7K-02　訪日外国人の趣向に合わせたイベント情報配信のためのソーシャルストリームの解析

今井 美希 †　　榎 美紀 ††　　小口 正人 †
† お茶の水女子大学　　††IBM Research - Tokyo

1. はじめに

近年日本を訪れる外国人観光客は急激に増加しており，2020 年に開催される東京オリンピック・パラリンピックを考慮すると更なる増加が見込まれる．観光客の増加に伴い，有名な観光スポットなどの情報はガイドブックや WEB サイトから見受けられるようになってきた．しかしながら，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知り得るのは，現状難しい．また興味のあるイベント情報を自身の手によって見つけるのは手間がかかる．そこで我々は SNS にある情報に着目した．本研究では，ソーシャルストリームに基づき，イベント情報配信のために訪日外国人の趣向を推定する手法を提案する．旅行者などの時間とともに移動していく人々に有用な情報を SNS の代表である Twitter から抽出し，その情報をユーザの過去のツイート内容から推定した趣向に合わせて配信していくことを目指す．

2. 先行研究

タイムリーな観光情報提示のための SNS を用いたイベント抽出（dicomo 2017 ）という題で同研究室工藤がイベント情報の収集について研究を行っている．システムの概要を図 1 に示す．

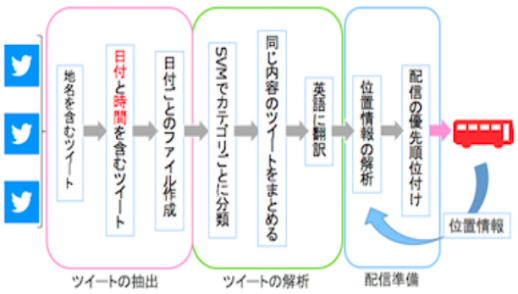

図 1: 提案システムの概要

この研究では，イベント情報の収集を行っている．そのイベント情報を使い，配信に向けて研究を進めていくとする．

3. 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 2 に示す．

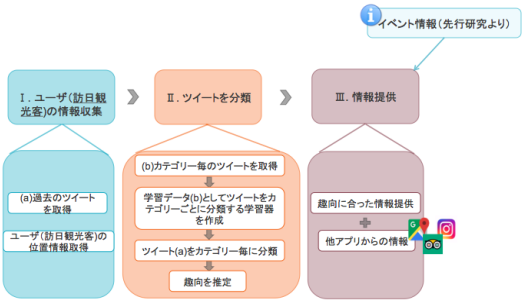

図 2: 提案システムの概要

1. ツイートの抽出
   (a) 過去のツイートを取得
   (b) ユーザ（訪日観光客）の位置情報取得
2. ツイートを分類
   (a) カテゴリー毎のツイートを取得
   (b) 学習データとしてツイートをカテゴリーごとに分類する学習器を作成
   (c) ツイートをカテゴリー毎に分類
   (d) 趣向を解析
3. 情報提供
   (a) 趣向に合った情報提供
   (b) 他アプリからの情報

4. 課題

システムを実現する上で課題となっていくことを挙げる

4.1 適切な情報提供

時間が限られる観光客にとって膨大な情報の中から興味がある情報を自力で探し出さなければならないのは，非常に煩わしい．　観光客それぞれの趣向にあった情報提供が必要だ．また，移動範囲が限られる旅行者には，その時その場で役立つ情報が必要であるため，移動している方向（交通手段）の推定が必要となってくる．

4.2 プライバシ

過去にツイートを収集し，また他のアプリと連携をしていくと考えると，どこまでの情報を扱い開示するのかプライバシについて考えていく必要がある．

4.3 リアルタイム性

時間，行動範囲が限られる旅行者には，「その時」「その場」で役立つ情報を発信していく必要がある．また，システムをより良いものにするために，多くのデータを扱うことになる．システムのパフォーマンスを考えたストリームのデータ処理が必要となってくる．

Analysis of social stream for event information distribution according to preference of foreigners visiting Japan
† Ruriko Kudo, Masato Oguchi
†† Miki Enoki
Ochanomizu University (†)
IBM Research - Tokyo(††)

5. 趣向の解析

趣向の解析方法について説明していく.

趣向の判定にはユーザの過去のツイートを使う. 過去のツイートをイベントのジャンルごとに分類し, 最も多くツイートが分類されたジャンル, つまり, 最も多くツイートをしていたジャンルが興味を持っているジャンルと定める. ジャンルに関しては, 先行研究で情報を収集していた, 舞台, 展示, ライブ, 映画の 4 ジャンルとする. 分類の流れについて説明していく. 今回は機械学習の手法の 1 つであるランダムフォレストという手法を使い, 分類していく. ジャンル毎に英語のツイートを取得する. それを学習データとしてツイートをジャンル毎に分類する学習器をランダムフォレストによって作成する. 分類精度を向上させるため, 形態素解析やチューニングを行っていく.

6. 趣向の解析の詳細

6.1 ツイート抽出

Twitter API のキーワード検索で各ジャンルごとのキーワードを設定, また言語を英語に設定することで英語のツイートを取得する. ツイートは各ジャンル 300 ツイートずつ取得し, キーワードは以下のように設定する.

舞台 musical

展示会 art, museum, gallery, anime

ライブ concert

映画 movie, movie theater, cinema, film

上記以外のツイート

位置情報による取得も試みたがほとんど取得できなかった. また取得する際, ツイートをする人が多くいる時間帯を見計らうことで多くのツイートを取得することが可能となった.

6.2 学習

6.2.1 学習器作成

6.1 で取得したツイートを使い, ユーザのツイートを分類するための学習器を作成していく. まず特徴語辞書を作成していく. 各ジャンルのツイートに関する特徴を捉えるためだ. ツイートを単語に分割, 更に小文字に直す. そこから is, a, RT など出現回数が多い単語を予め設定しておき取り除く.

特徴ベクトル (出現頻度のベクトル) に変換するため, 単語と ID, ID と頻度にマッピングした後, ベクトルにする. オーバーフィッティングを避けるためクロスバリデーションを行い, 学習データとテストデータを 7 対 3 にわける. 学習データを使いランダムフォレストによって学習する. テストデータを入れた結果 72.4%の正答率となった.

6.3 精度の向上

6.2.1 の結果, 精度は 72.4 %だった. 更に精度をあげるため, チューニングを行っていく.

6.3.1 形態素解析

英単語には過去分詞 (ed), 現在分詞 (ing), 複数形があり, 分類の妨げとなると考えた. そこで, 形態素解析を行い英単語を過去分詞 (ed), 現在分詞 (ing), 複数形などを標準形でまとめた. その結果正答率は, 3.1 %向上し 75.5 %となった.

6.3.2 チューニング

scikit-learn のランダムフォレストには多くのパラメタがある.

一部を説明する

$\mathrm{num}_t rees$: いくつ決定木を作成するか

$\mathrm{max}_d epth$: どの深さの決定木を作成するか

$\mathrm{num}_f eatures$: 目的変数のサンプリング時に, いくつの目的変数をサンプリングするか

これらを調整することで, よりランダムフォレストで正確な分類をできるようになる. そこでグリッドサーチという自動的な最適化ツール, を使い与えたパラメータの中で最も精度の良いものを選ぶ. 次に, その選んだパラメータの付近でグリットサーチを行いより最適なパラメータを選ぶ.

結果 7.3 %向上し, 82.8 %となった.

7. 結果

ランダムフォレストによって学習器を作成し, ツイートを分類してきた. はじめは 72.4 %だった分類精度が, 形態素解析の結果 75.5 %となった. 更にチューニングをし, パラメータを改めたところ 82.8 %となり 10.4 %精度をあげることができた.

8. まとめと今後の課題

今回, 形態素解析やチューニングなど最適化を重ねることで 80 %以上の精度で分類を成功することができた.

今後はより学習器の精度を上げていくために, tfidf という文書中に含まれる単語の重要度を評価する手法を使い, 単語に重み付けることで精度の向上を目指す.

また, 実際に訪日観光客のツイートの趣向を判定できるか検証していこうと考えている.

謝辞

本研究の一部はお茶の水女子大学と日本 IBM との共同研究契約に基づくものである.

参考文献

[1] 日本政府観光局
https://www.jnto.go.jp/jpn/statistics/data$_i nfo_l isting/pdf/$

[2] http://scikit-learn.org/stable/modules/generated/sklearn.ensemb