

研究指導 中澤 真 准教授

# ホテルサービス改善のためのテキストマイニングを用いた 宿泊予約サイトのクチコミ分析～花巻温泉郷を事例として～

澤藤 なつ美

## 1. はじめに

近年、インターネットの普及とさらなるIT化の進歩が、観光ビジネスにも影響を与えている[1]。その影響を与えたものの一つが宿泊予約サイトである。このサイトではクチコミを投稿することができ、実際に利用したことのあるユーザの声を見ることができるのが特徴である。それをもとに旅行の計画を立て、予約をし、施設に足を運ぶ人も多い。このような消費者行動の変化が広がる一方で、企業側も実際に利用した利用客の生の声取り入れ、有用なマーケティングデータとして宿泊予約サイトを活用している。

クチコミのホテルに対する評価情報には、5段階で優劣を表した数値情報と、自由記述欄の自然言語で記述されたテキスト情報がある。数値情報はデータとして取り扱いやすく、分析もしやすい。しかし、なぜそのような評価になったのかという要因を明らかにすることは難しい。一方で、テキスト情報は分析をする際に手間がかかってしまうが、データとしては多くの情報を得ることができる。しかし、企業側がテキスト情報をマーケティングデータとして活用しようとするには、何万字もの分析をする必要が生じ、これを手作業で行うのは現実的ではない。実際、ホテル側はクチコミに対して個別に返信をするなどの対応はしているが、どのようなクレームが多いかというような傾向については担当者の感覚に頼ることになってしまう。

このような問題に対し、自然言語処理やテキストマイニングの技術を活用し、その分析結果をマーケティングに活かす試みも行われている[2][3][4]。しかし、サービス改善につながる研究では高評価のホテルのみで比較されており、サービス改善へとつなげるためには、低評価のクチコミを取り入れるべきであると考えられる。

そこで本研究では、ネガティブな単語に着目したテキストマイニングをすることで、ホテルサービスの改善点をクチコミから発見する手法を提案する。

## 2. 宿泊予約サイトの現状と課題

宿泊予約サイトとはインターネット上からホテルの予約を可能とするウェブサイトのことを指す。JTB総合研究所の調査によると[1]、最初に旅行プラン等を探す方法として52.8%の人が「オンライン専門の宿泊予約サイトで探す」ということが示されている。宿泊予約サイトは、複数のホテルの比較がしやすいことや、ユーザが求める各種条件に合致した宿泊施設を瞬時に探すことができるため利用率が高いと考える。

さらに、宿泊予約サイトは予約機能に加えてクチコミを投稿・閲覧できるというクチコミサイトとしての機能も持っている。実際、ユーザが宿を選択する情報として、クチコミを最重視している人が8割近くもいるという結果も報告されている[5]。

そのため、ホテル側はクチコミに対して敏感に反応せざるを得なくなっており、クチコミに対する丁寧かつ迅速な個別の返信をすることが求められる。しかし、このような対応をすることで個々の問題点は把握することが可能となるが、全体的な問題点の傾向は感覚的にしかわからないため、マーケティングとしては不十分である。

このような問題に対し、自然言語処理やテキストマイニングの技術を活用し、その分析結果をマーケティングに活かす試みも行われている。次節では、これらの既存の方法について述べる。

## 3. テキストマイニングによるクチコミ分析と課題

テキストマイニングとは、テキストを対象としたデータマイニング技術であり、「確率・統計的な手法を用いて大量のテキストから有用な情報を引き出す技術の総称」のことである[6]。

田邊ら[2]は宿泊予約サイトからクチコミのテキストデータを収集し、テキストマイニングにより、主成分をいくつかの層に分けている。その結果からホテルに期待するユーザの期待要因を見つけ、個々のホテルがとるべき方向性を示している。さらに二種類のテキストマイニングをしており、その結果を比較し考察することで、研究の精度を明らかにしている。しかし、この研究では、宿泊客がネガティブな意見を言っている単語は抽出できているが、その単語に対しなぜネガティブな意見となったのかは特定できていなかった。

石橋[3]は伊豆地域にある旅館・ホテルを対象にクチコミを用いた数値データ分析とテキストマイニングに取り組んでいる。数値データ分析ではエリアごとや、総合点数に基づいた男女比、年代の比率などが示されている。一方、テキストマイニングではエリアと性別による比較分析をしている。しかし、いずれの分析結果でもホテル別の特徴は示されていない。

久保田[4]は宿泊予約サイトからクチコミを抽出しテキストマイニングによって単語頻度等を示すことによりホテルの特徴を抽出している。さらに、データを再構築することでホテル間の関係やグループごとに見たホテルの特徴が示されている。しかしこの研究では、ビジネスホテルを3種類ピックアップしその関係性や特徴を明らかにしているが、ポジティブな意見の傾向を示しているのみで、マーケティングに結びつける具体的な考察が十分になされていない。

3つの先行研究では、いずれもホテルごとの特徴は明らかにされておらず、ホテルグループや性別等でグループ分けされた分析であった。さらに、ネガティブな単語に注目した分析はされておらず、サービス改善へとつながる各ホテルの問題点を発見できるような結果も示していなかった。

そこで本研究では、ホテルのサービス改善にクチコミを用いる場合、ネガティブな内容の傾向についてホテル別に他ホテルとの違いを比較すれば、宿泊客の視点からのホテルの問題点を洗い出すことができると考えた。その方法としてまずクチコミを形態素解析<sup>1</sup>して頻度の高いネガティブな単語の傾向を抽出、さらに抽出したネガティブな単語の近くに出現している単語の傾向を分析する。この結果から、シンプルな5段階評価では見つけることができない宿泊客からの印象を明らかにし、ホテルサービス改善へとつなげる分析・考察をする。

#### 4. 抽出したクチコミの分析と考察

本節では、宿泊予約サイトから抽出したクチコミを分析し、その結果を考察する。

##### 4.1 調査方法

本研究では国内の大手宿泊予約サイトの「じゃらん.net<sup>2</sup>」（以下じゃらん）のクチコミデータを利用した。このデータの中から、今回は岩手県花巻市にある花巻温泉郷の総クチコミ数が900以上ある7ヶ所<sup>3</sup>のホテルを対象に分析を進めた。なお、2016年11月から2017年10月までの一年分の971件のデータを分析対象とする。

テキストマイニングをする際に使用したソフトウェアは、形態素解析ソフト「MeCab<sup>4</sup>」、統計分析をする「R<sup>5</sup>」、「MeCab」と「R」を繋ぐ「RMeCab」である[7]。

##### 4.2 ホテルごとの出現頻度

まず、抽出したクチコミを形態素解析して単語単位に分割する。この単語の出現頻度をホテルごとに算出することにより、そのホテルのクチコミを投稿した宿泊客が、どのようなことを重視してホテルに宿泊しているのかを定量的に分析・比較する。

例として花巻温泉のクチコミに現れた名詞単語の中で、出現頻度の上位15位までの単語を表1に示す。品詞を名詞に限定したのは、宿泊客が注目している対象物を発見することが目的であるためである。花巻温泉のクチコミにおける最頻度の名詞は「部屋」であり、宿泊客は「部屋」に一番多く注目していることがわかる。部屋は宿泊客がホテルに宿泊した際に最も長い時間を過ごす場所であるため、必然的に細かいところまで注目してしまうからと考えられる。2位の「の」という名詞は「非自立」の名詞であり、単独では意味をなさないため、このような名詞は分析対象から除外する。その他に「風呂」や「バイキング」など上位に現れている単語はいずれも宿泊客が重要視しているポイントと考えられるが、5位の「温泉」については、ホテル名に「温泉」が入っているため、必ずしも入浴する「温泉」を重視していると直ちに結論づ

けることはできない。

このように、出現頻度が高い名詞に注目することで、宿泊客がホテルを利用した際に、どのような点に注目しているのかが定量的にわかるようになった。これは他のホテルでも同様の分析をすることが可能である。

表1:花巻温泉(ホテル紅葉館)の単語頻度表

順位	単語	品詞	品詞細分類	頻度
1	部屋	名詞	一般	181
2	の	名詞	非自立	162
3	風呂	名詞	一般	150
4	利用	名詞	サ変接続	141
5	温泉	名詞	一般	137
6	バイキング	名詞	一般	129
7	ホテル	名詞	一般	108
8	満足	名詞	サ変接続	94
9	夕食	名詞	一般	88
10	こと	名詞	非自立	76
11	宿泊	名詞	サ変接続	66
12	食事	名詞	サ変接続	66
13	朝食	名詞	サ変接続	62
14	子供	名詞	一般	61
15	残念	名詞	形容動詞語幹	56

##### 4.3 各ホテルの頻度表比較

4.2ではホテル単体の出現頻度を示したが、これだけではサービス改善につなげることができない。そこでAホテルとその他のホテルというように、1対6でホテルのクチコミの単語頻度を比較する。この比較から、それぞれのホテル固有の特徴的な単語が明らかになると考えた。

テキストマイニングでは名詞、形容詞、動詞を対象に分析するのが一般的である。名詞はさらに、一般、接尾、非自立というようにさらに細分類できるが、単独で意味をなさないものも含まれるため、5種類<sup>6</sup>の名詞のみを対象とした。同様に形容詞においても、単独で意味をなさない細分類区分を除き「自立」のみを対象とした。

表2は、鉛温泉とその他のホテル6箇所のクチコミを集約したものを比較し、表にしたものである。鉛温泉とその他のホテル両方に示した出現率は、その単語のクチコミ1件当たりの出現率であり、単語の出現頻度を各ホテルのクチコミ総数で除したものである。表を見ると、鉛温泉の名詞の5位である「湯治」という単語は、その他6箇所のホテルの出現頻度の上位には現れていないことがわかる。実際に鉛温泉には湯治部という、湯治をするために利用する温泉がある。このことから湯治を目的とした宿泊客が少なくはないことが予想される。加えて、「湯」という単語から、「お湯」が特徴的であることも推察される。

形容詞では「深い」という単語が鉛温泉では6位、その他のホテルでは53位と出現率の順位に大きな差が現れている。これは、鉛温泉には日本一深い自噴天然岩

<sup>1</sup> 言語を構成する最小単位である単語を切り出す技術[6]

<sup>2</sup> <https://www.jalan.net/>

<sup>3</sup> 結びの宿愛隣館, 山の神温泉 優香苑, 湯の杜 ホテル志戸平, 大沢温泉「山水閣」, 鉛温泉・藤三旅館, 花巻温泉ホテル紅葉館, 渡り温泉 別邸楓, を分析対象としている。

<sup>4</sup> <https://www.mlab.im.dendai.ac.jp/~yamada/ir/MorphologicalAnalyzer/MeCab.html>

<sup>5</sup> <https://cran.r-project.org/>

<sup>6</sup> 一般, サ変接続, 形容動詞語幹, 代名詞, 副詞可能

温泉があり、ホテルの名物として知られている。この温泉を目的に利用する宿泊客が多いため、このような結果になったといえる。

形容詞に「古い」という単語が入っているが、この場合、ポジティブな意味であるのか、ネガティブな意味なのかは判断がつかない。ホテルや旅館に関しては「古い」という単語はポジティブな意味でもネガティブな意味でも使われる。そこで、単語の共起関係を用いて、これらの単語をポジティブかネガティブか判別できるようにする。

表2:鉛温泉とその他ホテルの頻度比較

名詞	形容詞			
	鉛温泉		その他ホテル	
	順位	出現率	順位	出現率
温泉	1	97%	4	66%
湯	2	97%	28	14%
部屋	3	84%	1	89%
風呂	4	83%	2	77%
湯治	5	72%	該当なし	0%
利用	6	41%	3	66%
食事	7	40%	9	37%
時間	8	35%	24	18%
満足	9	29%	5	46%
最高	10	29%	17	24%

#### 4.4 ネガティブな単語の共起

単語の出現頻度を比較した際、どのホテルにも少なからず表3に示すような明らかにネガティブな単語は出現していた。しかし、出現頻度の高いネガティブな単語を抽出しただけでは、その要因を突き止めることはできず改善策へとつなげられない。そこで、改善する点を見つけるため、ネガティブな単語に注目した単語の共起頻度を分析する。共起とは、分析対象とする単語の近くに高い頻度で現れる単語のことである[8]。

表3:対象ホテルすべてのネガティブ単語上位10個の出現頻度と出現率

単語	品詞	品詞細分類	出現頻度	一クチコミ当たりの出現率
残念	名詞	形容動詞語幹	207	21.3%
不満	名詞	一般	16	1.6%
無い	形容詞	自立	55	5.7%
悪い	形容詞	自立	50	5.1%
少ない	形容詞	自立	46	4.7%
寒い	形容詞	自立	36	3.7%
遅い	形容詞	自立	25	2.6%
うるさい	形容詞	自立	12	1.2%
暗い	形容詞	自立	10	1.0%
臭い	形容詞	自立	10	1.0%

共起関係を分析するための前処理として、Rを用いてN-gramによる単語の連りを取り出した。N-gramとは連続する単語の連りのことをさし[8]、Nを5とした場合、文中の前後5つの連続する単語の集合関係が抽出される。今回はネガティブな単語から前後5つ分までの範囲の

単語を共起関係のある単語と考え、5-gramの中でネガティブな単語と同時に出現している単語を抽出した。

まず、ネガティブな単語の抽出として、7箇所すべてのホテルのクチコミに対し、出現頻度の高い名詞と形容詞の中から、ネガティブに捉えられると思われる単語を筆者が手動で取り出した。その結果を表3に示す。

次に、表3に示したネガティブ単語と共起回数の高い単語を抽出した。表4は山の神温泉の共起表である。列方向にネガティブ単語を配置し、行方向に出現頻度の高い名詞を配置した。表内の数値はそれぞれの名詞がネガティブ単語と共起した回数を表している。合計は10個のネガティブ単語との共起回数の和であり、これをその名詞の出現頻度で割ったものが出現率である。この出現率が高い名詞ほど、ネガティブな文内で使われていることを意味するため、そのホテルの問題点を指し示している可能性が高い。このような共起関係の出現率をホテル別に算出し、比較分析することでホテルの問題点の洗い出しを試みる。

表4:山の神温泉のネガティブな単語との共起回数と共起出現率

名詞	残念	不満	無い	悪い	少ない	寒い	遅い	うるさい	暗い	臭い	合計	共起出現率
部屋	7	0	0	0	0	0	0	4	0	0	11	15%
風呂	4	0	0	0	0	0	0	0	0	0	4	13%
温泉	0	0	0	0	0	0	0	0	0	0	0	0%
利用	2	0	0	0	0	0	0	0	0	0	2	7%
満足	0	0	4	0	0	0	0	0	0	0	4	17%
夕食	3	0	0	0	0	0	0	0	0	0	3	11%

#### 4.5 ネガティブ単語の分析・考察

まずホテル別に、出現率上位の名詞に対し、ネガティブ単語との共起関係の出現率の結果を表5に示す。

表5:ホテル別のネガティブ単語と共起関係の強い名詞の共起出現率ランキング

名詞	単語の出現頻度	愛隣館	鉛	花巻	山の神	志戸平	大沢	渡り
部屋	858	19%	38%	34%	15%	<b>61%</b>	11%	15%
風呂	751	16%	9%	23%	13%	<b>37%</b>	10%	16%
温泉	675	9%	10%	15%	0%	19%	2%	24%
利用	614	6%	15%	7%	7%	15%	10%	0%
夕食	405	35%	4%	17%	11%	29%	18%	<b>63%</b>
宿	401	9%	10%	6%	8%	4%	0%	0%
食事	360	22%	17%	18%	17%	44%	16%	33%
バイキング	320	17%	0%	31%	20%	25%	0%	0%
宿泊	302	7%	11%	6%	14%	8%	19%	50%
料理	268	11%	14%	42%	16%	52%	17%	50%
旅行	259	1%	29%	0%	0%	0%	6%	21%
今回	252	15%	0%	9%	13%	25%	0%	0%
湯	235	11%	12%	0%	14%	50%	0%	0%
朝食	232	22%	0%	35%	25%	24%	39%	11%
家族	217	0%	50%	8%	7%	3%	9%	0%
子供	217	11%	0%	3%	0%	3%	10%	0%
スタッフ	202	12%	18%	36%	8%	0%	8%	22%
ホテル	196	22%	0%	16%	0%	62%	0%	0%
対応	195	5%	25%	39%	44%	24%	38%	33%

この表中の「志戸平温泉」に着目すると、「部屋」、「風呂」という単語がネガティブ単語と共起する割合が他のホテルと比較しても高く、サービスの改善点として検討すべき項目であることを導き出すことができる。

7 全ホテルのクチコミでの出現頻度の上位の名詞であり、かつ 4.2 に述べた細分類に属することを条件としている。

また、5位の「夕食」に着目すると、「渡り温泉」が60%を超え、「夕食」に対する評価が低いことがわかった。これを実際のクチコミと比較してみると、「部屋」という単語を含むクチコミは約6割がネガティブな評価をしていたため、共起関係の出現率によって問題点の抽出ができることを確認できた。

さらに、ホテル別に名詞とネガティブ単語との共起出現率を降順で並べた結果を表6に示す。ただし、共起出現率が高い名詞の中には出現頻度が低いものも含まれており、1人のクチコミに左右されてしまっているものもある。このため、出現頻度が10未満の名詞は共起出現率が高くても除外している。

表6:ホテル別のネガティブ単語と共起関係の強い名詞の共起出現率ランキング

愛隣館	鉛	花巻	山の神
夕食 35%	部屋 38%	種類 86%	対応 44%
時間 34%	プラン 31%	料理 42%	接客 38%
食事 22%	対応 25%	対応 39%	プラン 20%
ホテル 22%	スタッフ 18%	スタッフ 36%	食事 17%
朝食 22%	食事 17%	朝食 35%	料理 16%
露天風呂 20%	利用 15%	予約 35%	部屋 15%
部屋 19%	料理 14%	部屋 34%	湯 14%
種類 18%	湯 12%	露天風呂 33%	宿泊 14%
バイキング 17%	宿泊 11%	バイキング 31%	風呂 13%
大変 17%	宿 10%	時間 29%	露天風呂 13%
志戸平	大沢	渡り	
ホテル 62%	朝食 39%	夕食 63%	
部屋 61%	対応 38%	料理 50%	
露天風呂 53%	予約 31%	対応 33%	
料理 52%	宿泊 19%	食事 33%	
食事 44%	夕食 18%	大変 27%	
風呂 37%	料理 17%	温泉 24%	
予約 33%	食事 16%	旅行 21%	
夕食 29%	時間 12%	風呂 16%	
残念 28%	部屋 11%	部屋 15%	
時間 27%	風呂 10%	家族 0%	

この結果をホテル別に考察してみる。例えば、渡り温泉では「夕食」、「料理」、「食事」などの単語がネガティブ単語と多く共起し、その値が他のホテルに比べても高いことがわかる。これは渡り温泉の宿泊客の中に、料理や食事、特に夕食に悪い印象を持ってしまった人が多くいたことを示唆している。さらに「対応」といった単語も入っているため、料理だけでなく、その際の対応にも悪いイメージを持ってしまったのではないかと考えられる。一方、花巻温泉では「種類」が一位であり、ついで「料理」、その下には「朝食」、「バイキング」とあること、また、ホテルでは朝食でのバイキングが多いことから、朝食バイキングの種類についてネガティブなイメージを持ったと考えられる。

以上の結果から、共起出現率を算出することで、各ホテルの宿泊客が何に対してネガティブと考えているのかを明らかにすることができた。

### 5. むすび

本研究では、ホテルサービス改善のため、宿泊予約サイトのクチコミからテキストマイニングを用い、形態素解析を行った。その結果、特定のホテルとその他のホテルの単語の出現頻度の傾向には差異があり、ホテルごとの特徴を抽出できることを示した。さらに、ネガティブな単語を抽出し共起関係を分析することで、各単語に

対するネガティブな単語との共起出現率を算出した。これは宿泊客がネガティブなイメージを何に対して持っているのかをホテルごとに明らかにすることができる情報であり、5段階評価からは得ることができないものである。以上のことから、本研究ではテキストマイニングを用いることで目視では分析することが容易ではないテキストデータからサービス改善へとつながる結果を効率的にかつ定量的に示すことが可能となった。

今回の分析ではホテルの件数は7件と少なく、収集したクチコミ数も差が生じてしまった。そのため、単語や共起の出現率に特徴が見られなかった部分や、他のホテルと比較した際に違いが見られない場合があった。また、クチコミ本文と一緒に性別や年代といったパーソナル情報も収集したが、今回の分析ではその情報を使用するにはいたらなかった。以上のことから、ホテルの件数やクチコミ数を増やすとともに、性別や年代といった項目も加味した分析をすることが今後の課題である。

### 参考文献

- [1] 株式会社JTB総合研究所, スマートフォンの利用と旅行消費に関する調査, 2017, <https://www.tourism.jp/wp/wp-content/uploads/2017/12/smartphone-2017.pdf>
- [2] 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", 武蔵工業大学環境情報学部情報メディアセンタージャーナル (9), pp.91-101, 2008
- [3] 石橋 太郎, "e-ロコミのテキスト・マイニング分析に向けて(その1):伊豆地域におけるホテル・旅館を対象として", 静岡大学経済研究・17(2), pp.1-11, 2012
- [4] 久保田貴文, "旅行情報サイトにおけるホテルロコミデータのテキストマイニング", 多摩大学経営情報学部, 2016.
- [5] TripAdvisor, ロコミ利用の動向に関するユーザー12,000名以上への調査を発表, 2014, [http://tg.tripadvisor.jp/news/wp-content/uploads/2015/03/20140213\\_TripAdvisorPressRelease.pdf](http://tg.tripadvisor.jp/news/wp-content/uploads/2015/03/20140213_TripAdvisorPressRelease.pdf)
- [6] 小町 守監修, 奥野 陽, グラム・ニュービック, 萩原正人, "自然言語処理の基本と技術", 翔泳社, 2016
- [7] 石田基広, RMeCabの使い方, 2008, <http://rmecab.jp/wiki/index.php?plugin=attach&refer=RMeCab&openfile=manual.pdf>
- [8] 小林雄一郎, Rによるやさしいテキストマイニング, オーム社, 2017
- [9] 石田基広, Rによるテキストマイニング入門, 森北出版, 2017
- [10] 石田基広, Rによるテキストマイニング入門-初級, 2009, <http://rmecab.jp/wiki/index.php?plugin=attach&refer=RMeCabFunctions&openfile=ism091124doc.pdf>