e-口コミのテキスト・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

論　説

e-口コミのテキスト・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―

石　橋　太　郎

Ⅰ．はじめに

インターネットが普及した現在，多くの旅行者（消費者）は，インターネット上の旅行代理店（virtual travel agent）のWebページにアクセスし，過去の利用者の評価ポイントが高いホテル・旅館を検索する．そして，やはり過去に利用したことがある人のレビュー記事を参考にして，旅行先のホテル・旅館を決定しようとする．

元来，ホテル・旅館が提供するサービスの品質は，消費者にとって実際に消費するまでは不確かである．その品質は，消費して（経験して）初めてわかる．すなわち，ホテル・旅館が提供するサービスは，「経験財（experience goods/services）」である[(1)]．経験財の購入にあたって，過去には，その財・サービスを購入消費した経験がある（と思われる）家族，友人，知人から，その（品質）情報を事前に入手しようとした．家族，知人，友人からの情報を「口コミ（word of mouse）」という．よい口コミを得た消費者は，そうでない消費者に比べてその財を購入する確率は高くなる．経験財を提供する売り手にとって良い口コミは，売り手の良い「評判（reputation）」となる．その評判に基づいて消費者がその財を購入することになれば，良い評判の形成は将来の利益を生み出す資産形成に等しい[(2)]．

かつての口コミは，今ではインターネット上での商品・サービスのレビュー記事となり，「e-口コミ（e-word-of-mouse）」と言われている．すなわち，口コミのデジタル化である．デジタル化された口コミに関する関心は，対象となる商品（財）・サービスを購入することを検討している消費者だけのもではない．Dellarocas（2003）は，Webページに書き込まれたe-口コミは，売り手にフィードバックされ，ブランドの構築，評判の形成，顧客の獲得・保持，製品開発と品質保証をするうえで大きな可能性があることを示した．

---

(1) 経験財については，最初に言及したNelson（1970）を参照．

(2) 例えば，Shapiro（1983）を参照．Kean（1997）は，Shapiro（1983）の「評判」の考え方に基づき旅行者の旅行先決定モデルを構築している．Ledesma et al.（2005）は，旅行者の旅行先の決定において，そして旅行者がリピータとなるうえで旅行先の「評判」の重要性を実証した．

e-口コミは，デジタル化されたテキスト情報である．情報を有効に活用するためには，その情報を分析しなければならない．テキスト情報を分析する手法として，テキスト・マイニング（text mining）という手法がある．本稿が分析対象とするホテル・旅館についてのe-口コミの分析事例としては，例えば，Lee and Hu（2004）がある．彼らは，テキスト・マイニングにより，インターネットに構築されたフォーラム上で表明されるホテルの顧客の不満について分析を行った．Lee *et al.*（2011）は，香港のホテルの顧客に対して実施された聞き取り調査のテキスト情報をテキスト・マイニングにより分析している．本稿は，こうした先行研究と同様の問題関心を持つものである．すなわち，ホテル・旅館が提供したサービスについての顧客による評価を含めた（e-）口コミ情報を分析し，テキスト・マイニングにより分析することである．分析結果は，ホテル・旅館のサービスの改善，ブランド・評判の改善等に貢献し，分析の対象地域とした静岡県伊豆地域の観光産業の発展にも寄与しうることが期待される．本稿は，こうした目標を達成するための出発点として，e-口コミの準備的分析を行うものである．

Ⅱ．準備的分析

１．データの収集

本稿が分析対象とするのは，静岡県伊豆地域のホテル・旅館である．これらのホテル・旅館の顧客による評価ならびにe-口コミは，インターネット上の旅行代理店（virtual travel agent）のWebページより収集した．収集先としたのは，リクルート社が運営する「じゃらんnet」（http://www.jalan.net/）で，収集の対象としたのは，ホテル・旅館の滞在時期が（ゴールデンウィークを含む）５月である投稿記事である[(3)]．さらに付け加えるならば，実際に収集したのは2010年であり，東日本大震災（2011年３月11日）の前年のデータである．東日本大震災は甚大なる人的・物的被害をもたらしたとともに，日本経済にも大きな影響を及ぼした．大震災後，さまざまな祭りや催し物が中止され，観光へも影響を与えた．こうした影響が大震災前と後でe-口コミにも現れたのかは，１つの研究課題となろうが，本稿では大震災前のデータだけを使っている．

収集したデータは，次のように整理した．まず，伊豆地域（半島）全体を大エリアとし，熱海，伊東・宇佐美・川奈，伊豆高原，東伊豆，中伊豆，西伊豆，南伊豆，下田・白浜の７地域を中エリアとした．これらの分類は，基本的に「じゃらんnet」に従っている．「じゃらんnet」では，小エリアとしてさらに地域分割されているが，本稿ではその分類は採用していない．そして，各中エリアごとにホテル・旅館を検索し，投稿記事（e-口コミ），投稿者の年代・性別，投稿者による

(3) ただし，実際の投稿は，滞在後に入力されるため，投稿日が６月になっているものもある．

ホテル・旅館の評価点を収集し，データベース化を行った．データ収集の対象となったホテル・旅館数は171軒であり，e-口コミ数は1,828件である．

次に，このデータベースから見えてくる基本的特徴について整理しよう．

２．データの特徴

表１は，１ホテル・旅館当たりのe-口コミ回数の記述統計量をまとめている．表１からわかるように，e-口コミの投稿回数は１ホテル・旅館当たり10.63件となっている．しかし，中央値，最頻値からわかるように，半数以上が10回未満であり，歪みを持った分布となる．この歪みは，投稿回数57件の旅館が存在することによるものである．半数以上が10件未満であることを考えると，57件の投稿は突出している．

なお，e-口コミの投稿者を男女別にみると，女性が884件（48.36％），男性が943件（51.59％）であった（性別不明の投稿が１件あった）．

次に，評価点数別にe-口コミの状況を整理しよう．評価点数は，５段階評価であり，「５」が最も高い評価を表す．表２から表６は，総合評価点数ごとに，中エリア・男女別の投稿件数，年代・男女別の投稿件数をまとめたものである．

総合評価点数を５点と付けた投稿件数は946件あり，全体の51.75％を占めている．総合評価点数を４点と付けた投稿件数（702件）と合わせると1,648件となり，総合評価点数を５点と４点と高い評価をつけた投稿だけで90.15％にも上る．

このような結果については，さらに検討が必要であろう．「日本人は他人のことを悪く言う傾向

表１　１ホテル・旅館当たりのe-口コミ回数に関する記述統計量

平　　均	10.63
標準誤差	0.75
中央値（メジアン）	8.00
最頻値（モード）	5.00
標準偏差	9.83
分　　散	96.73
尖　　度	4.48
歪　　度	1.91
範　　囲	56
最　　小	1
最　　大	57
合　　計	1,828
標 本 数	172

表２－１　総合評価５点の中エリア・男女別投稿件数

	女性	男性	合計
熱　　海	57	49	106
伊東・宇佐美・川奈	56	61	117
伊豆高原	107	108	215
東 伊 豆	52	93	145
中 伊 豆	38	30	68
西 伊 豆	84	101	185
南 伊 豆	11	10	21
下田・白浜	45	44	89
合　　計	450	496	946

表２－２　総合評価５点の年代・男女別投稿件数

	女性	男性	合計
10　代	4		4
20　代	128	82	210
30　代	159	164	323
40　代	116	147	263
50　代	33	69	102
60　代	9	20	29
70　代	1	10	11
80　代		4	4
合　　計	450	496	946

表３－１　総合評価４点の中エリア・男女別投稿件数

	女性	男性	合計
熱　　海	55	44	99
伊東・宇佐美・川奈	49	63	112
伊豆高原	41	38	79
東 伊 豆	46	47	93
中 伊 豆	36	32	68
西 伊 豆	68	81	149
南 伊 豆	3	8	11
下田・白浜	47	44	91
合　　計	345	357	702

表３－２　総合評価４点の年代・男女別投稿件数

	女性	男性	総計
10　代	1		1
20　代	89	57	146
20　代	1		1
30　代	125	110	235
40　代	81	96	177
50　代	43	66	109
60　代	4	24	28
70　代	1	3	4
80　代		1	1
総　　計	345	357	702

表４－１　総合評価３点の中エリア・男女別投稿件数

	女性	男性	合計
熱　　海	8	9	17
伊東・宇佐美・川奈	9	17	26
伊豆高原	2	3	5
東 伊 豆	9	9	18
中 伊 豆	6	5	11
西 伊 豆	9	10	19
南 伊 豆	1	2	3
下田・白浜	8	2	10
合　　計	52	57	109

表４－２　総合評価３点の年代・男女別投稿件数

	女性	男性	総計
10　代	1		1
20　代	20	5	25
30　代	19	20	39
40　代	10	22	32
50　代	2	7	9
60　代		3	3
総　計	52	57	109

表５－１　総合評価２点の中エリア・男女別投稿件数

	女性	男性	合計
熱　　海	10	1	11
伊東・宇佐美・川奈	4	7	11
伊豆高原	2	0	2
東 伊 豆	2	4	6
中 伊 豆	1	1	2
西 伊 豆	3	2	5
南 伊 豆	1	0	1
下田・白浜	1	4	5
合　　計	24	19	43

表５－２　総合評価２点の年代・男女別投稿件数

	女性	男性	総計
20　代	6	3	9
30　代	11	11	22
40　代	2	3	5
50　代	4		4
60　代	1	2	3
総　計	24	19	43

表６－１　総合評価１点の中エリア・男女別投稿件数

	女性	男性	合計
熱　　海	3	3	6
伊東・宇佐美・川奈	4	1	5
伊豆高原	2	0	2
東 伊 豆		1	1
中 伊 豆	1	3	4
西 伊 豆	1	3	4
南 伊 豆			0
下田・白浜	0	1	1
合　　計	11	12	23

表６－２　総合評価１点の年代・男女別投稿件数

	女性	男性	総計
20　　代	2		2
30　　代	6	5	11
40　　代	2	6	8
50　　代	1	1	2
総　　計	11	12	23

はない」ことを反映しているなどと考えるのは，短絡的である．この点について，Lehto *et al.*（2007）は１つの考えを示唆している．Lehto *et al.*（2007）によれば，インターネット上の旅行代理店は顧客がオンライン上で表明する不満や賛辞を分析し，それらをもたらした要因を明らかにすることができる．とりわけ不満をもたらす要因の改善は，オンライン上での顧客のレビュー（e-口コミ）を管理することができることを示した．Lehto *et al.*（2007）に従うならば，「じゃらんnet」は，高い評価を得るようなサービスを提供していると考えることができる．すなわち，「じゃらんnet」は，顧客が満足し，高い評価を与えるようなホテル・旅館を厳選している，と考えることができる．しかしそう結論するには，他のインターネット上の旅行代理店との比較分析が必要であろう．この点については，今後の課題としたい．

今，「じゃらんnet」の利用者が総じて高い評価を与えていることを確認した．次に，そうした高い評価と結びつく口コミ情報の特徴を検討しよう．

３．テキスト・マイニングによるe-口コミの分析準備

本稿でのテキスト・マイニングは，フリーの統計ソフトR[(4)]と工藤拓氏が開発した日本語形態素解析ソフトMeCab（和布蕪）[(5)]，そして石田基広氏が開発したRMeCabパッケージ[(6)]を利用した．

(4) 本稿で用いた統計ソフトRは，バージョン2.13.1．ダウンロード先は，例えば，http://cran.md.tsukuba.ac.jp/bin/windows/base/（閲覧日2012年７月26日）を参照．
(5) MeCabのダウンロード先は，http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html（閲覧日2012年７月26日）を参照．

本稿は,「１．はじめに」の最後で述べたようにe-口コミの準備的分析を目的としている．そこで，今回準備したデータベースにまとめた全てのe-口コミ情報を分析するのではなく，最も特徴的な部分を取り出し，初歩的なテキスト・マイニングを行う．なお，最も多くの投稿がなされていて，かつ総合評価点５点が付いている伊豆高原を，ここでは最も特徴的な部分と見なした（表２－１を参照）.

総合評価点５点が付けられた伊豆高原のホテル・旅館に対して投稿されたe-口コミの最初のテキスト・マイニングは，形態素解析を行い，その頻度を数えることである[7]．ここでは，解析結果をそのまま示すのではなく，e-口コミを理解するうえで必要な情報に絞って解説しよう.

e-口コミを理解するうえで必要な情報とは，ホテル・旅館の評価を表現する形容詞と評価の対象を示す名詞である．なお，形態素解析は，男女別に行った.

まずは，女性のe-口コミにおける形容詞の数は，78個．男性のe-口コミにおける形容詞の数は，77個．形容詞の語彙数としては男女差はないように見える．より頻繁に出現する形容詞が，ホテル・旅館の評価の程度を主に規定すると考えられる．なぜならば，頻度が高いということは，その形容詞を使っている人が多いことを意味し，逆に言えば，多くの人がその形容詞を使って評価しているからである．他方で，頻度が低い形容詞は，多くの人が経験するホテル・旅館のサービスの特徴をとらえているとは考えにくい．なお，女性のe-口コミ件数107件中３回以下しか出現しない形容詞は50個．２回以下しか出現しない形容詞は43個．１回以下しか出現しない形容詞が33個ある．男性については，e-口コミ件数108件中３回以下しか出現しない形容詞は，50個．２回以下しか出現しない形容詞は46個．１回以下しか出現しない形容詞が33個ある．しかし，例えば,「美味しい」と「おいしい」，あるいは「良い」と「よい」は別の形容詞として数え上げられていて，これは明らかに重複である．こうした重複は排除して数え上げるべきだが，ここでは，頻度が高い形容詞について重複を整理し，その結果を示そう.

表７－１ならびに表７－２は，それぞれ女性と男性のe-口コミに現れた形容詞の出現回数が最も高いものから数えた上位10のリストである．女性と男性のe-口コミに共通に現れる形容詞は,「美味しい」,「良い」,「ない」,「いい」,「嬉しい」,「楽しい」,「広い」,「気持ち良い」の８個である．この中で,「美味しい」は食事以外に形容するものはないので,（「じゃらんnet」が提供する）伊豆高原のホテル・旅館は，料理で評価が高いという特徴を持つと考えられる．しかし，その他の形容詞は「何が」,「良い」のか.「何が」,「嬉しい」のか.「何が」,「楽しい」のか，わからない．１つの手掛かりは，形容する対象の「何が」は，名詞であることが考えられる．次に，名詞につい

[6] RMeCabパッケージのダウンロード先は，http://rmecab.jp/wiki/index.php?RMeCabを参照．また，石田（2008）も参照.
[7] 形態素とは言語学の専門用語で,「意味の最小単位」と説明される．石田（2008），p.45を参照.

表７－１　女性の形容詞頻度上位10

	頻度
美味しい	94
良い	74
ない	26
嬉しい	25
いい	23
楽しい	19
広い	19
多い	15
気持ち良い	14
温かい	10

表７－２　男性の形容詞頻度上位10

	頻度
良い	85
美味しい	78
広い	27
ない	26
いい	16
楽しい	13
気持ち良い	20
嬉しい	11
高い	9
素晴らしい	8

ても同様の整理を行い，その結果を示そう[8]．

表８－１ならびに表８－２は，それぞれ女性と男性のe-口コミに現れた名詞の出現回数が最も高いものから数えた上位10のリストである．ただし，同じ頻度のものがあることにより，男性については11個の名詞を上げている．女性と男性のe-口コミに共通に現れる名詞は，「部屋」，「風呂」，「オーナー」，「最高」，「夕食」，「雰囲気」の６個である．前述した形容詞の「美味しい」は，「夕食」が「美味しい」と結びつく．形容詞「良い」は，「部屋」が「良い」．「風呂」が「良い」．「オーナー」が「良い」．「雰囲気」が「良い」など，日本語として自然に結びつく．しかし，実際のe-口コミではどう結びついているのかは，これだけでは分からない．

また，共通していない名詞については，それぞれ女性と男性の評価の対象が異なることに起因すると予想される．例えば，表８を見る限り，女性の上位10位の中には「母」が，男性の上位10位の中には「酒」が入っていて，どのような旅行滞在をしていたかを想像させる．この想像が正しいかは，この語の前後につながる語を検討しなければならない．

４．N-gramと語の共起関係の分析

RMeCabパッケージを利用した統計ソフトRでは，語の結びつきや，あるキーワードにつながる語の関係を分析することができる．

[8] 名詞においても重複があると考えた．例えば，「部屋」と「宿」は同じものと考えた．他に，「風呂」，「露天風呂」，「露天」，「温泉」も同じものと考えた．

表８－１　女性の名詞頻度上位10

	頻度
部屋	165
風呂	96
オーナー	31
最高	28
夕食	38
量	22
口コミ	19
感じ	17
母	17
雰囲気	15

表８－２ 男性の名詞頻度上位10

	頻度
部屋	160
風呂	129
夕食	41
最高	30
オーナー	29
夫婦	21
ボリューム	18
雰囲気	16
気	15
酒	15
貸切	15

例えば,「美味しい夕食」というテキストは,「美味しい」－「夕食」という組み合わせ（結びつき）を取り出すことができる．文字（あるいは，形態素，品詞情報）がN個つながった組み合わせをN-gramという．次に，出現回数が大きい順に10個のN-gramを１例として示しておく．

表９－１　女性の10個のN-gram

	頻度
［オーナー－さん］	11
［オーナー－夫妻］	11
［私－達］	10
［二－人］	10
［口コミ－通り］	9
［２－人］	8
［(－笑］	7
［スタッフ－方］	7
［記念－日］	7
［広い－さ］	7

表９－２　男性の10個のN-gram

	頻度
［二－人］	20
［夕食－朝食］	11
［食事－美味しい］	9
［オーナー－さん］	8
［伊勢－海老］	8
［部屋－風呂］	8
［とも－満足］	7
［オーナー－夫妻］	7
［チェック－アウト］	7
［金目－鯛］	7

表９－１，表９－２で示された語の組み合わせ（結びつき）は，先に想像したような結果を示していない．単純にRMeCabパッケージのNgram（・）関数を適用しても所望の結果は得られない．RMeCabパッケージの開発者である石田基広氏は，この問題を次のように指摘している．「言語は本質的に曖昧であり，コンピュータによる解析結果も完全ではないし，決して完全にはなりえない．（中略）ユーザーは日本語形態素解析機の出力を，分析方針に合わせて修正することも必要になる」（石田（2008），８－９頁）．また「解析結果が分析者の研究意図にそぐわない場合は，独自の辞書」（石田（2008），57頁）を整備する必要がある，という．こうした問題への対応は，今後の課題としたい．

言語学では，ある語（キーワード）が別の特定の語と隣接して出現することを共起（collocation）という．RMeCabパッケージのcollocate（・）関数は，個の共起関係を解析することができる．本稿で利用したデータセットに対して試験的に解析を行った．キーワードとして「美味しい」を設定すると，確かにその前に隣接する語として「料理」が出現することが最も頻度が大きい．しかし，解析は完全であるかというと，前述したとおり，「美味しい」と「おいしい」は同じ誤であるわけだから，「おいしい」と共起する語を解析できていない．ここでも，さらなる工夫・修正が求められる．この問題についても今後の課題としたい．

Ⅲ．今後の課題

本稿を作成する中で，課題は明確となった．１つは，技術的問題である．技術的問題とは，所望の結果が必ずしも直ちに入手することができないことにある．これは先に指摘した通り，コンピュータによる言語解析は，完全なものではない．より完全なもの，より精度が高い解析を目指すためには，１次的な出力結果を精査し，どのような修正が必要であるかを検討することである．２つ目は，研究対象，研究意図の拡充である．本稿では，2010年５月に「じゃらんnet」を利用した消費者の投稿e-口コミを分析対象とした．研究対象の拡充としては，e-口コミの時期の拡張である．季節によっては，e-口コミの内容も異なるものもあるかもしれない．あるいはインターネット上の旅行代理店の違いは，そこを利用するユーザーの特性の違いや，代理店が提供するホテル・旅館の品質の違いももたらすかもしれない．こうした違いがe-口コミの違いを生み出す可能性を否定できない．「じゃらんnet」とは異なる旅行代理店のe-口コミとの比較が必要となろう．

次稿では，ここにあげた課題を改善した結果を報告する．

追記

本稿で使ったデータベースは、科学研究費補助金（基盤研究（C））課題番号22530223によって

作成したものである。

参考文献

１．邦文文献

石田基広（2008)『Rによるテキストマイニング入門』森北出版

２．英文文献

Dellarocas, Chrysanthos (2003), “The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms”, Management Science, 49 (10), pp. 1407-1424.

Keane, Michael J. (1996), “Sustaining quality in tourism destinations: an economic model with an application”, *Applied Economics*, 28 (12), pp. 1545-1553.

Ledesma, Francisco J. , Manuel Navarro and Jorge V. Pérez-Rodríguez (2005), “Return to tourist destination. Is it reputation, after all?”, *Applied Economics*, 37 (18), pp. 2055-2065.

Lee, Charles C. and Clark Hu (2004), “Analyzing Hotel Customers’ E-Complaints from an Internet Complaint Forum”, *Journal of Travel and Tourism Marketing*, 17 (2-3), pp. 167-181.

Lee, Myong Jae, Neha Singh and Eric S.W. Chan (2011), “Service failures and recovery actions in the hotel industry: A text-mining approach”, *Journal of Vacation Marketing*, 17 (3), pp. 197-207.

Nelson, Philip (1970), “Information and Consumer Behavior”, *Journal of Political Economy* 78 (2), pp. 311-329.

R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Shapiro, Carl (1983), “Premiums for High Quality Products as Returns to Reputations”, *The Quarterly Journal of Economics*, 98 (4), pp. 659-679.