

地理情報推薦のための観光スポットと 位置情報付きユーザ生成コンテンツの対応付け手法の提案

奥 健太[†] 橋本 拓也[†] 上野 弘毅^{††} 服部 文夫[†]

[†] 立命館大学情報理工学部 〒525-8577 滋賀県草津市野路東1丁目1-1

^{††} 立命館大学大学院情報理工学研究科 〒525-8577 滋賀県草津市野路東1丁目1-1

E-mail: [†]oku@fc.ritsumei.ac.jp, ^{††}{is0033hr,is006088}@ed.ritsumei.ac.jp, ^{†††}fhattori@is.ritsumei.ac.jp

あらまし 我々は、膨大な観光スポットの中から適切な観光スポットを提供する観光スポット推薦システムの開発を行っている。効果的な観光スポット推薦を実現するためには、季節や時間帯に応じて魅力が大きく変化する観光スポットを適切に特徴化しておく必要がある。本稿では、近年、膨大に発信されている位置情報付きツイートに基づき観光スポットの特徴化を行う手法を提案する。具体的には、領域判別問題の解法の一つである One-Class SVM を用い、対象スポット名をタイトルにもつ位置情報付き写真および対象スポット名をテキストに含む位置情報付きツイートを基に、対象スポットの実質的な活動領域を推定し、その活動領域内に含まれる位置情報付きツイート集合を対象スポットに対応付ける。また、観光スポットに対応付けられたツイートを基に、観光スポットの特徴を抽出する手法を提案する。さらに、提案手法の有効性を確認するために、主要な観光スポットが集中している京都における実データを基に行った定性分析の結果を示す。

キーワード 地理情報推薦, 位置情報付きユーザ生成コンテンツ, ツイート, 観光情報

Kenta OKU[†], Takuya HASHIMOTO[†], Koki UENO^{††}, and Fumio HATTORI^{†††}

[†] College of Information Science and Engineering, Ritsumeikan University,
1-1-1 Nojihigashi, Kusatsu-city, Shiga, 525-8577, Japan

^{††} Graduate School of Information Science and Engineering, Ritsumeikan University,
1-1-1 Nojihigashi, Kusatsu-city, Shiga, 525-8577, Japan

E-mail: [†]oku@fc.ritsumei.ac.jp, ^{††}{is0033hr,is006088}@ed.ritsumei.ac.jp, ^{†††}fhattori@is.ritsumei.ac.jp

1. はじめに

情報技術により観光産業を活性化させることの需要が高まってきた [1]。観光は、非日常的な体験を楽しむ人間の行動であり、風景や観光施設、地域の名物（特産品やイベントなど）などといった観光資源に触れることができる。観光を通して、人々は新たな発見が得られ、生活がより豊かなものになると共に、地域にも潤いをもたらす、地域経済の活性化につながることから、観光産業を活性化させることの重要性は高い。特に世界中には、膨大かつ多種多様な観光資源が存在するため、利用者へ適切な観光資源を提供する手段として、情報検索技術や情報推薦技術を適用することの意義は大きい。

本研究では、観光資源の中でも観光スポットに焦点をあてる。観光スポットの効果的な検索や推薦を実現させるためには、観光スポットに関する情報を検索および推薦可能な形にすることが必要不可欠である。現在、Web 上には多くの観光情報サイ

ト^{(注1)(注2)}が存在するものの、観光情報サイトに掲載されているような情報は、ある時点において取材された静的な情報であるということが多い。

しかし、観光スポットに関する情報の特質として、季節や時間帯、その時期の流行や行事などによって、観光スポット自体の特徴が動的に大きく変動することが挙げられる。たとえば、紅葉で有名なスポットであれば、見ごろである秋季に魅力が最も高くなるし、夜景で有名なスポットであれば、夜の時間帯の方が魅力が高くなる。さらには、お祭りなどが行われる地域であれば、それが開催される時期には急激に魅力が高くなる。しかしながら、既存の観光情報サイトは人手で管理されていることが多く、このように動的に特徴が大きく変化する観光スポットの情報を定期的に更新するには大きなコストがかかる。

(注1): じゃらん: <http://www.jalan.net/>

(注2): 楽天トラベル: <http://travel.rakuten.co.jp/>

このような問題を解決するために、本研究では、近年 Web 上に膨大に発信されている位置情報付きユーザ生成コンテンツ (Geotagged User Generated Content; G-UGC) に注目する。G-UGC を発信できるサービスとしては、foursquare^(注3)や Twitter^(注4)、Panoramio^(注5) などが挙げられる。G-UGC サービスの中でも、特に Twitter では、ユーザが気軽に発信できるという利点から、日々膨大な量の記事 (ツイート) が発信されている。さらに、近年では、GPS 付き携帯端末の普及により、位置情報付きツイートも膨大に発信されている。位置情報付きツイートには、発信時の位置情報 (緯度・経度) に加え、発信時刻や発信ユーザの情報、発信記事の内容 (つぶやき) などが含まれる。このような位置情報付きツイートを観光スポットに対応付けることにより、動的に変化する観光スポットの特徴を抽出することができると考える。

ここで、観光スポットと位置情報付きツイートとを適切に対応付けるには課題がある。ツイートのテキストや緯度・経度からだけでは、どの観光スポットに関するツイートであるのか判断することが困難である。たとえば、ツイートのテキストに「清水寺なう」など、観光スポット名が明示されていれば対応付けも容易ではあるが、観光スポット名を明示していないツイートがほとんどである。また、観光スポットの緯度・経度を中心にした任意の大きさをもつ領域内に含まれるツイートを観光スポットに対応付けることも考えられるが、観光スポットの領域は小ささまざまであり、どのスポットにも対応できる適切な領域を定義することは容易でない。

上記の課題に対し、本研究では、対象とする観光スポット周辺において、対象スポット名をタイトルにもつ位置情報付き写真および対象スポット名をテキストに含む位置情報付きツイートを基に、対象スポットの実質的な活動領域を推定する。観光スポットの実質的な活動領域とは、観光スポットの登録上の緯度・経度や住所で表される領域とは異なり、人々が実際にその観光スポットを楽しむために訪れる領域と定義する。この活動領域の推定には、領域判別問題の解法の一つである One-Class SVM (One-Class Support Vector Machine; OC-SVM) [2] を用いる。推定された活動領域内に含まれる位置情報付きツイート集合を対象スポットに対応付ける。

本稿の構成は以下のとおりである。第 2. 章では、関連研究について取り上げる。第 3. 章では、本研究で用いる位置情報付きユーザ生成コンテンツについて説明する。第 4. 章では、観光スポットと位置情報付きツイートの対応付け手法について説明し、第 5. 章では、対応付けられたツイートを基に観光スポットの特徴を抽出する手法を説明する。第 6. 章では、上記の対応付け手法および特徴抽出手法の有効性を確認するため、実データを用いた定性分析について述べる。最後に、第 7. 章で、本稿をまとめ、今後の展望について述べる。

2. 関連研究

位置情報付きユーザ生成コンテンツから地域情報を抽出する研究として次のようなものが挙げられる。

木俣ら [3] は、時空間情報をもった音声マイクロブログなどで得られたデータから、その時間と場所の特徴、つまり雰囲気を表す語句を抽出する手法を提案している。具体的には、ある時区間に含まれるデータの位置情報に基づきクラスタリングを行い、各クラスタに含まれる語句から、その時空間の雰囲気を表す語句を生成する。

Crandall ら [4] は、flickr に投稿される位置情報付き写真に対し、クラスタリング手法である Mean-shift 法を適用することで、多くの人に写真が撮影されやすいようなランドマークを抽出し、地図上にそのランドマークの写真を提示するシステムを提案している。Zheng ら [5] [6] は、独自に収集した GPS 軌跡データから、多くの人々が滞在する地点を抽出し、クラスタリングを行うことで、関心地点 (Point of Interest; POI) の抽出を行っている。

山中ら [7] は、携帯電話などから発信された時空間情報付きのテキストデータを要約することで、その地域の管理者が状況把握を行うことを支援するシステムを提案している。あらかじめ注目する観点をカテゴリとして用意しておき、新たに発信されたテキストデータを適宜クラスタリングしながら、そのカテゴリに自動分類していくものである。

Lee ら [8] は、位置情報付きツイートを情報源とし、そこから地域イベントを抽出する手法を提案している。特に、地域ごとの人々の通常活動状態との差異を考慮することで、通常とは異なる特別な活動が行われている地域を抽出することで地域イベントの抽出を行っている。Lee らは、地域イベントの抽出を目的としていることから、重要な手掛かりとなる対象地域の時間的变化に着目している。

李ら [9] は、位置情報付きツイートをを用いて都市空間における群集行動をモニタリングし、都市の地域特徴を抽出する手法を提案している。具体的には位置情報付きツイートに基づいて作成された群集行動ベクトルを分析することで、都市における群集行動パターンを抽出している。

以上のように位置情報付きユーザ生成コンテンツから特徴的な地域を抽出しようという研究は多く行われている。これら既存の研究では、クラスタリング手法などにより、ユーザ生成コンテンツが地理的に集中している地域を抽出することを試みているのに対し、本研究では、まず、位置情報付きツイートを対象スポットに対応付けたうえで、対象スポットの特徴を抽出することを目指している。

3. 位置情報付きユーザ生成コンテンツ

本研究では、位置情報ベースの SNS の一つである foursquare から観光スポットの情報を取得する。そして、マイクロブログサービスの一つである Twitter 上に発信された位置情報付きツイートを観光スポットに対応付ける。また、観光スポットと位置情報付きツイートとを対応付けるための手掛かりとして、写

(注3): <https://foursquare.com/>

(注4): <https://twitter.com/>

(注5): <http://www.panoramio.com/>

表 1 観光スポットとして対象とするカテゴリ名

* 文献 [10] を参考

対象カテゴリ名	大分類 *	小分類 *
Hiking Trail	自然資源	1. 山岳
Mountain	自然資源	1. 山岳
Lake	自然資源	5. 湖沼
River	自然資源	8. 河川
Beach	自然資源	9. 海岸
History Museum	人文資源	1. 史跡
Cemetery	人文資源	1. 史跡
Historic Site	人文資源	1. 史跡
Religious Center	人文資源	1. 史跡
Shrine	人文資源	2. 寺社
Temple	人文資源	2. 寺社
Campground	人文資源	4. 庭園・公園
Dog Run	人文資源	4. 庭園・公園
Farm	人文資源	4. 庭園・公園
Garden	人文資源	4. 庭園・公園
Garden Center	人文資源	4. 庭園・公園
Playground	人文資源	4. 庭園・公園
Park	人文資源	4. 庭園・公園
Sculpture Garden	人文資源	6. 碑・像
Monument / Landmark	人文資源	6. 碑・像
Bridge	人文資源	7. 建造物
Harbor / Marina	人文資源	7. 建造物
Boat or Ferry	人文資源	7. 建造物
Pier	人文資源	7. 建造物
Zoo	人文資源	8. 動物園・植物園
Museum	人文資源	9. 博物館・美術館
Art Gallery	人文資源	9. 博物館・美術館
Art Museum	人文資源	9. 博物館・美術館
Arts & Entertainment	人文資源	9. 博物館・美術館
Science Museum	人文資源	9. 博物館・美術館
Aquarium	人文資源	10. 水族館
Scenic Lookout	人文資源	11. 景観

真共有サイトの一つである Panoramio 上に発信された位置情報付き写真を用いる。

本章では、これら 3 種類の位置情報付きユーザ生成コンテンツについて説明し、これらの収集方法について述べる。

a) 観光スポット

foursquare は、位置情報ベースの SNS の一つであり、2009 年 3 月にサービスが公開された。ユーザは、携帯端末を用いて、venue とよばれる特定の場所で check-in を登録することで、サービスに参加できる。

本研究では、foursquare が公開している foursquare API^(注6)を用いて、venue に関する情報の収集を行った。venue には、“University” や “Train Station,” “Bookstore” などさまざまなカテゴリに属するものが含まれるが、本研究では、表 1 に示すカテゴリに属する venue を観光スポットとして扱う。なお、表 1 に示すカテゴリは、溝尾による観光資源の分類 [10] を参考

(注6): <https://developer.foursquare.com/>

表 2 観光スポットテーブルの属性

属性	説明
id	観光スポット ID
name	観光スポット名
address	住所
latitude	緯度
longitude	経度
category_name	カテゴリ名
url	URL

表 3 ツイートテーブルの属性

属性	説明
id	ツイート ID
user_id	投稿者 ID
user_name	投稿者名
text	テキスト
year	発信された年
week_of_year	発信された週
hour	発信された時刻
latitude	緯度
longitude	経度

に選定した。

収集された観光スポットに関する情報を、次の *spot* テーブルに格納した。

- *spot(id, name, address, latitude, longitude, category_name, url)*

ここで、各属性の説明は、表 2 のとおりである。

b) 位置情報付きツイート

Twitter は、マイクロブログサービスの一つであり、2006 年 7 月にサービスが公開された。ユーザは、ツイートとよばれる記事を発信することが可能である。また、GPS 付き携帯端末の普及により、近年では位置情報付きツイートも膨大に発信されている。発信ツイート数は年々増大しており、2010 年には 1 日 3,500 万ツイートを記録している。

本研究では、Twitter が公開している streaming API^(注7)を用いて、位置情報付きツイートの収集を行った。収集された位置情報付きツイートに関する情報を、次の *tweet* テーブルに格納した。

- *tweet(id, user_id, user_name, text, year, week_of_year, hour, latitude, longitude)*

ここで、各属性の説明は、表 3 のとおりである。

c) 位置情報付き写真

Panoramio は、写真共有サイトの一つであり、2005 年 10 月にサービスが公開された。ユーザは、地図上に写真をアップロードすることが可能である。

本研究では、Panoramio が公開している Panoramio API^(注8)を用いて、位置情報付き写真の収集を行った。収集された位置情報付き写真に関する情報を、次の *photo* テーブルに格納する。

(注7): <https://dev.twitter.com/docs/streaming-apis>

(注8): <http://www.panoramio.com/api/data/api.html>

表 4 写真テーブルの属性

属性	説明
photo_id	写真 ID
photo_title	写真タイトル
photo_url	写真 URL
longitude	経度
latitude	緯度
owner_id	投稿者 ID
owner_name	投稿者名

表 5 位置情報付きユーザ生成コンテンツの特徴の比較

ユーザ生成コンテンツ	データ量	対象の明確さ	対象の領域	テキスト情報	時間情報
観光スポット	少	明確	なし	なし	なし
位置情報付きツイート	膨大	不明確	あり	あり	あり
位置情報付き写真	少	明確	あり	なし	なし

• $photo(photo_id, photo_title, photo_url, longitude, latitude, owner_id, owner_name)$

ここで、各属性の説明は、表 4 のとおりである。

以上の 3 種類の位置情報付きユーザ生成コンテンツの特徴をまとめると表 5 のようになる。

Twitter では、一度に発信できる字数を 140 字に制限することで、ユーザにとっては気軽に発信できるという利点があり、このような利点により日々膨大な量のツイートが発信される。一方で、位置情報付き写真に関しては、気軽に発信されるツイートと比べ、投稿者が良いと思い撮影された写真の中から、特に良いと思った写真がアップロードされる傾向があるため、Panoramio 上には、特に選りすぐれた写真がアップロードされているといえる。また、位置情報付き写真のタイトルにはスポット名が明記されていることも多いため、対象とするスポットが明確であるという利点もある。位置情報付きツイートに比べ、投稿されるデータ量は少ないものの、位置情報付き写真自体の質は高いものといえる。

位置情報付きツイートは膨大なデータ量を有し、さらに付加情報としてテキスト情報や時間情報が含まれるため、観光スポットの特徴抽出には有効な情報源であるといえる。しかしながら、対象とするスポットが不明確であるという欠点をもつため、対象が明確である観光スポットおよび位置情報付き写真とを効果的に連携させることが課題となる。第 4 章では、観光スポットと位置情報付きツイートの対応付け手法について述べる。

なお、本稿では、収集した観光スポット集合、位置情報付きツイート集合、位置情報付き写真集合を、それぞれ次のように表す。

$$S = \{s_1, s_2, \dots\} \quad (1)$$

$$T = \{t_1, t_2, \dots\} \quad (2)$$

$$P = \{p_1, p_2, \dots\} \quad (3)$$

また、各属性を指すときには、 $t_1.text$ のように、[オブジェクト. 属性名] の形式で表記する。

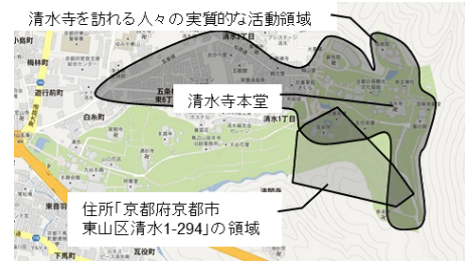


図 1 清水寺の実質的な活動領域



図 2 One-Class SVM による領域判別の例

4. 観光スポットと位置情報付きツイートの対応付け手法

観光スポットと位置情報付きツイートを対応付けることにより、対応付けられたツイートを基に、観光スポットに関する特徴を抽出することが可能となる。しかし、ツイートのテキストや緯度・経度からだけでは、どの観光スポットに関するツイートであるのか判断することが困難である。たとえば、ツイートのテキストに「清水寺なう」など、観光スポット名が明示されていれば対応付けも容易ではあるが、観光スポット名を明示していないツイートがほとんどである。また、観光スポットの緯度・経度を中心にした任意の大きさをもつ領域内に含まれるツイートを観光スポットに対応付けることも考えられるが、観光スポットの領域は大小さまざまであり、どのスポットにも対応できる適切な領域を定義することは容易でない。

以上の問題を解決するために、本研究では、対象とする観光スポット周辺において、対象スポット名をタイトルにもつ位置情報付き写真および対象スポット名をテキストに含む位置情報付きツイートを基に、対象スポットの実質的な活動領域を推定する。そして、その活動領域内に含まれる位置情報付きツイートを対象スポットに対応付ける手法を提案する。

ここで、観光スポットの実質的な活動領域とは、観光スポットの登録上の緯度・経度や住所で表される領域とは異なり、人々が実際にその観光スポットを楽しむために訪れる領域と定義する。たとえば、清水寺を訪れる人々は、清水寺本堂だけでなく、周辺の建造物や参道等にも訪れることが多く、実質的にはこれらの領域も清水寺の領域であるとみなすことができる。清水寺の住所は「京都府京都市東山区清水 1-294」であるが、この実質的な活動領域は必ずしも住所と一致しない。図 1 には、例として清水寺の住所に一致する領域（薄い灰色の領域）と、実質

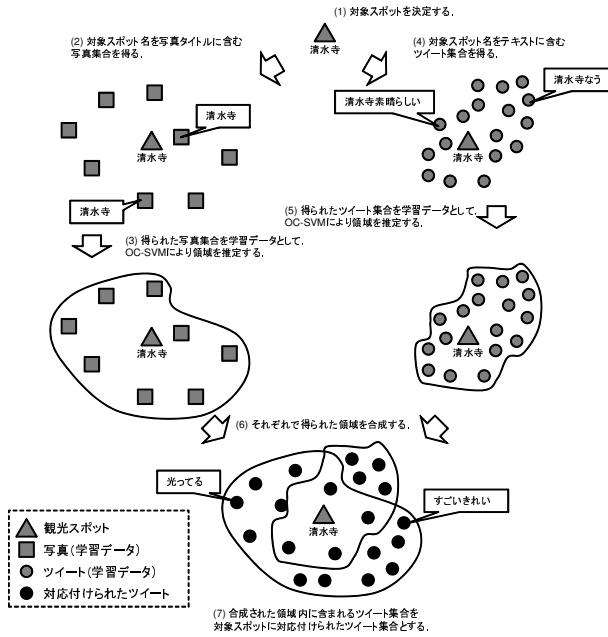


図3 観光スポットと位置情報付きツイートの対応付け手法の流れ

的な活動領域（濃い灰色の領域）を示している．

観光スポットの活動領域の推定には、領域判別問題の解法の一つである OC-SVM [2] を用いる．OC-SVM では、図 2 のように、与えられた学習データ群から高密度領域を推定することが可能である．SVM は、学習データが比較的少ない場合においても、高い汎化性能を示すという特徴をもつ

本研究では、対象とする観光スポット周辺において、対象スポット名をタイトルにもつ位置情報付き写真および観光スポット名をテキストに含む位置情報付きツイートを学習データとして、OC-SVM により領域推定を行う．

以下、図 3 を例に、対応付け手法の具体的な手順について説明する．

- (1) 対象とする観光スポット $s_i \in S$ を決定する．
- (2) 対象スポットから半径 r 以内において、対象スポット名 $s_i.name$ を写真タイトル $p_i.photo_title$ に含む写真集合 $P_i^* \subseteq P$ を得る．
- (3) 写真集合 P_i^* を学習データとして、OC-SVM により領域 $R_{P_i^*}$ を推定する．
- (4) 対象スポットから半径 r 以内において、対象スポット名 $s_i.name$ をテキスト $t_i.text$ に含むツイート集合 $T_i^* \subseteq T$ を得る．
- (5) ツイート集合 T_i^* を学習データとして、OC-SVM により領域 $R_{T_i^*}$ を推定する．
- (6) 領域 $R_{P_i^*}$ および領域 $R_{T_i^*}$ を合成し、領域 R_i を得る．この領域 R_i を観光スポット s_i の活動領域とする．
- (7) 領域 R_i 内に含まれるツイート集合 T_i を観光スポット s_i に対応付けられたツイート集合とする．

5. 観光スポットの特徴抽出手法

観光スポット s_i に対応付けられたツイート集合 T_i に基づき、

観光スポット s_i の特徴抽出を行う．本研究では、次の 2 種類の特徴の抽出を行う．

- (a) 観光スポットの時間的特徴
- (b) 観光スポットの特徴語句

以下、それぞれの抽出方法について説明する．

5.1 観光スポットの時間的特徴の抽出

ツイート集合 T_i 内のツイート発信時間 ($t_i.year$, $t_i.week_of_year$, $t_i.hour$) の分布を解析することで、観光スポットの時間的特徴を抽出することができる．たとえば、清水寺は紅葉で有名であるため、紅葉の時期には特別多くのツイートが発信される．また、特にこの時期は夜間ライトアップのイベントが催されているため、夜間に多くのツイートが発信される．

まず、 $week_of_year$ に関する時間的特徴の抽出方法について述べる．なお、 $week_of_year$ に関する時間的特徴は、年単位で抽出するものとする．ここでは、2011 年の清水寺を対象スポットの例として説明する．

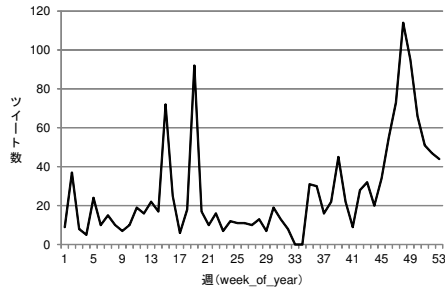
(1) 週ごと ($week_of_year = 1, 2, \dots, 53$) に発信されたツイート数を集計する．ここでは、週ごとのツイート数を要素とした特徴ベクトル $W_i = (w_{i1}, w_{i2}, \dots, w_{i53})$ を定義する．特徴ベクトル W_i の内容をグラフ化すると図 4(a)（横軸は週、縦軸はツイート数を表す）のようになる．

(2) ツイート集合 T_i 内の全ツイート数 $|T_i|$ が 1 となるように、週ごとのツイート数を正規化する．ここで正規化するのは、知名度のある観光スポットとそうでない観光スポットの間では、対応付けられるツイート数に大きな違いがみられるという格差を解消するためである．正規化後の週ごとのツイート数を要素とした正規化特徴ベクトル $W_i^* = (w_{i1}^*, w_{i2}^*, \dots, w_{i53}^*)$ を定義し、これをグラフ化すると図 4(b) のようになる．

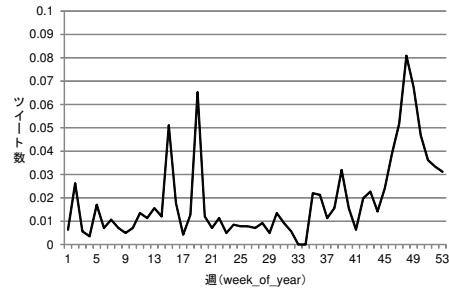
(3) (1), (2) を全スポットについて行う．全スポットについて得られた正規化後のツイート数の平均を得る．全スポットについて得られた正規化後のツイート数の平均を要素とした平均特徴ベクトル $W_i^{all} = (w_{i1}^{all}, w_{i2}^{all}, \dots, w_{i53}^{all})$ を定義し、これをグラフ化すると、図 4(c) のようになる．

(4) 正規化特徴ベクトル W_i^* と平均特徴ベクトル W_i^{all} との差分をとった、差分特徴ベクトル $W_i^{diff} = W_i^* - W_i^{all}$ を得る．ここで、差分特徴ベクトルを求めるのは、全体的に時期により発信ツイート数に偏りがみられるため、その偏りを消去するためである．差分特徴ベクトル W_i^{diff} をグラフ化すると、図 4(d) のようになる．この例では、清水寺は、特に第 19 週（ゴールデンウィークの時期）や第 49 週（紅葉の時期）などにおいて魅力が高くなるといえる．

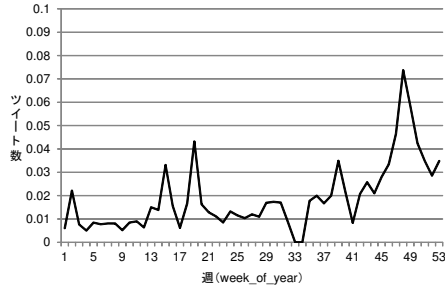
同様の手順で、 $hour$ に関する時間的特徴の抽出を行う．なお、 $hour$ に関する時間的特徴は、年および週単位で抽出するものとする．週単位で抽出することにより、夏場は昼間が人気があり、冬場は夜間が人気があるなど、季節ごとの特徴を抽出することができる．詳細な手順は割愛するが、清水寺の 2011 年第 19 週および 2011 年第 49 週における $hour$ に関する時間



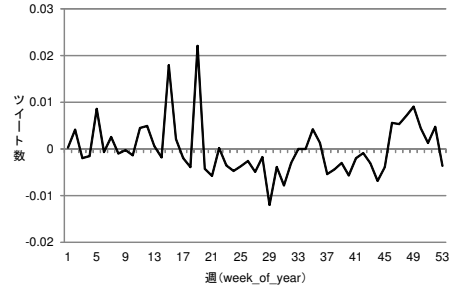
(a) 清水寺における週別ツイート数 (2011 年)



(b) 清水寺における正規化後の週別ツイート数 (2011 年)

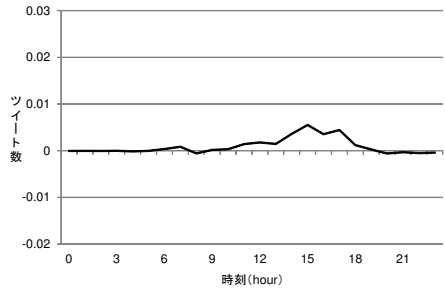


(c) 全スポットにおける平均週別ツイート数 (2011 年)

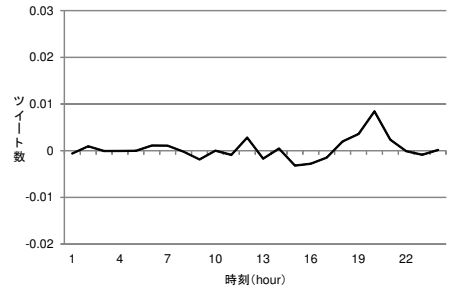


(d) 清水寺における週別ツイート数の差分 (2011 年)

図 4 清水寺における *week_of_year* に関する時間的特徴 (2011 年)



(a) 清水寺における時刻別ツイート数の差分 (2011 年第 19 週)



(b) 清水寺における時刻別ツイート数の差分 (2011 年第 49 週)

図 5 清水寺における *hour* に関する時間的特徴 (2011 年)

的特徴をグラフ化したものを、それぞれ、図 5(a), (b) に示す。図のように、第 19 週においては、15 時前後が高くなっているのに対し、第 49 週においては、20 時前後が高くなっていることがわかる。これは、第 49 週においては、清水寺ライトアップの催しが行われていたためであり、それを観賞するために多くの人々が訪れていたことを反映している。

なお、ここでは、先述と同様の手順にしたがって、それぞれ、特徴ベクトル H_i 、正規化特徴ベクトル H_i^* 、平均特徴ベクトル H_i^{all} 、差分特徴ベクトル H_i^{diff} が得られているものとする。

5.2 観光スポットの特徴語句の抽出

ツイート集合 T_i 内のテキスト ($t_i.\text{text}$) から形態素解析によ

り観光スポットの特徴語句を抽出する。本研究では、形態素解析器として、茶筌^(注9)を用いる。抽出対象の品詞を表 6 のとおりとした。ただし、「清水寺本堂」や「清水の舞台」などのような複合語に対応するため、連続する名詞や未知語、「の [助詞-連体化]」で連結される語句は一つの語句として抽出している。

特徴語句の抽出は週別に行う。観光スポット s_i の第 j 週目に抽出された語句集合を、

$$L_{ij} = \{l_{ij1}, l_{ij2}, \dots\} \quad (4)$$

と表す。

(注9): <http://chasen.naist.jp/hiki/ChaSen/>

表 6 形態素解析による抽出対象の品詞

抽出対象の品詞
名詞-一般
名詞-数
名詞-固有名詞-一般
名詞-固有名詞-人名-一般
名詞-固有名詞-人名-姓
名詞-固有名詞-人名-名
名詞-固有名詞-組織
名詞-固有名詞-地域-一般
名詞-固有名詞-地域-国
名詞-接尾-サ変接続
未知語
形容詞-自立
名詞-接尾-一般
名詞-接尾-地域
記号-一般

表 7 清水寺の特徴語句（上位 10 件）

順位	特徴語句（() 内は tf-idf 値）	
	第 16 週	第 49 週
1	Kiyomizudera(52.8365)	清水寺 (161.3635)
2	清水寺 (41.5633)	東山区清水 (99.0543)
3	東山区清水 (29.3494)	善光寺堂 (22.9248)
4	清水の舞台 (13.9454)	清水の舞台 (20.9181)
5	京都市 (5.9212)	京都市 (19.9841)
6	さくら (5.5048)	清水寺本堂 (15.0961)
7	自主桜 (3.8918)	地主神社 (11.3221)
8	ザ・舞台 (3.8918)	清水寺境内 (7.5481)
9	八重 (3.8918)	ライトアップ (6.1777)
10	夜間参拝 (3.8918)	ganbareeeee(4.5850)

抽出された各語句 l_{ijk} の重みを算出する．ここでは，文書検索においてよく用いられる $tf-idf$ を用いる． tf とは，語句の出現頻度（term frequency）であり，語句 l_{ijk} の tf は，観光スポット s_i の第 j 週目における語句 l_{ijk} の出現回数となる．また， idf とは，語句の逆文書頻度（inverse document frequency）であり，語句 l_{ijk} の idf は，次式より算出される．

$$idf = \log \frac{|S|}{n_{ijk}} \quad (5)$$

ここで， $|S|$ は全スポット数であり， n_{ijk} は語句 l_{ijk} を，対応付けられたツイートのテキストに含むスポット数である．語句 l_{ijk} の $tf-idf$ は次式より算出される．

$$tf-idf = tf \times idf \quad (6)$$

最終的に，語句集合 L_{ij} 内で $tf-idf$ の高い順に整列しておく．こうしておくことで，各観光スポットの週ごとの特徴語句を提示することが可能となる．たとえば，清水寺の第 16 週目および第 49 週目における上位 10 件の特徴語句は表 7 のとおりとなる．表のように，桜の時期においては「さくら」という語句が抽出されており，後者のように紅葉の時期においては「ライトアップ」という語句が抽出されていることがわかる．

表 8 実験で収集したユーザ生成コンテンツ

ユーザ生成コンテンツ	コンテンツ数	収集期間
観光スポット	1,006	2012 年 6 月 13 日
位置情報付きツイート	389,579	2011 年 1 月 1 日から 12 月 31 日
位置情報付き写真	12,480	2012 年 6 月 30 日

6. 定性分析

本章では，第 4. 章で述べた対応付け手法および第 5. 章で述べた特徴抽出手法の有効性を確認するために，実データを用いた定性分析を行う．

6.1 データセット

データセットとしては，第 3. 章で述べた，観光スポットおよび位置情報付きツイート，位置情報付き写真の 3 種類のユーザ生成コンテンツを用いた．

本実験では，日本国内において主要な観光スポットが集中している京都を対象として選んだ．ここでは，南西緯度経度 (34.87069, 135.566713) から，北東緯度経度 (35.12967, 135.935152) の矩形領域を対象エリアとして設定した．この矩形領域内の各ユーザ生成コンテンツを収集した．収集した各コンテンツ数およびその収集期間は表 8 のとおりである．

6.2 対応付け手法の分析

対応付け手法の有効性について定性的に分析する．ここでは，(a) 清水寺，(b) 金閣寺，(c) 哲学の道の事例を取り上げる．

図 6 は，各観光スポットに対応付けられた位置情報付きツイート集合を示している．

(a) 清水寺に関しては，第 4. 章で述べたように，清水寺の公式の住所以外の領域においてもツイートが対応付けられていることがわかる．また，清水寺本堂以外にも周辺の建造物や参道におけるツイートも対応付けられている．

(b) 金閣寺に関しては，金閣寺周辺以外にも，最寄りのバス停「金閣寺道」から金閣寺に向かうまでの道中におけるツイートも対応付けられている．実際に「いざ金閣寺へ！」や「歩いて金閣寺へ」など，これから金閣寺へ向かうような内容のツイートが多い．

(c) 哲学の道に関しては，南北に伸びる道に沿ってツイートが対応付けられた．観光経路は線形状になっているが，提案手法では，このような観光経路の抽出も可能であることを示した．

以上，一例ではあるが，提案の対応付け手法により，観光スポットの実質的な活動領域を考慮しながら，観光スポットと位置情報付きツイートとを適切に対応付けることが可能であることを示した．

6.3 特徴抽出手法の分析

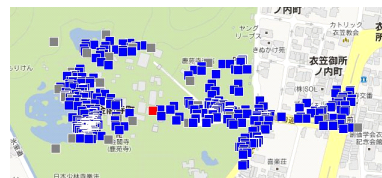
特徴抽出手法の有効性について定性的に分析する．ここでは，(a) 金閣寺，(b) 哲学の道，(c) 八坂神社の事例を取り上げる．

各観光スポットの $week_of_year$ に関する時間的特徴を図 7 に示す．

(a) 金閣寺に関しては，他のスポットに比べ，特に冬の時期（第 53 週から第 3 週まで）において，多くのツイートが発信さ



(a) 清水寺

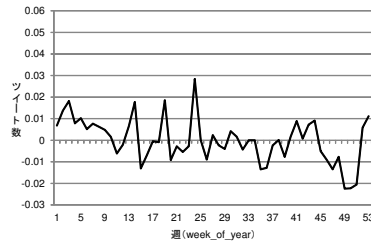


(b) 金閣寺

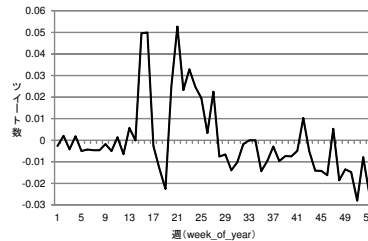


(c) 哲学の道

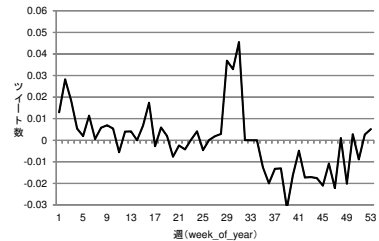
図 6 各観光スポットに対応付けられた位置情報付きツイート集合



(a) 金閣寺



(b) 哲学の道



(c) 八坂神社

図 7 各観光スポットの *week_of_year* に関する時間的特徴

れていることがわかる．実際に，この時期に抽出された特徴語句に着目すると，「金閣の雪景色」という語句がみられた．

(b) 哲学の道に関しては，4 月（第 15 週から第 16 週まで）および 6 月（第 20 週から第 27 週まで）において，多くのツイートが発信されていた．それぞれの時期における特徴語句に着目すると，4 月の時期には「さくら」や「さくら満開」，6 月の時期には「紫陽花」が抽出されていた．

(c) 八坂神社に関しては，7 月（第 29 週から第 31 週まで）において，多くのツイートが発信されていた．実際にこの時期は祇園祭が開催されていた時期であり，この時期における特徴語句に着目すると，「祇園祭観覧」や「御幸祭の神事」，「祇園祭神輿渡御」など祇園祭に関する語句が抽出されていた．

以上のように，観光スポットの時間的特徴および特徴語句を考慮することにより，時節に応じた観光スポットを推薦することが可能となる．

7. ま と め

本稿では，OC-SVM により，対象スポットの実質的な活動領域を推定し，観光スポットと位置情報付きツイートを対応付ける手法を提案した．また，対応付けられたツイートを基に，観光スポットの時間的特徴および特徴語句を抽出する手法を提案した．京都における実データを用いた定性分析により，提案手法の有効性を示した．今後は，抽出された観光スポットの特徴に基づいた，観光スポット推薦システムの設計を検討する．

謝 辞

本研究の一部は，文部科学省科学研究費補助金若手研究（B）「コンテキスト限定価値を考慮した情報推薦方式」（研究代表者：奥健太，課題番号 23700132）による．ここに記して謝意

を表します．

文 献

- [1] 川村秀憲，鈴木恵二，山本雅人，松原仁．新しい〇〇情報学：2. 観光情報学．情報処理，Vol. 51, No. 6, pp. 642–648, 2010.
- [2] 高畠泰斗，香田正人．1 クラス SVM と近傍サポートによる領域判別（＜特集＞ SVM の周辺：One-Class SVM と領域判別）．オペレーションズ・リサーチ：経営の科学，Vol. 51, No. 11, pp. 677–682, 2006.
- [3] 木俣豊，是津耕司，河合由起子，水口充，宮森恒，柏岡秀紀．4. ライフログに基づく実世界でのコンテンツ利活用（＜特集＞ライフログ）．情報処理，Vol. 50, No. 7, pp. 613–623, 2009.
- [4] David Crandall, Lars Backstrom, Daniel Huttenlocher, and Jon Kleinberg. Mapping the World's Photos. In *WWW '09: Proceedings of the 18th international conference on World wide web*, 2009.
- [5] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-ying Ma. Mining Interesting Locations and Travel Sequences from GPS Trajectories. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pp. 791–800, 2009.
- [6] Vincent W Zheng, Yu Zheng, Xing Xie, and Qiang Yang. Collaborative Location and Activity Recommendations with GPS History Data. In *WWW '10: Proceedings of the 19th international conference on World wide web*, pp. 1029–1038, 2010.
- [7] 山中努，田中祐也，土方嘉徳，西田正吾．時空間情報を伴うテキストデータを用いた状況把握支援システム．知能と情報（日本知能情報ファジィ学会誌），Vol. 22, No. 6, pp. 691–706, 2010.
- [8] Ryoung Lee and Kazutoshi Sumiya. Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks - LBSN '10*, pp. 1–10, New York, New York, USA, 2010. ACM Press.
- [9] 李龍，若宮翔子，角谷和俊．Tweet 分析による群衆行動を用いた地域特徴抽出．情報処理学会論文誌：データベース（TOD），Vol. 5, No. 2, pp. 36–52, 2012.
- [10] 溝尾良隆（著），日本観光研究会（監修）．観光光学全集＜第 1 巻＞観光光学の基礎．原書房，2009.