ISS-SP-036

Flickr を利用した撮影スポットの分析

陳　嘉穎　　新妻　弘崇　　太田　学
岡山大学大学院自然科学研究科

1. はじめに

ソーシャルメディアサイトに投稿されたジオタグ付き写真から撮影スポットを抽出する研究は数多く行われている．本稿では，Flickr[1]に投稿された写真に付与されている緯度経度情報をクラスタリングする．また各写真の url から閲覧数，お気に入り数とコメント数を抽出し，撮影スポットについて考察する．

2. YFCC100M のクラスタリング

YFCC100M は，写真投稿サイト Flickr に投稿されたおよそ 1 億枚のラベル付き写真と 80 万個の動画のデータセットのことである[2]．本稿では，京都市内の撮影スポットを分析するため，YFCC100M のメタデータから，京都駅を中心に半径 8 キロの範囲内で撮られた写真 200 件の緯度経度情報と写真の url を抽出した．そしてクラスタ数を 70 に設定した k-means 法を用いて，緯度経度情報をクラスタリングした．クラスタリング結果を Google Map 上で表示するために，Google Maps API[3]を用いた．その結果の一部を図 1 に示す．赤いマーカの数値はクラスタ番号，黒い丸はそのクラスタの重心の位置である．

生成されるクラスタは，それぞれ 1 つの観光スポットに対応しているのが理想的な状態である．よってこの結果が理想的な状態にどれだけ近いかを以下の手順で評価した．まず 70 のクラスタに観光スポット名の割り当てを試みた．この割り当てができなかったクラスタは評価から除外し，70 のクラスタは 56 となった．6 件の写真は見られなかったため，ノイズとして除外した．また同じユーザが同じ緯度経度で複数の写真を投稿している場合は 1 件に統合した．これらの処理の結果 200 件のメタデータは 143 件になった．この 143 件のうち 110 件が理想的なクラスタに属していた．

3. 撮影スポットの分析

理想的なクラスタに割り当てられた 110 件に対して，Web スクレイピングを用いて，各写真の url から閲覧数，お気に入り数とコメント数を抽出し，次のような人気度を用いて分析する．また，同じユーザが同じ場所で複数の写真を撮る場合には，その複数の写真の閲覧数，お気に入り数とコメント数の平均を利用して人気度を計算する．

人気度=log(閲覧数)+お気に入り数+コメント数

撮影スポットの写真数が多い上位 10 件を図 2 に示す．ただし，平均閲覧回数は大きいため，1/10 の値である．

図1. クラスタリング結果の一部

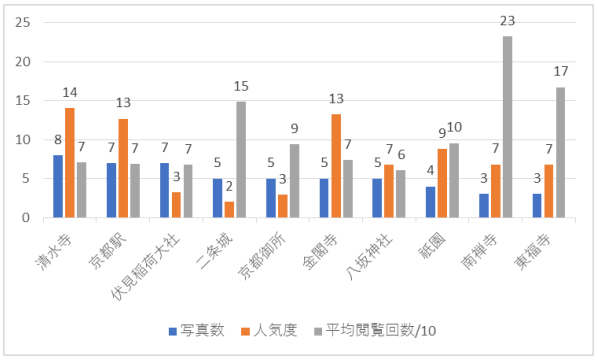

図 2. 写真数が多い上位 10 件の撮影スポット

図 2 において，「南禅寺」と「東福寺」の写真数が少ないが，平均閲覧回数は，この 10 件中 1 位と 2 位である．また，写真数の多い「伏見稲荷大社」，「二条城」，「京都御所」より人気度も高い．一般的に知名度が高いスポットで多くの人々が写真を撮るため，写真数の多いスポットは知名度の高いスポットといえる．したがって，「南禅寺」と「東福寺」は「伏見稲荷大社」，「二条城」，「京都御所」より知名度が低いが，人気があるといえる．

4. まとめと今後の課題

京都駅周辺で撮られた写真の緯度経度情報をクラスタリングして，その結果を分析した．今後データを増やした実験を行う予定である．また，知名度が低くても，人気がある撮影スポットを発見する方法について検討したい．

参考文献

[1] Flickr, http://www.flickr.com/
[2]“1 億枚ラベル付き画像データセット Yahoo Flickr Creative Commons 100M(YFCC100M) を使う”，Qiita，2017-4-18，http://qiita.com/_akisato/items/66deb481ea3cedf388fa
[3] Google Maps API, https://developers.google.com/maps/