ソーシャルデータを用いた季節感のある用語の自動検出

難波 英嗣，竹澤 寿幸
広島市立大学大学院情報科学研究科
{nanba, takezawa}@hiroshima-cu.ac.jp

概要：旅先の見どころには，神社仏閣や遺跡などのように，その場所に行けばいつでも見ることができるものと，花見や花火や犬ぞり体験などのように，限られた季節でなければ見る/経験することができないものに分けられる．本研究では，ソーシャルデータを用いて，後者に関する情報，すなわち季節感のある用語を自動的に検出する手法を提案する．
Keyword：季節感，ジオタグ，機械学習，素性選択

1. はじめに

旅先の見どころには，神社仏閣や遺跡などのように，その場所に行けばいつでも見ることができるものと，花見や花火や犬ぞり体験などのように，限られた季節でなければ見る/経験することができないものに分けられる．本研究では，ソーシャルデータを用いて，後者に関する情報，すなわち季節感のある用語を自動的に検出する手法を提案する．期間限定の旅の見どころの有無は，旅行者が計画を立てる上で重要な情報となりうるが，こうした情報を網羅的，体系的に集める試みは，著者が知る限りこれまでに存在しない．

「限られた季節でなければ見る/経験することができない旅の見どころ」には，(1)日本における夏の花火のように特定の地域においてのみ季節感があるものと，(2)毎年 11 月に解禁されるボジョレーヌーボーに関するイベントや毎年 10 月に世界各地で開催されるオクトーバーフェストのように季節感はあるが地域性があまりないものの 2 種類に分けることができる．また，(1)に関しては，例えば桜前線の移動のように，イベントの発生する時期が地域によって変わるものもある．従って，季節感のある用語を自動的に検出するには，時間的な範囲だけでなく，地域的な範囲も同時に考慮する必要がある．

以上をふまえ，本研究では，ジオタグ付きのソーシャルデータを利用し，季節感のある用語を自動的に抽出する手法を提案する．ジオタグ付きソーシャルデータには様々なものが存在するが，本研究では，画像共有サイト Flickr のデータを用いて季節感のある用語の自動検出を試みる．

2. 関連研究

これまでに，ジオタグ付き，あるいは場所情報と紐付いたソーシャルデータを用いた知識発見に関する数多くの研究が行われている．例えば，Sakaki ら[4]は，Twitter を用いて地震の震源地や台風の進路予測を行う手法を提案している．Aramaki ら[2]は，Twitter を用いてインフルエンザの流行を察知する手法を提案している．Sakai ら[3]は，特定地域の魅力を抽出している．これらの研究では，特定地域で急に注目を集める話題を検出することに主眼を置いているのに対し，本研究では，特定地域における話題の周期性に着目し，特定の時期(季節)に注目される話題の検出を目指す．

遠藤ら[1]は，ジオタグ付き Twitter を情報源として，桜や紅葉の見頃を推定する手法を提案している．ジオタグ付きソーシャルデータに時間情報を加えて分析するという点では本研究と共通するが，遠藤らは，桜や紅葉などのトピックが与えられた状況から分析を開始しているのに対し，桜や紅葉のような季節感のある用語そのものの検出を試みる点が遠藤らの研究と異なる．

3. ジオタグ付きソーシャルデータからの季節感のある用語の自動検出

画像共有サイトのひとつである Flickr (www.flickr.com)では，このサイトにユーザが画像を投稿する際，その画像の内容を示す 1 つ以上の単語をタグとして付与する．投稿された画像の一部には，その画像を撮影した場所(ジオタグ)や撮影日時も含まれていることがある．これらの情報を用いれば，前節で述べた季節感のある用語の自動検出が実現可能になる．

ここで，ある地理的範囲内のジオタグおよびタグ付き画像が一定数あった場合に，季節感のある用語を自動検出する手順を，図 1 に示すデータを例に用いて説明する．

月	画像に付与されたタグ
1	snow, cold
1	snow, ski
1	snow, resort
2	snow, freezing
2	snowman, cold
3	plum, beautiful
4	cherry, blossom, bloom, april
4	cherry, blossom, bloom, beautiful
4	cherry, blossom, bloom, plum
5	may, fresh
5	may, bloom
12	christmas, present, home
12	christmas, party

図 1　ある地域で投稿されたジオタグおよびタグ付きデータの例

図 1 は、13 件の画像が投稿された月および各画像に付与されているタグを示している．本研究の目的は，図 1 のようなデータが与えられた時，例えば 4 月であれば，4 月に集中的に現れる“cherry,” “blossom,” “bloom”といった語を，季節感のある語と考えて抽出することである．

季節感のある語を抽出するには，様々な方法が考えられる．例えば，月ごとにタグを集計し，出現頻度の高いものを，その月の季節感のある語とする方法である．しかし，この方法の場合，どの月にも満遍なく高頻度で出現する語も，季節感のある語として抽出されてしまうことになる．そこで，1 つの月にだけ着目するのではなく，ある月には頻出するが，別の月にはほとんど出現しない語を優先的に抽出する必要がある．

本研究では，ひとつの画像に付与された複数のタグから，その画像が撮影された月を推定するのに有用な単語を，季節感のある用語として抽出する．例えば，図 1 の 1 件目の画像の場合，この画像に付与された 2 つのタグ“snow”と“cold”から“1 月”を推定するような分類器を構築し，もし“snow”という単語が“1 月”を推定するのに有用であると判断されれば，“snow”を季節感のある語とみなす．これは，文書分類における有効な素性(単語)を選択する課題と基本的に同じである．一般に，素性選択には自己相互情報量や情報利得などの統計的な手法が用いられるが，本研究でもこれらの手法を適用する．これに加え，図 1 のデータを訓練用データと考えてランダムフォレストを用いて分類器を構築し，各タグの特徴量としての重要度を計算し，重要度の高いタグを季節感のある語として抽出する．なお，図 1 のデータに対し，ランダムフォレストで季節感のある用語(4 月)を抽出すると，図 2 の結果が得られる．タグと共に記載されている数値は重要度を示している．

重要度	タグ
0.378585858586	cherry
0.267166666667	blossom
0.241785714286	bloom
0.0606060606061	beautiful
0.0385223665224	snow
0.0133333333333	plum

図 2　図 1 のサンプルに対しランダムフォレストを用いて季節感のある語(4 月)を抽出した結果

なお，本研究では Flickr データを利用して提案手法の検証を行うが，例えばジオタグ付き Twitter データがあれば，まったく同様の手順で季節感のある用語を抽出することができる．

4. 実験

提案手法の有効性を確認するため，実験を行った．

実験データ

Flickr データを研究用に提供されている YFCC100M データセット[5]を用いた．このデータ

には約 1 億件の画像のメタデータから構成されているが，このうち約 1/4 の画像にはジオタグ情報が付与されている．この中で，北緯 34 度 0 分以上 35 度未満および東経 135 度 0 分以上 136 度未満の範囲にある画像のメタデータ 40,852 件を用いた．各メタデータには，画像の URL，画像が撮影された年月日，人手で付与されたメタタグ，機械的に付与されたメタタグなどが含まれている．本研究では，人手で付与されたメタタグおよび機械的に付与されたメタタグを季節感のある語の候補とし，各画像の撮影月を用いて季節感のある語を選定する．なお，月別の画像データの内訳を表 1 に示す．

表 1　季節感のある語の抽出実験に用いた画像データの月別件数

月	件数	月	件数
1 月	2,366	7 月	3,513
2 月	1,376	8 月	3,021
3 月	5,082	9 月	3,047
4 月	4,906	10 月	3,234
5 月	5,446	11 月	3,345
6 月	2,870	12 月	2,646

計 40,852 件

比較手法

以下の 3 種類の手法で比較を行う．

- ランダムフォレスト(提案手法)
- 自己相互情報量(ベースライン手法)
- 月別出現頻度(ベースライン手法)

評価方法

4 月，8 月，12 月について各手法で抽出された用語を nDCG@10 で評価する．

実験結果

結果を表 2 に示す．表からわかるとおり，ランダムフォレスト(提案手法)の結果がベースライン手法よりも非常に高い値となった．表 3 に 3 手法による抽出結果(4 月)を示す．表において，正解と判定したものは太字で示している．

表 2　3 手法による抽出精度の比較

	ランダムフォレスト	自己相互情報量	月別出現頻度
nDCG@10	0.916	0.167	0.168

表 3　3 手法による抽出結果の比較(4 月)

順位	ランダムフォレスト	自己相互情報量	月別出現頻度
1	**sakura**	naniwa	japan
2	**cherry**	scenes	osaka
3	osa	fall	kyoto
4	**blossoms**	tenjin	**cherry**
5	airport	club	square
6	osaka	nanba	temple
7	japan	march	asia
8	britton	mount	kansai
9	umeda	hommachi	**blossoms**
10	**april**	neighbourhood	app

表からわかるとおり，ランダムフォレストは上位に正解を抽出することができている．月別出現頻度は，どの月の結果を見ても上位に“japan”や“osaka”が出現しており，全体的にランダムフォレストよりも抽出精度が低くなっている．逆に捉えれば，ランダムフォレストは，他の月にも頻出する用語は重要度を下げる効果があることが確認できる．

他の月に頻出する用語の重要度を下げるという点では，自己相互情報量の上位には“japan”や“osaka”が入っていないので一定の効果があったと言える．しかしながら，極端に出現頻度の低い用語は除外するなど，自己相互情報量の結果の改善につながると思われるフィルタリングなどの処理を適用しても，結果に改善が見られなかった．

5.　季節感のある旅行ブログエントリの検索

4 節でランダムフォレストにより獲得した用語リストを用い，季節感のある旅行ブログエントリを検索した．月ごとに得られた用語の上位 5 語および各語の重みベクトルと類似度の高い旅行ブログエントリを，同じ地理範囲内で同じ月のブログから検索した．検索対象のブログエントリは，TravelBlog(www.travelblog.org)のものを用いた．TravelBlog の各エントリには緯度経度情報は付与されていないが，各ブログエントリ中の画像を，Google cloud vision API を用いて解析し，ジオタグを付与することができた 24,040 件を実験に用いた．

実験の結果，以下のような季節感のある旅行ブログエントリが検索された．なお，カッコ内の単語は，検索クエリの中でブログエントリ本文中に出現した単語を示す．

4 月：

- 桜が満開の大阪城を訪れる．（cherry & blossoms）

https://www.travelblog.org/Asia/Japan/Osaka/Osaka/blog-779237.html

- 大阪城周辺で花見．桜茶を飲む．（cherry & blossoms）

https://www.travelblog.org/Asia/Japan/Osaka/Hirakata/blog-52850.html

7 月

- 京都祇園祭（matsuri）

https://www.travelblog.org/Asia/Japan/Kyoto/Kyoto/blog-796573.html

11 月

- 南禅寺と永観堂のもみじ（fall）

https://www.travelblog.org/Asia/Japan/Osaka/blog-766488.html

6. おわりに

本研究では，ジオタグ付きソーシャルデータのひとつである YFCC100M データセットを用い，季節感のある用語の自動検出を行った．実験の結果，ランダムフォレストを用いた場合に高い精度で季節感のある用語を検出できることがわかった．

参考文献

[1] 遠藤雅樹，三富恵佑，佐伯圭介，江原遥，廣田雅春，大野成義，石川博：ツイートを用いた生物季節観測の見頃推定手法による情報提供の検討，観光と情報，第 12 巻，第 1 号，pp. 47-60，2016.

[2] Eiji Aramaki, Sachiko Masukawa, and Mizuki Morita: Twitter Catches the Flu: Detecting Influenza Epidemics using Twitter, Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp.1568-1576, 2011.

[3] Tatsuhiro Sakai, Keiichi Tamura, and Hajime Kitakami: Extracting Attractive Local-Area Topics in Georeferenced Documents using a New Density-based Spatial Clustering Algorithm, IAENG International Journal of Computer Science, Vol. 41, No. 3, pp. 185-192, 2014.

[4] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo, Earthquake Shakes Twitter Isers: Real-time Event Detection by Social Sensors", Proceedings of the 19th International Conference on World Wide Web, pp.851-860, 2010.

[5] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li: YFCC100M: the New Data in Multimedia Research, Communications of the ACM, Vol. 59, Issue 2, pp. 64-73, 2016.