人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第21回)
SIG-AM-21-19

宿泊施設のレビューの時系列分析による季節を表す特徴語の抽出

Extracting Feature-Words Depending on Seasons by Time-Aeries Analysis of Web Reviews for Accommodations

佐藤裕次郎　　山西良典　　西原陽子 *

立命館大学情報理工学部
College of Information Science and Engineering, Ritsumeikan University

Abstract: The users sometimes concern some keywords which reprenent the seasonal events/features such as “fireworks” and “colored leaves,” when they look for the accommodations. Such keywords can be also helpful for the accommodation sides; those keywords can be the good advertisements appealing the features of that place. To find those keywords, we propose a time-series analysis of Web reviews for accommodations in this paper. The review for the accommodation is divided into monthly units, and the nouns characteristically appear in the monthly time-series are extracted as the keywords depending on the seasons. Through the evaluation tasks, it was confirmed that the proposed method extracted the nouns appearing in some specific seasons well.

1　はじめに

宿泊施設をユーザが決める際，ウェブレビューが決定要因になることが多い．このとき，特に，宿泊する季節の内容が含まれたレビューを参考にすることがある．例えば,「花火」や「紅葉」などの季節のイベントや，時期によって混雑具合が変わる周辺施設の情報のレビューは，旅行者にとっては充実した滞在をする上で有用な情報となりうる．また，宿泊施設側の観点からも，これらの情報は効果的な広告・宣伝として活用できる．宿泊施設のレビューを対象として，宿泊施設の戦略構築 [1] や宿泊施設サービス改善のための情報抽出 [2, 3] などの研究も報告されている．

ウェブレビューの解析，情報の抽出や比較の研究は数多く存在しており，様々な方法で分析が行われている．中山らは，レビューテキストから条件付き意見の抽出 [4] と評価条件の抽出 [5] を行っている．これらの研究では利用者の宿泊施設の選択の意思決定支援に有益な情報の抽出を目的としている．しかし，同一のレビュー対象についての時系列変化に着目した分析とはなっていない．松尾らは，レビューのユーザの評価の根拠を提示しており [6]，ユーザの評価の根拠の中には本研究で着目する季節によって変化する特徴が含まれる可能性がある．旅行は時期と場所が重要な要素となるアクティビティである．宿泊施設のレビューを対象とする本研究では，ユーザの評価の根拠の中でも特に時期によって変化する特徴の抽出を目指す．乾らはテキストから評価辞書を使用した評価情報の抽出を行っており [7, 8]，評価を記述するもの，要求，提案，認識，印象などの意見の分類を問題としている．

本稿では，レビュー対象としては宿泊施設を扱い，時系列変化する評価情報 [9] の中でも特に一定の季節において盛り上がりを見せるような季節を表す特徴語の抽出を目指す．提案手法によって抽出された単語が，主観評価実験によっても季節を表す単語であるかを評価する．

2　提案手法

提案手法では，ある宿泊施設のレビュー文から季節の特徴を表す単語の抽出を目的としている．単語は名詞とする．レビューデータとして，楽天トラベルのデータ[1]を使用する．宿泊施設に対して書かれたレビューを月ごとにまとめ一つの文書として扱い，12ヶ月分の文書に出現する名詞の 1 ヶ月毎の出現回数を算出する．その出現回数の時系列変化の割合が大きい名詞を，その宿泊施設の季節の特徴をあらわす名詞として抽出する．月単位の文書内に出現する名詞の出現回数を全単語の出現回数で割った tf 値と，全文書数を単語 t の出現する文書数で割り，その値の対数である idf 値を掛け合わせて $tf-idf$ 値を計算する．$tf-idf$ 法を用いるこ

*連絡先：〒 525-0072 滋賀県草津市野路東 1 − 1 − 1
E-mail: {is0309he@ed, ryama@fc, nisihara@fc}.ritsumei.ac.jp

[1]楽天株式会社 (2016): 楽天トラベルデータ. 国立情報学研究所情報学研究データリポジトリ. (データセット). https://doi.org/10.32130/idr.2.2

とにより，出現した名詞のその宿泊施設の特定の月においての重要度が計算できる．また時系列変化の割合が大きいものを出力するために，正規化を行い比較できるように処理する．レビューに出現した名詞の月単位の変化として，12ヶ月分の $tf-idf$ 値の平均値と最大値の差が大きいものから順に出力する．この処理により，特定の季節にのみ $tf-idf$ 値が上昇した名詞が出力される．

処理手順をまとめると，以下のようになる；

1. レビューの分割を行う．月単位でレビューを分割し，1つの宿泊施設に12個の文書が存在するとする．
2. レビューの分かち書きの際に形態素解析器MeCabと辞書にNEologdを用いて出現した名詞を取得する．
3. 出現した名詞の月単位の $tf-idf$ 値を計算する．
4. 解析中の宿泊施設のレビューに1回でも出現した名詞に対し，12ヶ月分の $tf-idf$ が計算され，$tf-idf$ 値の変化の割合の大きい順に名詞を出力するため，正規化を行い比較できるようにする．最大値を0，最小値を1とし，12ヶ月分の $tf-idf$ 値に対し正規化を行う．
5. 正規化された12ヶ月分の $tf-idf$ 値の平均と最大値の差が大きい順に名詞を出力する．
6. 正規化の最大値が1なので，平均の小さい順に並び替え出力する．

2.1 $tf-idf$ 値の計算

$tf-idf$ 値は，その文書内で使われている単語の重要度を表す．対象の文書内での出現頻度が高い単語ほど重要であるという考えに基づく tf 値と，いくつもの文書で横断的に出現する単語は重要でないという考えに基づく idf 値の掛け合わせで $tf-idf$ 値を計算する．

以下，本稿における tf 値と idf 値の計算方法を記す．式(1)に tf 値，式(2)に idf 値の計算式を記す．

$$tf(t,d,month)=\frac{n_{t,d,month}}{\sum_{s\in d} n_{s,d,month}}, \tag{1}$$

$$idf(t,month)=\log\frac{N}{df(t,month)}, \tag{2}$$

ここで，$n_{t,d,month}$ は，ある名詞 t の解析する月 $month$ の文書 d 内での出現回数を示し，$\sum_{s\in d} n_{s,d,month}$ は，解析する月 $month$ の文書 d 内での全ての単語の出現回数を示す．また，N は分析対象とする全文書数を示し，$df(t,month)$ はある単語 t が解析する月 $month$ に出現する文書の数を示す．つまり，df 値を計算する文書集合を12ヶ月のレビューとすることで，各月に特徴的に現れる単語の抽出を図る．

2.2 正規化と単語の抽出

出現した名詞の $tf-idf$ 値の変化の割合を比較するために正規化を行う．データの最大値を1, 最小値を0にする，式(3)に従って正規化を行う．

$$SX=\frac{X-x_{(min)}}{x_{(max)}-x_{(min)}}, \tag{3}$$

ここで，SX，X はそれぞれ正規化後の値と正規化前の値を示し，$x_{(min)}$ と $x_{(max)}$ はそれぞれデータの最小値と最大値を示す．

この正規化された数値グラフから，特定の月においてのみ，その $tf-idf$ の値の割合が大きくなっている名詞を出力する．特定の月における値の上昇を検知するために，12ヶ月分の平均値を参照とする．12ヶ月分の $tf-idf$ の平均値と最大値の差が大きい名詞を宿泊施設の季節を表す特徴語として抽出する．

ただし，以下の条件で分析を行うものとした；

- レビュー数の上位200件の宿泊施設を使用する
  十分にレビュー数が存在するデータを分析対象として使用する
- 数字や，記号など宿泊施設の季節の特徴を明らかに表さない名詞を除外する
- レビュー内で出現回数が9回以下の名詞を除外する
  出現回数の少ない名詞では，相対的に12ヶ月分の $tf-idf$ の平均値と最大値の差が大きくなる傾向にあるが，このような名詞は特定のユーザのみが記述した単語である可能性が高く季節の特徴を表すとは言えない
- 「1月」や「2月」などそれ単体で特定の月を表す名詞はあらかじめ除外する

3 季節を表す特徴語の抽出結果

表1に,「神戸ポートピアホテル」「横浜桜木町ワシントンホテル」「天然温泉　斛の湯 ドーミーイン富山」「ホテル阪急インターナショナル」「グランドホテル浜松」のレビューから，抽出された名詞の上位10件を示す．抽出された名詞のうち、「年末」「お正月」「お盆」「GW」「盆」「クリスマス」といった名詞は，特定の季

表 1: 5 件の宿泊施設のレビューから抽出した名詞の上位 10 件

	神戸ポートピアホテル	横浜桜木町ワシントンホテル	天然温泉　剱の湯 ドーミーイン富山	ホテル阪急インターナショナル	グランドホテル浜松
1 位	ルミナリエ	花火	盆	年末	浜松まつり
2 位	年末	要望	システム	名前	祭り
3 位	お正月	程度	煙草	椅子	ふぐ
4 位	お盆	開港	枕	レート	総合的
5 位	不要	不備	雪	クリスマス	空室
6 位	GW	不自由	風	クーポン	若干
7 位	屋外プール	込み	知人	こと	誕生日
8 位	キャンセル	近隣	一緒	タイミング	今年
9 位	渋滞	電車	朝ごはん	みたい	基本的
10 位	不愉快	ランク	おかげ	パノラマバスルーム	交換

節と時期を明確に表す名詞であることがわかる．また，「ルミナリエ」や「屋外プール」「花火」「浜松まつり」は開催される時期があり，通年で催されるとは考えにくい．そのため、特定の季節のレビューに特徴的に出現すると考えられる名詞の抽出において，一定の有効性が確認されたと考えられる．

4　評価実験

主観評価実験により，抽出された名詞が宿泊施設のレビュー文の特定の季節に特徴的に出現するかどうかを評価した．比較対象として提案手法によってランク付けされた下位 5 件の名詞も評価した．実験は以下の手順で行った;

手順１： 宿泊施設のレビュー数上位 200 件のうち，任意に 5 件の宿泊施設を選択

手順２： その宿泊施設から抽出された名詞の上位 5 件と下位 5 件を選択

手順３： 被験者には上位と下位を伝えず，研究の内容を伝え，抽出された名詞が「特定の季節に特徴的に出現すると思うか」を問う

手順４： 被験者 44 人に答えを「はい」，または「いいえ」の二択で回答してもらう

4.1　実験結果

表 2〜 6 に,「神戸ポートピアホテル」「横浜桜木町ワシントンホテル」「天然温泉　剱の湯 ドーミーイン富山」「ホテル阪急インターナショナル」「グランドホテル浜松」の結果をそれぞれ示す.

表 2: 「神戸ポートピアホテル」についての結果．表中の人数は,「特定の季節に特徴的に出現する単語であるか？」という質問への回答者数.

	名詞	YES	NO
上位 5 件	**ルミナリエ**	26 人	18 人
	年末	44 人	0 人
	お正月	44 人	0 人
	お盆	44 人	0 人
	不要	2 人	42 人
下位 5 件	一流	0 人	44 人
	用意	2 人	42 人
	電話	0 人	44 人
	対応	1 人	43 人
	南館	3 人	41 人

また，図 1 , 2 , 3 , 4 , 5 に，縦軸が $tf-idf$ 値、横軸が $month$ となる月単位の $tf-idf$ 値の推移グラフをそれぞれ示す.

4.2　考察

「宿泊施設の季節の特徴を表す名詞である」と抽出された名詞の 5 つの宿泊施設の上位 5 件ずつ，計 25 件の名詞のうち，被験者の過半数が「宿泊施設のレビューに特定の季節に特徴的に出現する名詞である」と思うと回答したのは 13 件（26 %）であった．この割合が十分に高くない理由として，ホテルチェーンが経営する宿泊施設は季節の特徴がレビューに書かれにくいこ

表 3: 「横浜桜木町ワシントンホテル」についての結果．表中の人数は,「特定の季節に特徴的に出現する単語であるか？」という質問への回答者数.

	名詞	YES	NO
上位 5 件	**花火**	43 人	1 人
	要望	2 人	42 人
	程度	0 人	44 人
	開港	14 人	30 人
	不備	0 人	44 人
下位 5 件	時間	4 人	40 人
	出張	2 人	42 人
	バイキング	3 人	41 人
	朝食	4 人	40 人
	立地	3 人	41 人

表 4: 「天然温泉　剱の湯 ドーミーイン富山」についての結果．表中の人数は,「特定の季節に特徴的に出現する単語であるか？」という質問への回答者数.

	名詞	YES	NO
上位 5 件	**盆**	43 人	1 人
	システム	0 人	44 人
	煙草	2 人	41 人
	枕	0 人	44 人
	雪	44 人	0 人
下位 5 件	料金	10 人	34 人
	フロント	0 人	44 人
	快適	5 人	39 人
	サウナ	11 人	33 人
	バス停	0 人	44 人

表 5: 「ホテル阪急インターナショナル」についての結果．表中の人数は,「特定の季節に特徴的に出現する単語であるか？」という質問への回答者数.

	名詞	YES	NO
上位 5 件	**年末**	43 人	1 人
	椅子	0 人	44 人
	レート	3 人	41 人
	クリスマス	44 人	0 人
	名前	1 人	43 人
下位 5 件	窓	3 人	41 人
	綺麗	9 人	35 人
	友人	3 人	41 人
	夜景	17 人	27 人
	リクエスト	1 人	43 人

表 6: 「グランドホテル浜松」についての結果．表中の人数は,「特定の季節に特徴的に出現する単語であるか？」という質問への回答者数.

	名詞	YES	NO
上位 5 件	**浜松まつり**	38 人	6 人
	祭り	42 人	2 人
	ふぐ	30 人	14 人
	総合的	0 人	44 人
	空室	5 人	39 人
下位 5 件	金額	9 人	35 人
	豪華	1 人	43 人
	デラックス	2 人	42 人
	親切	0 人	44 人
	駅	0 人	44 人

とも関係すると考えられる．評価実験を行った宿泊施設の「神戸ポートピアホテル」以外は全てチェーンの宿泊施設であった．ホテルチェーンが経営する宿泊施設は対象のユーザに年中一定の品質を提供するため季節の名詞が抽出されにくいと考えられる．

また出現回数が低いことによって，誤って季節を表す語句として抽出されている名詞が見られた．出現回数が少なく $tf - idf$ 値が年間を通じて小さい場合，現在の提案手法では抽出される順位が高くなってしまう．例えば,「要望」や「程度」は意味として季節を表す名詞として考えることは難しいが，これらの単語についての全レビューでの総出現回数が 12 回しかなく，出現回数が小さいことによって偶然に出現回数が増加した月があったため，抽出対象として選出されたと考えられる．今後は，評価式の再構成や例外処理の条件を設けることで，抽出の精度向上を目指す．

一方で,「ルミナリエ」や「クリスマス」といった他の宿泊施設に出現していないような名詞や,「年末」や「お正月」といった特定の季節を表す名詞の抽出には成功した．図 1 の「お正月」と図 4 の「クリスマス」を見ると抽出された名詞の $tf - idf$ 値が特定の月や、前後の月で上昇しているのがわかる．このことから特定の季節を表す名詞が，レビューに特徴的に出現する季節の $tf - idf$ 値が高くなっていることがわかる．

現段階では，提案手法によって抽出した名詞すべてについてを，宿泊施設の季節の特徴を表す単語とすることは難しい．抽出された名詞を含むレビュー文を解析することで，本研究の目的に活用できる季節の情報を抽出する可能性を探る．

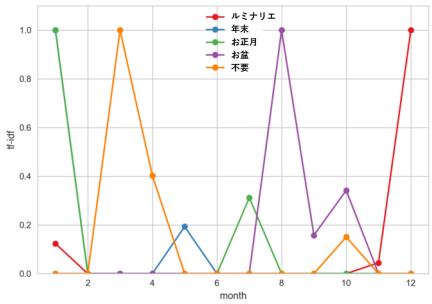

図 1: 「神戸ポートピアホテル」についての tf-idf 値の推移グラフ.

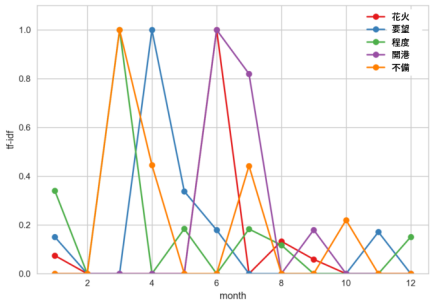

図 2: 「横浜桜木町ワシントンホテル」についての tf-idf 値の推移グラフ.

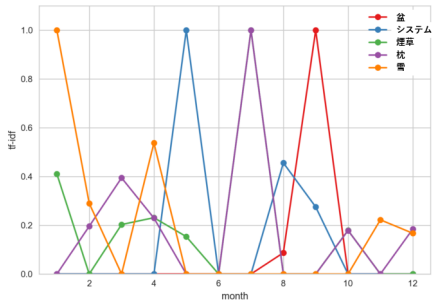

図 3: 「天然温泉　斛の湯　ドーミーイン富山」についての tf-idf 値の推移グラフ.

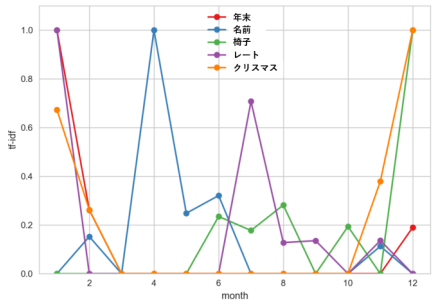

図 4: 「ホテル阪急インターナショナル」についての tf-idf 値の推移グラフ.

5　おわりに

本研究では宿泊施設のレビュー文から，ユーザが宿泊施設の決定要因になりうる情報，または宿泊施設の広告や効果的な宣伝に活用するために，季節の特徴を表す名詞の抽出を行った．レビューを月単位に分割し，$tf - idf$ 値の変化の割合が大きいものを抽出し，宿泊施設のレビューに出現した名詞に，季節の特徴を表しているかどうかの順位をつけた．評価実験の結果，抽出した名詞の中に，宿泊施設の季節の特徴を表していると考えられる名詞があることがわかった．

しかしながら，抽出方法の問題点もいくつか見つかった．今後は，提案手法の問題点の改善とともに，抽出された名詞を含むレビュー文を分析することによって，より詳細な季節の特徴抽出を目指す．また，抽出された情報について，ユーザが宿泊施設の決定要因になりうる情報，または宿泊施設の広告や効果的な宣伝に利用可能な情報といった情報の評価も行っていく．

謝辞

本研究では，国立情報学研究所の IDR データセット提供サービスにより楽天株式会社から提供を受けた「楽天データセット」を利用した．記して謝意を表す．

参考文献

[1] 田邊亘, 後藤正幸. 宿泊施設の戦略構築を支援するユーザレビュー分析に関する一考察. メディアセンタージャーナル, Vol. 9, pp. 91–101, 2008.

[2] 辻井康一, 津田和彦. 宿泊レビューを用いた宿泊施設サービス改善のための情報抽出. 人工知能学会全国大会論文集, pp. ROMBUNNO.2E1–5, 2013.

[3] 辻井康一, 津田和彦. テキストマイニングを用いた宿泊レビューからの注目情報抽出方法. 情報処理学会デジタルプラクティス, Vol. 3, No. 4, pp. 289–296, 2012.

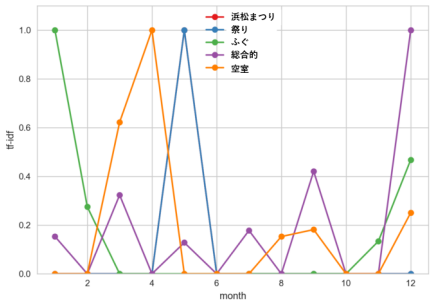

図 5: 「グランドホテル浜松」についての tf-idf 値の推移グラフ.

[4] 中山祐輝, 藤井敦. レビューテキストを用いた条件付き意見文の抽出. 言語処理学会 第 20 回年次大会発表論文集, pp. 888–891, 2014.

[5] 中山祐輝, 藤井敦. レビューテキストを対象とした評価条件の抽出手法. 言語処理学会第 19 回年次大会発表論文集, pp. 248–251, 2013.

[6] 松尾哉太, 新妻弘崇 1, 太田学. レビュー解析に基づくユーザ評価の根拠提示の一手法. 情報処理学会研究報告, Vol. 2014-DBS-160, No. 14, pp. 1–6, 2014.

[7] 乾孝司, 奥村学. テキストを対象とした評価情報の分析に関する研究動向. 自然言語処理, Vol. 13, No. 3, pp. 201–241, 2006.

[8] 乾孝司, 梅澤佑介, 山本幹雄. 評価表現と文脈一貫性を利用した教師データ自動生成によるクレーム検出. 自然言語処理, Vol. 20, No. 5, pp. 683–706, 2013.

[9] 打田裕樹, 吉川大弘, 古橋武, 平尾英司, 井口浩人. Web ユーザレビューにおける評価情報の時系列変化の可視化. 日本知能情報ファジィ学会誌, Vol. 22, No. 3, pp. 377–389, 2010.