

テキストマイニングによるブログ内容の把握可能性に関する研究

山梨大学 工学部土木環境工学科 学生会員 ○山下 宏明
山梨大学 大学院医学工学総合研究部 正会員 佐々木 邦明

1. 研究の背景と目的

これまで土木計画の分野では、アンケートの自由記述や討論など、構造化されていないデータを主な対象にして、テキストマイニングを活用した様々な分析がなされるようになってきた。観光地の景観評価のように、多様な評価軸が存在するような場合、これまでアンケートによって行われてきたが、分析者はあらかじめどういった評価がなされているのかを知る必要があり、適切な評価のための事前調査などが必要となる。一方観光はブログのトピックの中でもポピュラーなもので、観光ブログだけを取り扱った SNS など存在する。その中には観光行動や景観評価などの情報が含まれている。これらの情報をうまく分析することで、観光地における景観の評価がどのように行われているのかを、事前に把握可能であると考えられ、これらの情報を取り出すことが可能かを本研究の対象とする。

2. 本研究で用いるデータ

本研究では分析対象の観光地として湯布院を選択した。湯布院は観光客の増加によって、町の開発が進み本来の風情ある「むら」の景観が失われることが危惧されている。その対策のため平成 9 年には「ゆふいん建設・環境デザイン協議会」¹⁾という集まりも発足し、回遊を促進するための修景等の試みが始まっている。

そこで、先行研究の結果を参考に「湯布院 行った」を検索ワードとして、2014 年 8 月～12 月に Google Blog Search を用いて検索を行ない、約 769,000 件の検索結果を得た。今回はこれらをすべて扱うのではなく、内容を吟味したうえで、適切な分析を行うことが可能であることを検証することを目的としているため、内容を目視で十分検討可能な範囲として検索結果から上位 200 件のブログのデータを収集し用いた。目視によってブログの内容を把握し、その上でコーディングルール等によってこれらの内容を適切に抽出できるかを検討する。

収集したブログより得られた特性を以下にあげていく。

まずブログの記述者が湯布院を訪れた年月について、ブログの検索を行なった年が 2014 年であったためブログエントリが書かれた年は近い年のものが多く、図-1 のような分布となった。月については由布院温泉観光協会²⁾によれば湯布院に訪れる観光客は年末が多いこともあり、図-2 に示すように 12 月が多くなっている。また 6, 9, 10 月も多くなっている。さらにブログ記述者が湯布院へ行った際にそこへ訪れたのは初めてかどうかについても集計した。その結果を図-3 に示す。それによると初来訪の人が多いという結果がでた。由布院温泉観光協会²⁾曰く来訪が初となる理由としては旅館のほとんどが高級で宿泊費等の旅費が高額になるため頻繁に訪れる客が少数になるということが挙げられるそうだ。

キーワード 観光, ブログ, テキストマイニング, 湯布院

連絡先 〒400-8511 山梨県甲府市武田 4-3-11 山梨大学工学部土木環境工学科 TEL. 055-220-8671 E-mail : sasaki@yamanashi.ac.jp

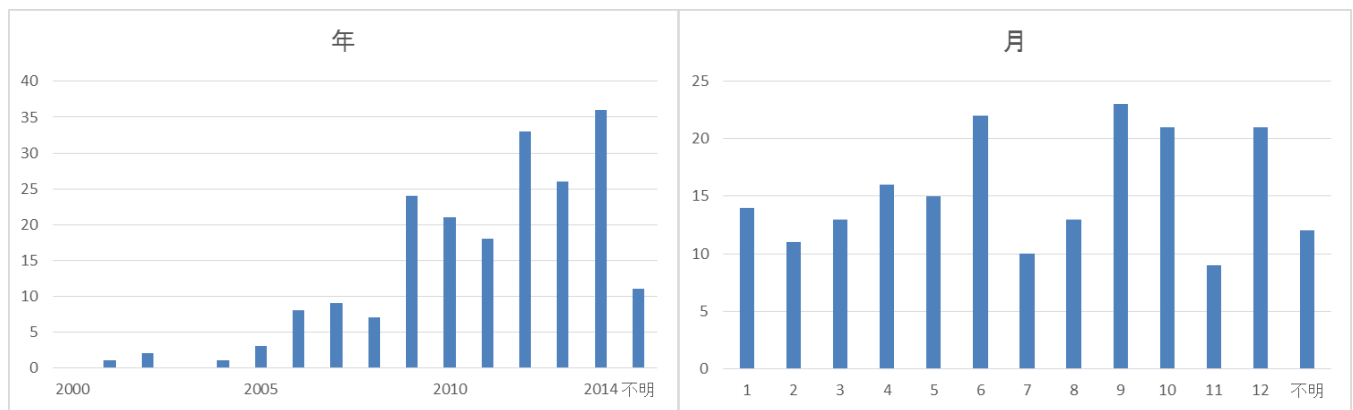


図-1 年別来訪者数

図-2 月別来訪者数

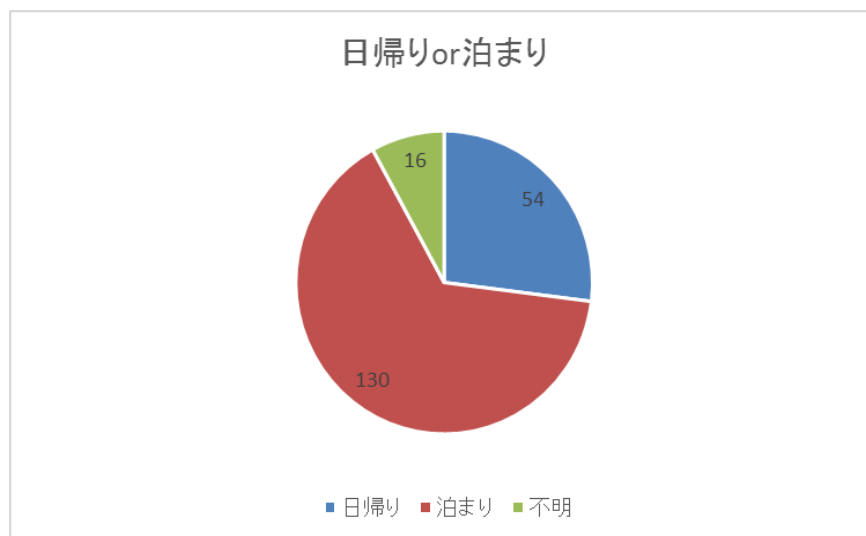


図-3 日帰り or 泊まり

3. ブログマイニングと基礎的特性

テキストマイニングは KH-Coder³⁾ というソフトウェアを利用する。これは多量のテキストデータに対して、形態素解析を行い、文章を意味のある最小の語句(形態素)にそれぞれ分解する。その上で、分析対象に応じて語句の出現頻度や共起性などの指標をグラフィカルに出力可能なソフトウェアである。これまでも土木計画だけでなく、多くの分野でテキストマイニングのツールとして利用されているソフトウェアである。

Kh-coder を利用して、抽出したブログデータにおける頻出単語を検索した結果が表-1 に示される。それぞれ、名詞、形容詞、固有名詞（地名）、動詞の中で出現数を示している。

表-1 頻出単語

名詞	出現数	形容詞	出現数	地名	出現数	動詞	出現数
温泉	653	美味しい	181	湯布院	1056	行く	501
旅館	199	良い	168	由布院	183	思う	305
部屋	175	多い	82	由布岳	135	食べる	266
写真	121	楽しい	77	金鱗湖	127	見る	229
場所	121	広い	73	玉の湯	86	入る	195
土産	106	高い	66	地獄	77	言う	192
感じ	104	素晴らしい	46	亀の井	55	来る	121
人気	101	少ない	44	由布院駅	46	向かう	115
夕食	94	無い	43	無量塔	45	走る	109
雰囲気	78	美しい	41	阿蘇	38	買う	109

これらから、地名としてはまず検索ワードにもある「湯布院」があり次いで別名で駅名でもある「由布院」、名所である「由布岳」や「金鱗湖」があり、形容詞としては「美味しい」「良い」が多くあげられている。よって各地名に対しての良い印象や感想が得られていると考えられる。しかしより頻出のものだけを見れば湯布院に訪れた人のブログは旅行等についての楽しい思い出のみを書き連ねたものばかりということになるだろうが、形容詞においてある程度の出現数があるもので「少ない」や「無い」といったマイナスな感情と思われるものも出現しているため、湯布院を訪れた人の中に悪い印象や感想を持った人もいるということも考えられる。よってその場所への印象や感想の良し悪しが分かればその情報は修景等への手助けとなるだろう。

4. コーディングによるブログ特性の判定

このようなデータに対して、コーディングで特性の把握が可能かを試みた。まずは基礎的特性として性別の判定が可能かを検証した。手順として、すべてのブログを目視にて確認し、掲載されている写真・プロフィール等を含めて性別の判定を行う。その上で、目視にて確認した際の判定基準をコード化し、単語の出現の組み合わせに変換する。その変換したルールを適用し、当初目視にて判定した性別をどの程度再現できるかを検証した。目視にて判定した性別を図-4に示す。

これを再現する最適化されたコーディングルールとしては「男」「俺」「僕」「嫁」「かみさん」という各ワードのいずれかが登場し、かつ「女」「夫」「彼」「メイク」「化粧」という各ワードが同文中に無ければ男性、その逆を女性とするというルールを用いる。これにより当初の判定の65.5%は再現され、誤判定として、性別を誤ったもの11.9%、判定できなかったものが22.6%存在する。またこれとは別に、プロフィールによって判定できたものが15.0%あり、プロフィールをテキストマイニングの対象とすることで、精度を97.1%まで向上することが可能であることが示された。誤判定の原因としてはブログ記述者以外の登場人物についての記述が主なものであり、これらは誤判定したワードの前後に登場している別のワード（例、「男の店員が」であれば「店員」）を、それがあればカウントしないというようにコーディングルールを追加することによって改善が可能であると考えられる。

同様の方式によって、観光行動の重要な特性である初来訪・リピーターの別を判定した。その結果が図-5である。

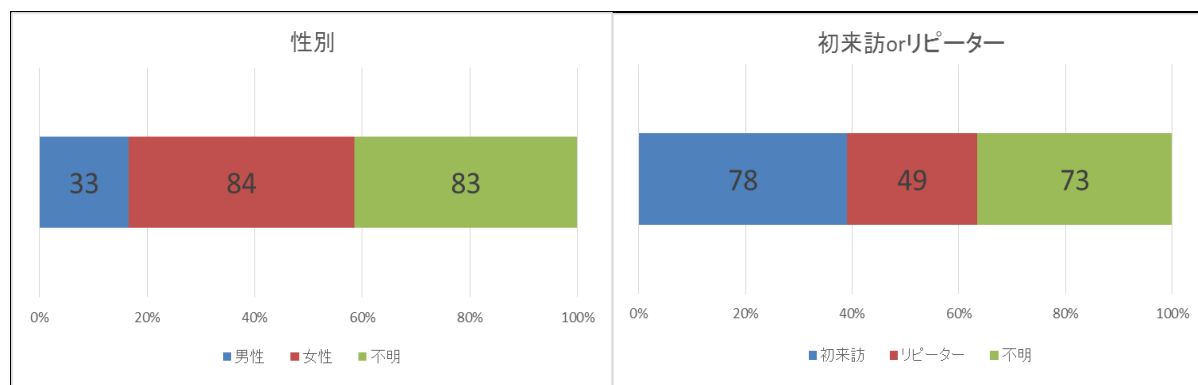


図-4 性別

図-5 初来訪 or リピーター

これを最大に再現できるコーディングルールは「初」「はじめて」「いままで」「思っていたより」「がっかり」「知らな」「いままで」のいずれかが登場していれば初来訪であり、「以前」「去年」「前回」「久」「数回」「何回」「熟知」のいずれかが登場していればリピーターとするというルールを用いる。これによって初来訪・リピーターの別を74.2%再現することが可能であった。ただし誤判定として、誤った判定となったものが16.7%、判定できなかったものが9.1%存在した。これはそれぞれ、景色等を初めて見たことについての感想がコードにはかからず初来訪と判定されなかったり、「以前」行った場所を湯布院と比較する記述があった場合にリピーターと判定してしまっていたりという理由から発生している。この誤判定を正すためにルールを追加する。前者は初来訪と判定する際のコード以外の、一見、初来訪かどうかの判定とは関係ないような単語を、ブログ

を直接読んだ際に初来訪であるした根拠となる文章から選びコードとして追加する。後者は以前行った場所の名前をコードとして、それが同文中に登場しなければリピーターと判定するといったものが考えられる。

先にあげたように、観光地の評価として景観は重要なものである。そこで、景観対象としての地物（地名）とその形容詞との関連性を、共起ネットワークにて分析した。その結果を図-6、図-7に示す。図-6は出現頻度がより高い地名と形容詞についてのネットワークであり、図-7は地名の出現頻度がそこまで高くない地名も含めたものと形容詞のネットワークである。これらから、「由布岳」については、「良い」「美しい」「少ない」のような形容詞が同時に使われる割合が高く、また「金鱗湖」については「多い」「よい」のような形容詞と同時に使われている割合が高いことが示される。

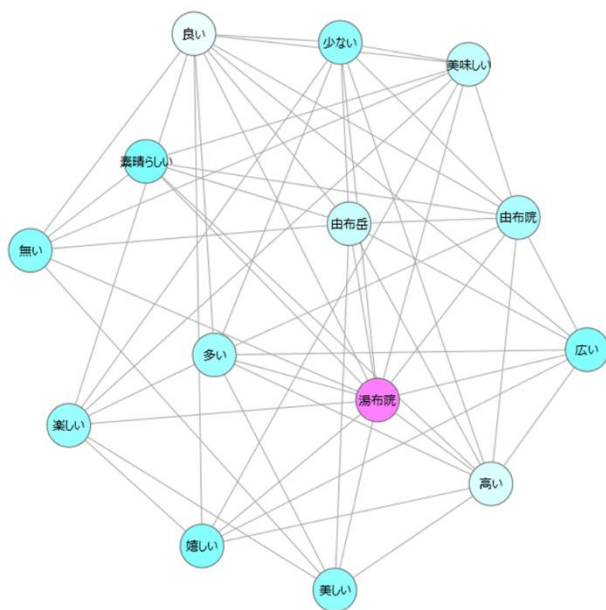


図-6 高頻出の地名と形容詞

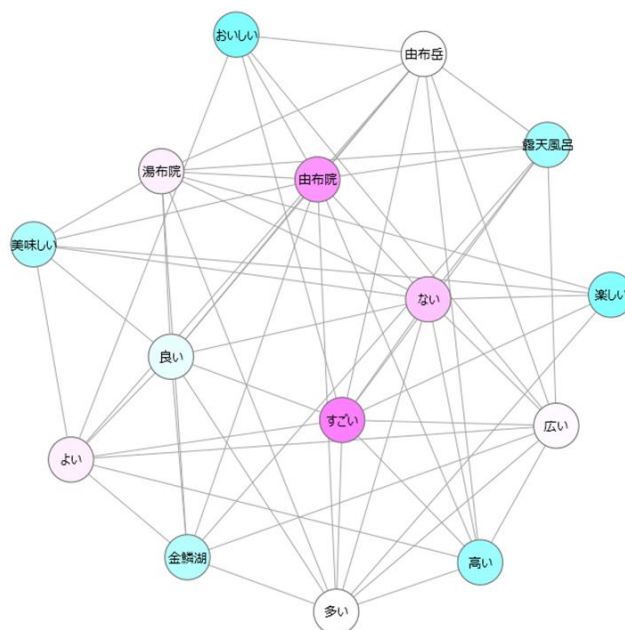


図-7 高、中頻出の地名と形容詞

5. まとめ

本研究では、コーディング等によってどの程度ブログ内容を適切に判定できるかを、あらかじめ目視によって判定とルールを作成し、それを用いて判定可能性について検討を行った。その結果、性別と初来訪・リピーターについては適合率の高い判定が可能なルールを抽出することができた。その一方、地物の評価については、共起ネットワークによって関連性を見ることで、どのような評価が行われているのかを示した。今後の課題として、地物の評価についても適切な判定ができているのかを目視との比較によって検証し、こういったルールやコーディングによってそれらを最適に判断できるかを示していきたい。

6. 参考文献

- 1) ゆふいん建設・環境デザイン協議会 <http://www.gappei-archive.soumu.go.jp/db/44ooita/10yuhu/index.html>
- 2) 由布院温泉観光協会
<http://www.yufuin.gr.jp/>
- 3) KH-Coder
<http://khc.sourceforge.net/>
- 4) 桜田朋也, 佐々木邦明: テキストマイニングによる観光行動分析に関する研究, 2011
- 5) 紀藤舞華, 佐々木邦明: 観光行動調査とブログマイニングの相互補完性に関する研究, 2014