分類する.

3.1 カテゴリの設定と実験環境

本論文における「有用なツイート」とは以下のカテゴリのイベントに関するツイートとする.

音楽イベント	寄席・お笑い	舞台	
映画	展覧会	ポケモン GO	その他

カテゴリは, 都市情報雑誌の代表格である「東京 Walker」[4] のイベントカテゴリを参考にし, 予備実験で地名と日時を含むツイートの種類を分析した際に, 一定数のツイートが得られると判明したカテゴリを設定した.

実験データには,2016 年 12 月 6 日〜2016 年 12 月 13 日に取得したツイートの中で, イベント開催日が 12 月 13 日のツイート 1000 件を使用した.

ツイートの分類は,2 クラスの分類を行うことができる学習機械である SVM(Support Vecter Machine) で行う. ソフトウェアには SVM-Light[5] を使用した.

ツイートの分類は, まず有用か有用でないかの一対一分類を行う. 次に有用なツイートをカテゴリごとに分類する一対他分類を行う. カテゴリ分類では, あるカテゴリとその他のカテゴリ全部で SVM を構築していき, 多カテゴリに分類する. 今回はカテゴリを 6 個設定したので,6 個の SVM を作成し判定を行った. どのカテゴリにも判別されなかったツイートはその他のカテゴリとする.

3.2 実験結果

有用か有用でないかに分ける一対一分類 (1), カテゴリごとに分ける一対他分類 (2) の結果を表 1,2 に示す. (1) の分類では,1000 件のうち 200 件で検証を行った.(2) の分類では,(1) で有用であると判定された 482 件のうち, カテゴリごとに 100 件で検証を行った.

(1),(2) の分類ともに高い Accuracy と Recall の結果が得られたが, 表 2 の Precision は数値に偏りが出てしまった.

[舞台] の Precision の値が極めて低いのは, 他のカテゴリの判別モデルの重みの大きな素性と共通している素性が,[舞台] の SVM 判別モデルの重みの大きな素性の中に存在していることが原因であった. そこで, カテゴリ [舞台] に多く混在していた [音楽イベント] との一対一分類の SVM 判別モデルを作成, 分類し,[舞台] でないツイートを排除した. 結果を表 3 に示す.

表 1: Performance of classification(1).

Accuracy	Precision	Recall	F 値
0.8650	0.9091	0.8333	0.8857

表 2: Performance of classification(2).

カテゴリ	Accuracy	Precision	Recall	F 値
音楽イベント	0.8700	0.8867	0.8703	0.8784
寄席・お笑い	0.9100	0.6500	0.8666	0.7428
舞台	0.8100	0.1363	1.0000	0.2399
映画	0.9300	0.4000	0.8000	0.5333
展覧会	0.9600	0.4285	1.0000	0.5999
ポケモン GO	1.0000	1.0000	1.0000	1.0000

表 3: Performance of classification(3).

カテゴリ	Accuracy	Precision	Recall	F 値
舞台	0.9439	0.3809	0.9411	0.5516

表 3 に示されているように,Recall の値が少し下がってしまったが,Precision の値を改善することができた.

4. まとめと今後の課題

場所と時間を考慮した「その場」「その時」に観光者などが利用出来る, スポット的な有用性の高い情報を Twitter から抽出した. また, 今後さらに増えていくと考えられるインバウンドへ向け多言語で配信を実現するために, 情報を整理し, 高い精度での翻訳を可能にした.

現在のシステムでは, 地名と日付と時間の 3 点をツイートの本文に含むもののみを対象としており, 解析の対象となるツイートの種類が少ない. 今後はこの 3 点の情報の一部が欠けている場合に情報を補う手法を考え, 解析するツイートの種類を増やし, より多様な情報を提示できるようにする.

参考文献

[1] 日本政府観光局 「年別訪日外客数、出国人数の推移」 http://www.jnto.go.jp/jpn/statistics/marketingdata-outbound.pdf

[2] http://www.u-tokyo.ac.jp/public/public01-261023-j.html

[3] https://mt-auto-minhon-mlt.ucri.jgn-x.jp

[4] http://www.walkerplus.com

[5] http://svmlight.joachims.org