穴場ワードを含むYelpのレビューを用いた新たな穴場スポットの発見

田中　拓人† 　石川　博†† 　廣田　雅春†

† 岡山理科大学 総合情報学部 情報科学科　〒 700–0003 岡山県岡山市北区理大町 1–1
†† 首都大学東京 システムデザイン学部 情報通信システムコース　〒 191–0065 東京都日野市旭が丘 6–6
E-mail: †i14i065nt@ous.jp, ††ishikawa-hiroshi@tmu.ac.jp, †††hirota@mis.ous.ac.jp

あらまし　近年，観光客の増加に伴い，Yelp や，ユーザの関心度などの口コミサイトに対する観光スポットに関するレビュー投稿数が増加している．観光スポットには，知名度の度合などによって，穴場と呼ばれる観光スポットが存在する．穴場な観光スポットを抽出するための研究では，観光スポットの知名度などの基準において，観光スポットをスコアリングする研究が多い．しかし，これらの研究では，知名度などによって穴場を表現しているが，レビュアーごとに穴場であると感じる観光スポットは異なると考えられることや，研究ごとに穴場の定義が異なるなど穴場の定義に抽象性がある．そこで，本研究では，Yelp のデータセットから，穴場ワードを含むレビューを抽出し学習することで，穴場に関するレビューを発見し，レビュアーが穴場と判断するスポットを抽出する手法を提案する．本研究では，ある観光スポットの全てのレビューについて穴場に関するものかどうかを穴場に関するレビューを学習した分類器を用いて分類し，その結果に基づいて観光スポットが穴場かどうかを評価する．本研究では，Yelp のデータセットを用いて，穴場ワードを含む分類器の性能を評価する．また，ユーザごとに穴場と感じるスポットが異なることについて，評価実験を行い，分析する．

キーワード　観光情報，SVM，評判情報，テキスト分類

1. はじめに

インターネットの発達に伴い，Yelp(注1) や，TripAdvisor(注2)，楽天トラベル (注3) などの口コミサイトなどを通じて，観光スポットや，レストラン，商業施設などに関するレビューを発信する機会が増えている．ここで，本論文では，観光スポットなどをまとめて Venue と呼ぶ．Venue に関するレビューは，その Venue の情報や，レビュアーの Venue に対する感想などが記述されている．そして，それらのレビューは，Venue についての情報の獲得や，旅行計画を立てる際，観光ルートを決定する際などに用いられている．

Venue の中には，”穴場”と呼ばれる Venue がある．穴場な Venue (以下，穴場スポット) は，観光客の満足度の向上や，リピータの獲得に繋がると考えられる．そのため，これまでに，穴場スポットを発見するための様々な手法や，その結果を観光客に推薦するための研究が行われている．

これまでの研究において，穴場について，研究ごとに様々な定義がなされている．たとえば，西脇ら [7] は，知名度は低いが他の地点と同等以上の満足が得られる地点と定義している．Chenyi ら [3] は，あまり知られていないが，まだ訪れる価値のある場所と定義している．片山ら [11] は，大衆的には知名度はあまり高くないが，観光客による満足度が高いスポットを隠れスポットと定義している．櫻川ら [12] は，イベントの中心地では見ることのできない景観を撮影できる地点と定義している．島田ら [9] は，必ずしもガイドブック類には紹介されていないような観光スポットと定義している．これらにおいて，穴場の定義は，知名度が低いという点は多くの研究において共通しているが，それ以外の点については様々である．

そこで，本研究でも，穴場スポットを抽出するのに取り組むが，本研究では，多くのレビュアーが”穴場”とレビューに記述したスポットを穴場スポットの定義とする．これについて，本研究では，穴場スポットについて定義を直接に与えずに，Venue のレビューに穴場を示す語句 (以下，穴場ワード) が含まれているレビューに着目する．Venue のレビューに穴場ワードが含まれていれば，その Venue が穴場であることを示唆していると考えられる．そして，そのレビューの対象の Venue は，穴場スポットであると考えられる．

本研究では，この考え方を用いて，レビューが与えられている Venue から，穴場スポットを抽出する手法を提案する．しかし，本研究では， Venue が穴場スポットかどうかの分類に取り組むが， Venue のレビュー数が十分でない可能性がある．そこで，穴場ワードを含まないレビューではあるが，穴場について記述していると考えられるレビュー (以下，穴場レビュー) を抽出するために，提案手法では，穴場ワードが含まれているレビューを学習することで，穴場レビューかそうでないかを分類するための分類器を作成する．そして，その分類器を用いて，Venue のレビューを分類させることにより， Venue が穴場かどうかを分類する．

また，レビュアーは Venue に対し様々なレビューを投稿している．これは，レビュアーごとに Venue への評価基準が異なるためと考えられるので，本研究では，レビュアーごとに穴場と感じる Venue は異なると考え，Venue ごとに穴場なレビュー

(注1)：https://www.yelp.com
(注2)：https://www.tripadvisor.com
(注3)：https://travel.rakuten.co.jp/

を投稿するレビュアーが異なることを示す．これにより，穴場の定義を直接に与えずに，穴場ワードを用いて”穴場”について学習することの有効性を検討する．

本論文の構成は，以下のようになっている．第 2 節では，穴場スポットを抽出するための既存手法などの関連研究について述べる．第 3 節では，Yelp のレビューを用いた穴場スポットの抽出手法について述べる．第 4 節では，提案手法による穴場スポットの抽出の性能について実験により示し，考察する．また，ユーザごとに穴場と感じる Venue が異なるという仮説についての分析を行う．第 5 節では，本論文で得られた成果をまとめ，今後の課題について述べる．

2. 関 連 研 究

2.1 レビューを用いた観光に関する分析

観光に関する情報を分析するために，ユーザがウェブ上に投稿したレビューをデータとして利用する研究が盛んに行われている．

吉田ら [4] は，分散表現を用いてレビューから生成したベクトルを用いて，ユーザの入力キーワードが表す主観的特徴を考慮した観光スポット検索を行う手法を提案している．山岸ら [6] は，観光スポットの人気度とユーザの行動から，ユーザ観光行動の確率モデルを生成する手法を提案している．阪井ら [5] は，耳よりキーワードを含むレビューを文単位でクラスタリングし，そのクラスタの中心ベクトルを構成する文との類似度がある程度低い文を耳より情報として抽出する手法を提案している．

これらの研究では，レビューを観光情報を分析するための有益な情報として扱っている．本研究でも，同様に，レビューを扱い，穴場という観光の満足度などの影響する要因についての分析を行う．

2.2 穴場スポットの発見

穴場スポットを発見，抽出する研究では，Flickr(注4) などにアップロードされている緯度経度情報が付与されている写真が用いられることが多い．

Chenyi ら [3] は，観光に関するキーワードのセットから，緯度経度情報が付与された画像を検索し，クラスタリングを適用することで穴場スポットを発見する手法を提案した．櫻川ら [12] は，写真に付与された緯度経度情報，テキストタグ，撮影時間を用いて，ある地域で行われるイベントの穴場スポットを発見する手法を提案した．島田ら [9] は，観光客が撮影した写真の緯度経度情報に基づいて，撮影者の特徴を分析することで，穴場な撮影スポットを訪れた観光客の特徴を抽出する手法を提案した．西脇ら [7] は，写真に付与された位置情報に基づいてクラスタリングを適用し，得られたクラスタごとにお気に入り数と写真数を用いて，穴場スポット度を算出することによって，知名度に対する満足度が高い地点を抽出する手法を提案した．片山ら [11] は，知名度と満足度の 2 つの指標を用いて，穴場スポットを抽出する手法を提案した．青山ら [8] は，写真の緯度経度情報と撮影日時を用いて，場所に対する人々の知名度と興味の度合いを算出し，それらに基づいて出発地から目的地までの移動中に立ち寄ることが可能な場所を発見する手法を提案した．藤井ら [10] は，写真に付与されている緯度経度情報，テキストタグ，撮影時間を用いて，意外な写真を撮影可能な撮影条件を発見する手法を提案している．

本研究では，たとえば，知名度や，興味など，研究ごとに異なる穴場の観点を包括的に扱うために，穴場ワードを含むレビューを学習させた穴場スポットの分類器を用いることで，穴場スポットを発見する．そのため，レビュアーが穴場であるという認識を持つ観光スポットを抽出する点に本研究の特徴がある．

(注4)：https://www.flickr.com

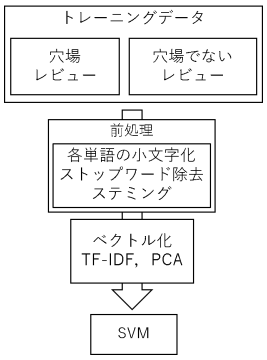

図 1　穴場レビューの分類器作成手順

表 1　穴場ワード

secret grate spot	secret grate place
kept secret place	kept secret spot
little known hot spot	secret spot
little known hot place	best kept secret
secret place	

3. 提 案 手 法

本研究では，Yelp のレビューから穴場ワードを含むレビューを抽出し，穴場かどうかを学習する．そして，観光スポットの全てのレビューに対してその分類器を適用し，その観光スポットが穴場スポットかを分類する．

3.1 穴場ワード

本研究では，穴場ワードを含むレビューから穴場スポットを学習するために，穴場ワードを定義する必要がある．本論文で扱う穴場ワードを表 1 に示す．穴場ワードの選択基準は，穴場を表現すると思われる英語の語句を選択し，その中から，穴場以外の意味を持たないと考えられる表現を人手により決定している．

3.2 前　処　理

本節では，レビューに対する前処理について述べる．本研究では，英語のレビューのテキストに対して，前処理として，各単語の小文字化，ストップワードの除去，ステミングを適用する．ストップワードの除去には，多くの文章で一般的に用いられる，”the”，”and”などの単語を 319 語指定したストップワードリストを用いる．

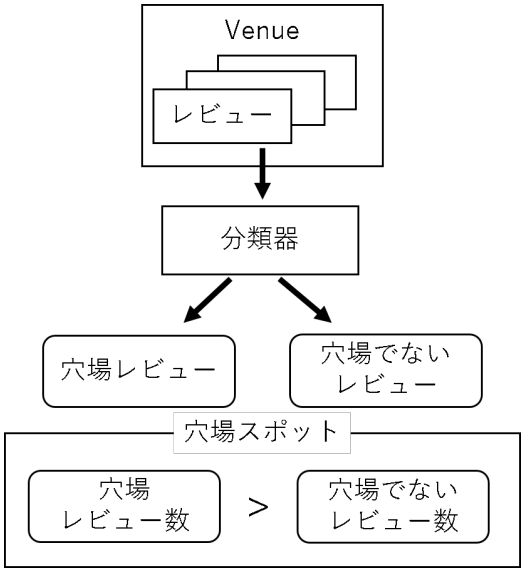

図 2　穴場スポットの発見手法

表 2　穴場レビューの分類器の分類結果

	Precision	Recall	F 値	Accuracy
穴場ワード有	0.92	0.73	0.81	0.98
穴場ワード無	0.96	0.99	0.98	
平均	0.95	0.96	0.95	

3.3　ベクトル化

本節では，レビューをベクトル化する方法について述べる．はじめに，3.2 節で前処理を行ったトレーニングデータのテキストに対して，TF-IDF を適用する．そして，TF-IDF の次元数を削減するために，PCA (Principal Component Analysis) [2] を適用し，ベクトルの標準化も行う．この全ての処理を適用した結果を用いて，以後の穴場ワードの学習を行う．

3.4　穴場レビューの分類器の作成

穴場レビューの分類器を作成する流れを図 1 に示す．穴場ワードを含まないが，穴場について書かれたレビューを発見するために，穴場ワードを含むレビューから穴場について学習する．そのため，穴場ワードを含むレビューと穴場ワードを含まないレビューから，レビューが穴場レビューかそうでないかの 2 値分類を SVM (Support Vector Machine) [1] を用いて行う．あらかじめ穴場ワードを含むレビュー，穴場ワードを含まないレビューを収集する．本節で使用する穴場ワードを含むレビューは，レビュー内容が穴場スポットを示していることを人手により分類したものである．そして，収集したレビューをトレーニングデータとして SVM で学習させ，穴場レビューの分類器を作成する．

3.5　穴場スポットの発見

次に，この穴場レビューの分類器を用いて，穴場スポットの発見を行う．図 2 に，穴場スポットの発見手法について示す．観光スポットの全てのレビューに対してこの分類器を適用し，穴場と穴場でないと分類されたレビューの数を求める．その結果，穴場と分類された数が多い観光スポットを穴場スポットとする．

4.　評 価 実 験

本節では，Yelp のデータセットを用いて，穴場レビューの分類器を作成し分類性能を評価する．また，穴場ワードを含まないレビューからも穴場レビューを抽出し，穴場スポットを発見し，その結果を考察する．

4.1　データセット

本論文では， Yelp Dataset Challenge (round 9)(注5) を用いる．このデータには，144,072 件の Venue と，4,153,150 件のレビューが含まれている．本論文で用いる穴場ワードを含むレビューは，1,978 件である．

4.2　穴場レビューの分類器の作成

本節では，穴場レビューの分類器を作成する手順について述べる．トレーニングデータとして，穴場ワードを含むレビューでかつ，人手によりそのレビューが穴場について記述されていることを確認したレビュー 140 件と，穴場ワードを含まないレビュー 1,000 件を SVM に学習させる．本実験では，SVM のカネール関数にガウスカーネルを用いる．また，SVM のハイパーパラメータである c と γ は，5 交差検定によるグリッドサーチを用い，F 値が最も高いものを用いる．その際に，PCA の次元数は，100 とした．分類性能を評価する為，評価指標として，Accuracy，Precision，Recall，および F 値を用いる．

表 2 に，前述した手順で作成した穴場レビューの分類器の評価結果を示す．表 2 において，穴場ワード有は穴場ワードを含むレビューを，穴場ワード無は穴場ワードを含まないレビューを示している．表 2 より，穴場ワード有と穴場ワード無の評価値を比較すると，すべての評価指標において，穴場ワード有の方が低い値となっている．特に，Recall は，値の差が大きい．これについて，穴場ワード有と穴場ワード無を比較すると，データの割合が不均衡であるため，穴場ワード有のレビューが誤って穴場ワード無と分類された場合が多いためと考えられる．しかし，本研究において，穴場ワードを学習させることで，穴場なレビューを抽出し穴場スポットを発見することが目的である．表 2 より，穴場ワード有の Precision が 0.92 と十分に高いことから，穴場に関して無いレビューを誤って穴場に関するレビューと分類することは少ないことがわかる．今後は，この穴場レビューの分類器を用いて，観光スポットのレビューを穴場かどうかを分類し，穴場スポットを発見することに取り組む．

4.3　穴場スポットの発見

本節では，穴場レビューの分類器を用いて穴場スポットを発見する．ここでは，1 つの Venue に対する全てのレビューについて，穴場レビューの分類器を適用し，穴場レビューと分類されたレビューの割合を求める．

表 3 に，穴場レビューと分類されたレビューの割合の高い Venue の上位 10 件の結果を示す．ここで，Venue 名は Venue の名前ではなく，カテゴリ名とそのカテゴリの通し番号で表している．

表 3 にある 10 件の Venue に対して投稿されたレビューを見

(注5)：https://www.yelp.com/dataset/challenge

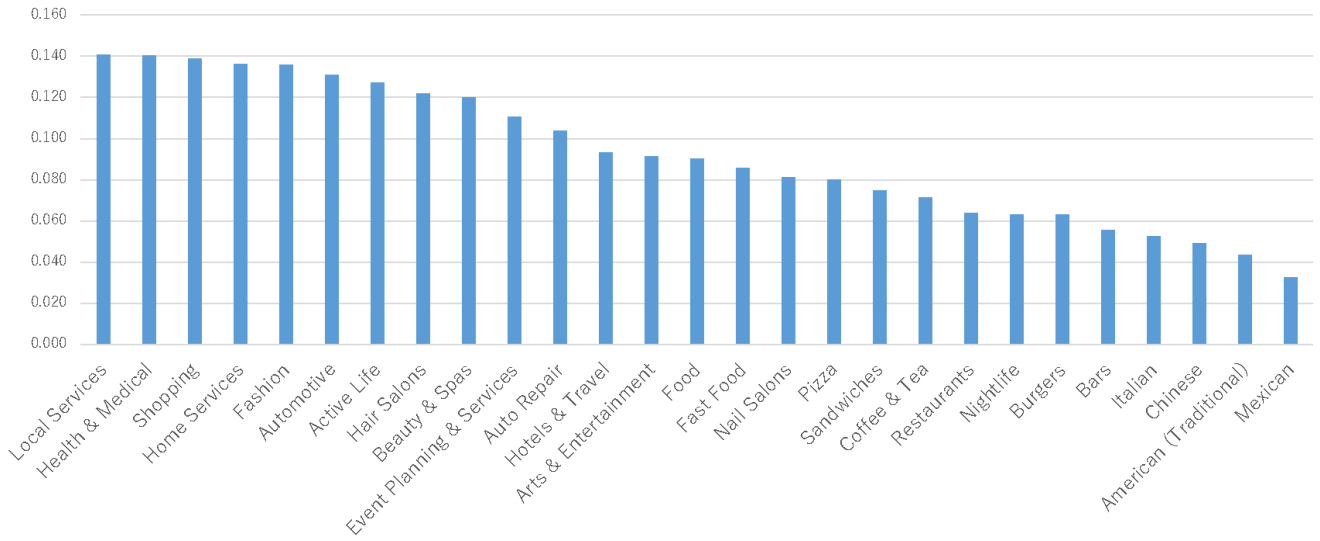

図 3　カテゴリ別の穴場スポットの分析

表 3　穴場レビューの割合が高い Venue の上位 10 件

Venue 名	穴場レビューの数	レビューの総数	穴場レビューの割合
Fashion1	4	5	0.8
Restaurants1	4	5	0.8
Fitness& Instruction1	4	5	0.8
Health& Medical1	4	5	0.8
Shopping1	4	5	0.8
Restaurants2	4	5	0.8
Home Services1	3	4	0.75
Shopping2	3	4	0.75
Beauty&Spas1	3	4	0.75
Restaurants3	3	4	0.75

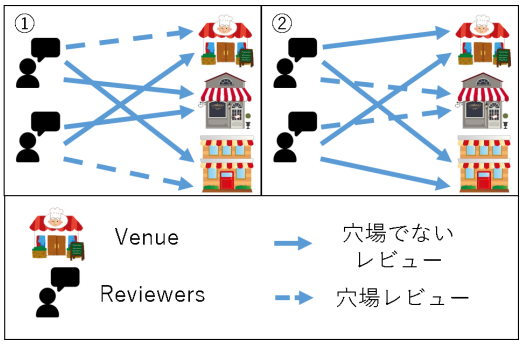

図 4　レビュアーの Venue に対する穴場の評価の違いのパターン

表 4　各パターンの件数とパターン①が全体に占める割合

パターン①	50
パターン②	883
①/(①+②)	0.053

ると，”初めて知った”，”アクセスしづらいがサービスが良かった”などが投稿されており，穴場に関連すると思われる語句が多く含まれていることを確認した．

4.4　カテゴリ別の穴場スポットの分析

本節では，Venue のカテゴリ別に穴場スポットの分析を行う．ここでは，穴場レビューの割合が 50%以上の Venue の カテゴリに注目し，カテゴリごとに何件が穴場スポットと分類されたかの割合を求める．図 3 に分析結果を示す．図 3 において，縦軸はカテゴリごとの穴場スポットと分類された Venue の全体に占める割合を，横軸はカテゴリに所属する Venue の件数が 1,000 件以上のカテゴリ名を示している．図 3 より，割合の高い上位 8 件を見ると日常生活で使用されるカテゴリの割合が高くなっている．これは，カテゴリ”Restaurants”や”Nightlife”のように，ユーザが人気のある Venue に集まる場合とは異なり，ユーザが個々にお気に入りの Venue を利用していると思われる．その結果，カテゴリに所属する Venue でレビューが分散し，穴場スポットとして判断された Venue の割合が高くなったと考えられる．

4.5　レビュアーごとに穴場と評価する Veune の違いの分析

本節では，1 節で述べたレビュアーごとに穴場であると感じる Venue の違いを分析する．これまでの穴場スポットの発見，抽出を行う研究では，様々な穴場の定義がなされているが，研究ごとに穴場の定義が異なるなど穴場の定義に抽象性がある．そこで，Venue ごとに穴場なレビューを投稿するレビュアーが異なることを示すことにより，穴場の定義を直接与えずに，穴場ワードを用いて”穴場”について学習することの有効性を検討する．

本研究では，穴場スポットについて，2 名のレビュアーが，2 つの Venue のペアに対して共通してレビューを投稿している場合に着目する．はじめに，図 4 のようなパターンを考える．図 4 のパターン①は，2 名のレビュアーが 3 つの Venue に対して穴場レビューと穴場でないレビューをそれぞれ別の Venue に投稿している場合である．これは，レビュアーが穴場と感じた Venue が異なる場合を表している．また，パターン②は，2 名のレビュアーが 3 つの Venue に対して投稿した穴場レビューと穴場でないレビューがいずれも重複している場合である．こ

れは，レビュアーが穴場と感じた Venue が同じである場合を表している．この場合，穴場スポットについて，パターン①が多い場合，レビュアーが穴場と感じるスポットが異なると言える．

本実験の手順を述べる．4.3 節において発見した穴場スポットについて，2 名のレビュアーが共通してレビューを投稿している穴場スポットを抽出する．そして，その Venue に対するそれらのレビュアーのレビューに 4.2 節において作成した穴場レビューの分類器を適用して，穴場レビューかそうでないかを分類する．その結果を図 4 のパターンに当てはめる．1,278 件の穴場スポットについて，パターンを求めてその中でパターン①の割合を求める．

本実験では，穴場スポットを 1,278 件を用いた．そして，その中で，前述した条件に該当するレビュアーは 696 人であった．表 4 に実験結果を示す．表 4 より，パターン①が全体の約 5.3%であった．結果として，レビュアーが穴場と感じるスポットが必ずしも同じでないことがわかる．そのため，穴場スポットを抽出するために穴場に関する基準を設けずに，穴場ワードを含むレビューを用いて穴場を抽象的に扱うというアプローチは，有効な可能性がある．

5. おわりに

本研究では，Yelp に投稿された穴場に関する記述を含むレビューを学習することで，穴場スポットを発見するための手法を提案した．穴場ワードを含むレビューを用いることで，レビュアーが穴場と認識するレビューを基準とした，レビューが穴場かどうかの分類器を作成した．実験結果より，穴場に関するレビューの分類器の評価実験を行い，穴場に関するレビューを発見するために有用であることを示した．作成した分類器に，Venue に対する全レビューを適用することで，穴場レビューの割合が高い Venue を抽出した．また，カテゴリごとに穴場と分類された Venue が何件存在するかを割合で示し，穴場として分類されるカテゴリの分析を行った．さらに，レビュアーごとに穴場と感じる Venue が異なるという仮説を立て，それを検証した．結果として，レビュアーが穴場と感じるスポットが必ずしも同じでないことがわかった．

今後の課題として，都市ごとの穴場スポットの数やカテゴリの種類の分析を行うことがあげられる．都市別に Venue のレビューを，この分類器に適用することで，都市固有の穴場スポットが発見できると考えている．

謝　辞

本研究は，首都大学東京傾斜的研究 (全学分) 学長裁量枠戦略的研究プロジェクト戦略的研究支援枠「ソーシャルビッグデータの分析・応用のための学術基盤の研究」，および JSPS 科研費 16K00157, 16K16158 による．

文　献

[1] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, Vol. 20, No. 3, pp. 273–297, 1995.

[2] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists*, Vol. 2, No. 1, pp. 37 – 52, 1987.

[3] Chenyi Zhuang, Qiang Ma, Xuefeng Liang, and Masatoshi Yoshikawa. Discover "Anaba" Sightseeing Spots Using Social Images. 電子情報通信学会技術研究報告 信学技報, Vol. 113, No. 150, pp. 49–54, 2013.

[4] 吉田朋史, 北山大輔, 中島伸介, 角谷和俊. ユーザレビューの分散表現を用いた主観的特徴の意味演算による観光スポット検索システム．第 9 回データ工学と情報マネジメントに関するフォーラム, 2017.

[5] 阪井奎伍, 灘本明代. 観光を対象としたレビューからの耳より情報抽出. 研究報告データベースシステム (DBS), Vol. 2015, No. 13, pp. 1–6, 2015.

[6] 山岸祐己, 斉藤和巳. 観光レビューデータから構築した確率ネットワークによる地域分析. 第 8 回データ工学と情報マネジメントに関するフォーラム, 2016.

[7] 西脇達也, 北山大輔. 写真共有サイトを用いた穴場スポットの抽出. 第 7 回データ工学と情報マネジメントに関するフォーラム, 2015.

[8] 青山賢, 廣田雅春, 石川博, 横山昌平. ジオタグ付き写真を用いた知名度が低いにもかかわらず興味の度合いが高い寄り道候補の発見. 第 7 回データ工学と情報マネジメントに関するフォーラム, 2015.

[9] 島田美優, 伊藤恵. 写真情報分析による観光客特徴集出の試み. 観光情報学会 第 16 回 研究発表会, 2017.

[10] 藤井慎太郎, 加藤大受, 遠藤雅樹, 荘司慶行, 廣田雅春, 石川博. ジオタグ付き写真を用いた意外な写真を撮影できる外れ値的な撮影条件の発見. 第 9 回データ工学と情報マネジメントに関するフォーラム, 2017.

[11] 片山晋, 礒川直大, 小渕幹夫, 西山勇毅, 大越匡, 米澤拓郎, 中澤仁, 高汐一紀, 徳田英幸. Spotrip: 観光客リピータ化促進のための隠れスポット情報提供システムの評価. 電子情報通信学会技術研究報告 信学技報, Vol. 116, No. 508, pp. 157–164, 2017.

[12] 櫻川直洋, 廣田雅春, 石川博, 横山昌平. ジオタグ付き写真を用いたイベントとその穴場スポットの発見. 第 8 回データ工学と情報マネジメントに関するフォーラム, 2016.