レビューデータにおける 評価の時系列的変化に着目したイベント検出

高橋 毅† 天笠 俊之††,‡ 北川 博之††

†筑波大学大学院システム情報工学研究科 〒305-8573 茨城県つくば市天王台 1-1-1
††筑波大学システム情報系 〒305-8573 茨城県つくば市天王台 1-1-1
‡宇宙航空研究開発機構 宇宙科学研究所 〒252-5210 神奈川県相模原市中央区由野台 3-1-1
E-mail: †s1120717@u.tsukuba.ac.jp, ††{amagasa, kitagawa}@cs.tsukuba.ac.jp

あらまし 近年では旅行先の宿泊施設を選ぶ際に宿泊施設レビューサイトを利用するユーザ数が増えつつある．宿泊施設レビューには宿泊者が宿泊施設に対して与える評価値の他に，宿泊施設に該評価を与えた理由が含まれている事がある．理由を発見し，利用者に提示する事は，業務改善など有益な利用につながると考えられる．また理由には宿泊施設の立地といった常に評価に影響を与える要因だけでなく，気温の変化や宿泊プランといった特定の時期にのみ発生するイベントが要因として挙げられるケースが存在する．本研究では，レビュアが宿泊施設に与える評価値が時間変化する点に着目し，変化のあった時間前後のレビューを分析する事によって，該宿泊施設に関連するイベントの検出を行う．

キーワード データマイニング，レビューデータ，時系列分析

Tsuyoshi TAKAHASHI† Toshiyuki AMAGASA††,‡ and Hiroyuki KITAGAWA††

† Graduate School of Systems and Information Engineering, University of Tsukuba
1-1-1 Tennnoudai, Tsukuba, Ibaraki 305-8573, Japan
‡ Division of Information Engineering, Faculty of Engineering Information and Systems, University of Tsukuba
1-1-1 Tennnoudai, Tsukuba, Ibaraki 305-8573, Japan
‡ Institute of Space and Astronautical Science, Japan Aerospace Exploration Agency
3-1-1 Yoshinodai, Chuo-ku, Sagamihara-shi, Kanagawa 252-5210, Japan
E-mail: †s1120717@u.tsukuba.ac.jp, ††{amagasa, kitagawa}@cs.tsukuba.ac.jp

1．はじめに

近年，レビューサイトの重要性が商業的にも学術的にも注目されている．例えば，宿泊施設のレビューサイトを例に取ると，そこには各施設に対する多数の利用者からの評価とレビュー文が蓄積されている．別の利用者が宿泊施設を探す際は，施設自身が提供する情報だけでなく，レビュー結果を参照する事により，利用者の視点に立った評価を参考にする事が出来，より満足度の高い施設選択を行えるようになる．

一方，レビューデータは利用者だけではなく，宿泊施設にとっても有益な情報になる．例えば，評価値から顧客の満足，不満な点を洗い出し，業務改善に活かす事が出来る．このため，膨大なレビューデータを分析し，そこから有用な情報を抽出することが重要となる．そのような分析方法には多様なアプローチが考えられるが，一つの考えとしてレビュー結果の時間的な変化に着目して，評価値が上がった（下がった）あるいは記述されたレビューの内容が変化した時刻に着目し，その原因を探る事が考えられる．

本研究では，レビュー結果の時間変化に着目し，その原因を分析した．このため，まずレビュアが宿泊施設に与える評価値に着目してレビュアの選別を行った．レビュアによっては，有効な評価をせず決まった値のみを入力する場合が考えられ，そのような評価値はノイズとしてあらかじめ取り除くことが望ましい．この目的のために評価値のエントロピーに着目した．次に，有効な評価結果に基づいて評価値の時系列的変化を観察し，変化のあった時刻に関連するレビューからその原因として発生したイベントを調査した．また，レビュー文に記述されている単語出現頻度の時系列的変化を観察し，変化のあった時刻に単語と関連するレビューから発生したイベントを調査した．

本論文の構成は以下の通りである．まず 2 節で関連研究について述べる．次に，3 節でレビューデータの詳細について述べ，4 節で提案手法について述べる．5,6 節では得られた結果について考察を与える．

2. 関連研究

レビューを対象にした研究は盛んに行われている．代表的なトピックとして，レビューの有効度を推定する研究や，レビュー文から効率的に意見を抽出する研究等が挙げられる．

レビューの有効度を推定する研究としては，小倉らの研究が挙げられる[1]．小倉らの研究では，他の利用者がレビュアの投稿したレビューに対して与えた評価を用いてレビューの有効度を算出した．

膨大なレビューから詳細な意見を抽出するための研究が自然言語技術の分野で盛んに行われている[2]．小林らは意見を｛評価対象，属性，評価値｝の三つ組の評価表現形式に構造化してレビュー文から抽出する事で意見を要約した[3]．藤井らは表現が好評と不評のどちらを表すかを示す評価表現辞書を作成し，評価表現を抽出．各評価表現を用いてホテルの評価値を予測した．更に予測したホテルの評価値と評価表現を二次元マッピングする事で，ホテルに与えられた評価の要因を効率的に分析する手法を提案した[4]．しかし藤井らの手法はレビューが時間変化する事を考慮していない．

本研究では，有用なレビューを投稿するレビュアを，登録した評価値のエントロピーで評価し，ある評価値（全て 3，全て 5 など）だけを登録するようなレビューを排除する．また，評価値の時系列的な変化に着目して，その原因を分析する．

3. レビューデータの詳細と投稿時間導出

本研究では楽天技術研究所から提供された楽天トラベルのレビューデータ[5]を用いた．レビューデータには，レビュー対象となる宿泊施設の名称を示す「施設名」，レビュアが自由に記述した「レビュー文」，マスクされた「利用者名」，レビューID を示す「投稿番号」等の他に，評価項目として{立地，部屋，食事，風呂，サービス，設備・アメニティ，総合}が与えられている．レビュアによって 0-5 の 6 段階の評価値が各評価項目に与えられる．

提供されたレビューデータには時間を示す情報は含まれていない．しかしながら，レビューの時間変化に着目する事で，より有益な情報を抽出する事が可能となる．このため，次の方法で投稿時間の導出を行った．まず投稿番号が時間順に振られていると仮定して，ランダムに抽出した 50 件について楽天トラベルサイト[6]上の投稿時間を Web 経由で調査し，投稿番号が時間順で割り当てられている事を確認した．本研究では，月毎にレビューの傾向を分析するため，月末の投稿番号を取得し，表 1 のようなデータを作成した．なお，文書集合からのタイムスタンプ推定については関連研究があり，これらを適用することも可能である[7]．

表 1：レビューデータの例

項目	例
施設名	ホテル A
レビュー文	「久々に利用しました…」
利用者名	user00001
投稿番号	7100000
立地評価値	4
部屋評価値	4
食事評価値	0
風呂評価値	3
サービス評価値	5
設備・アメニティ評価値	2
総合評価値	3
投稿年月	2009 年 12 月

4. 提案手法

本研究では実世界でのイベントを検出するために，評価値が時間変化した点に着目する．また，その原因を特定するため，変化点の前後におけるレビュー内容の変化にも着目する．

4.1 レビュアの選別

まず，レビューにおける数値評価の変化に着目したイベント検出を行う．このため，6 段階評価における数値の分布を調査した．この結果，レビュアが与える評価値は 4~5 に偏る傾向が強いことが分かった．これは楽天トラベルだけではなく，他のレビューサイトにも共通している性質だということが知られている．

このような結果が得られる原因の一つとして，偏った評価を与えるレビュアの存在が挙げられる．特に不満がない場合，少なくない割合のレビュアが最高得点かそれに近い値を一貫して与えることがある．また，レビュア間の慣例として，そのようなスコアの与え方が推奨されていることもある．

このような偏ったスコアを与えるレビュアからの評価は，正当な評価であるとは考えられないため，イベント検出処理の前に除外したい．このため本研究では，レビュアの与えるスコアからエントロピーを計算し，一定の閾値を基準にレビュアのフィルタリングを

行う．

レビュアが各評価項目に与える評価値の確率を用いてエントロピーを算出する．レビュアがレビューした回数を t, 評価項目 k に対して評価値 i を与える回数を t_{ki} とすると評価項目 k に対して評価値 i を与える確率 $p_{ki} = t_{ki} / t$ となる．よって，レビュアの評価項目 k におけるエントロピーE_k は式(1)となる．

$$E_k = \sum_{i=0}^{5}(-p_{ki} * \log_2 p_{ki}) \quad \text{式(1)}$$

レビュアのエントロピーE_{sum} を E_k の総和として式(2)に定義する．

$$E_{sum} = \sum_{k=0}^{6} E_k \quad \text{式(2)}$$

式(1), 式(2)によって得られた E_{sum} をレビュアのエントロピーとする．

得られたエントロピーを基準に，レビュアの選別を行う．本研究では，エントロピーが 0 でないレビュアを処理の対象とするが，しきい値はデータセットなどに応じて適切に設定する必要がある．

4.2 評価値の時間変化に着目したイベント検出

実世界で何らかのイベントが発生した場合，それがレビュアの評価結果に影響を与え，評価値が影響を受けることが考えられる．このように，評価値の変動に着目すれば、実世界で何らかのイベントの発生を検出することが期待される．評価値によるイベントの検出を行うには，単純にはある施設に関するレビューの評価値を時期ごとにグループ化し，その平均値の変動を利用することが考えられる．しかし，実際には平均値の変動は大きく，それだけを使ってイベントを検出することは難しい．

そこで本研究では、評価値の上下の変動ではなく，ばらつきを検出する．何らかのイベントによって一部のレビューが影響を受けたとすると，それは評価値のばらつきに反映されると考えられる．それを評価値の分散の変動によって検出する．

4.3 レビュー文の時間変化に着目した原因分析

4.3.1 手法の概要

評価値の分散の変動に着目することで，イベントを検出することができた．次は，そのイベントが発生した要因を分析する．これにはレビュー文を利用する．レビュー文を利用した原因分析には，おおまかに次の二つのアプローチが考えられる：

1） イベントが検出された時期に投稿されたレビューのうち，他と比べて評価値が低いレビューのレビュー文を参照する．これは，4.2 節で述べたように，投稿される評価値のうち，ほとんどが高い値を持つことが分かっているため，変動要因は低い評価値の投稿によるものである可能性が高いためである．

2） イベントが検出された時期に投稿されたレビュー文に形態素解析を行って形態素を抽出し，その出現頻度を求める．直前の時期についても同様の処理を行う．直前の時期と比べて頻度が増加したものが原因と関連付いている可能性が高い．ただし，レビュー文の投稿数にばらつきがあるだけでなく，「ホテル」，「宿泊」など，頻繁に使用される用語が多数存在するため，単純に頻度だけを比較することはできない．

以下では，2）の方法を説明する．

4.3.2 レビューデータの処理

提案手法の流れを説明する．

1） 4.1 節で説明した方法でレビュアを選択する．

2） 選択されたレビュアによって投稿されたレビューデータに対して形態素解析を行い，自立語を抽出する．

3） 一定期間の間に投稿されたレビューデータを 1 文書として，TF-IDF あるいはその他の方法で単語の重みづけを行い，その値を元に単語のランキングを行う．

4） イベントが検出された期間と直前の期間のランキングを比較し，ランクが大幅に変化（上昇）した単語をピックアップする．

本研究では，単語の重みづけに Okapi（sBM25）[8]を利用した．具体的な計算式を(3), (4)に示す．ただし，本研究では該当月に投稿された全レビューを R，投稿のあった月の数を N, N の内 w を含む月の延べ数を n(w), R 中の w が出現するレビュー数を f(w, R)，R の文書長を R_{len}, R の平均文書長を $avgR_{len}$ とし，k = 2, b = 0.75 とした．

$$IDF(w) = \log\frac{N - n(w) + 0.5}{n(w) + 0.5} \quad \text{式(3)}$$

$$sBM25(R, w) = IDF(w) * \frac{(k+1) * f(w,R)}{f(w,R) + k * (1 - b + b * \frac{R_{len}}{avgR_{len}})} \quad \text{式(4)}$$

5. 評価実験

提案手法の有効性を実験によって評価した．使用するデータセットは，楽天トラベルのレビューデータのうち 135,199 名によって記述された 177,057 件を対象とした．形態素解析器は Mecab[9]を利用し，自立語のみを抽出した．

表 1：分散に着目した分析における分析対象レビュー文の一例

施設名	投稿年月	評価値	レビュー文
施設 A	2010 年 3 月	1	隣の声や咳が聞こえまくりで眠れたものではなかった。ホテルの本来の目的が…
施設 B	2009 年 12 月	2	部屋の暖房の機器も悪いように感じました。
施設 C	2010 年 1 月	1	あれだけ音の出る壁工事をするなら、部屋周りをもっと…
施設 D	2009 年 12 月	2	26 日朝 8 時過ぎから工事音がうるさく、起こされてしまった。
施設 E	2009 年 11 月	1	一晩中、ホテルの前の工事の音がうるさくて眠れませんでした。
施設 F	2009 年 12 月	2	格安プランを選択したせいか、部屋が狭くて残念でした。
施設 G	2010 年 1 月	1	学生格安プランがあり、深夜まで騒いでいるので

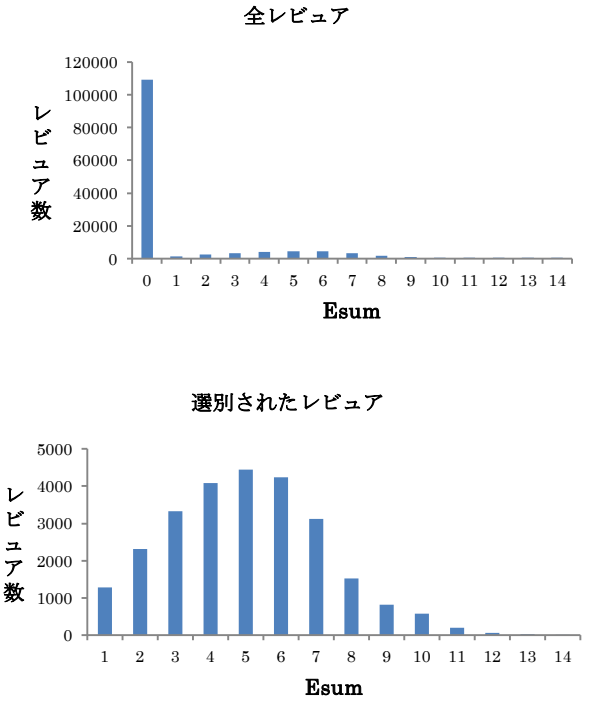

図 1：E_{sum} に対するレビュア数の分布

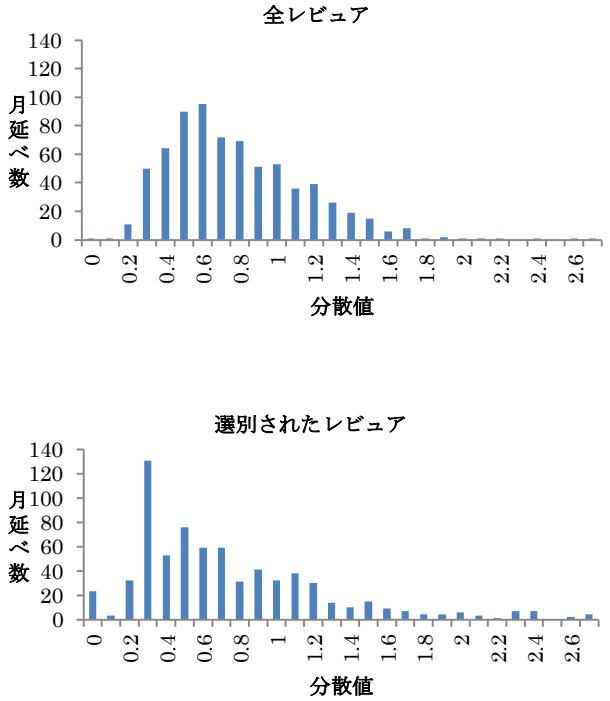

図 2：一月当たりの部屋に関する評価値の分散の頻度分布

5.1 レビュアの選別

エントロピーの算出結果を図 1 に示す．図 1 はエントロピーに応じたレビュアの頻度分布であり，上の図は全レビュアの頻度分布，下の図は $E_{sum} > 0$ のレビュアの頻度分布を示す．横軸は E_{sum}，縦軸はレビュア数を示す．図 1 より，全レビュアのうち，$E_{sum} = 0$ となるレビュアが大半である事が分かる．

次に，宿泊施設に対する評価値の分散に応じた月の延べ数の頻度分布を図 2 に示す．上は全レビュアによるレビューを用いた評価値の分散に応じた頻度分布であり，下は選別されたレビュアによるレビューのみを用いた一月当たりの評価値の分散に応じた頻度分布である．上下の図を比較すると，下の図の方が分散の高い月が比較的多く現れている事が分かる．本研究では評価値が変動した月のレビューを調査するため，選別されたレビュアによるレビューを調査する．

5.2 レビューの分析

5.1 の手続きにより，$E_{sum} > 0$ のレビュア 26,033 名が記述したレビュー67,040 件を抽出した．抽出したレビューに対し，評価値の時間変化に着目した分析とレビュー文の時間変化に着目した分析を行う．

5.2.1 低い評価値に着目した分析

分散の値が他の月に比べて高い月上位 1 割は分散が 1 より大きい月であったため，本研究では分散が 1 より大きい月をイベント発生時期として検出する．検出された月の中で評価値が低いレビュー文の内容について調査した．表 1 に検出例を示す．全ての施設において高い評価値を与えるレビュアが多く，低い評価値を与えるレビューは少なかった．

評価値が低いため，注目するレビューの内容は施設

表 2：得られた形態素の一例

施設名	投稿年月	得られた形態素	前の月からの順位変動
施設 E	2009_11	シングル	122
施設 E	2009_11	風呂	118
施設 E	2010_3	フロント	110
施設 E	2010_3	アメニティ	103
施設 E	2010_3	サービス	106
施設 E	2010_1	プラン	108
施設 F	2010_1	料金	107
施設 F	2010_3	送迎	133
施設 F	2010_3	飛行機	194

A に対して 2009 年 11 月に投稿されたレビュー文や施設 B に対して 2009 年 12 月に投稿されたレビュー文のように，イベントではなく施設の設備が原因の苦情が多い．しかし施設 C~E に投稿されたレビュー文からは，決まった時期に発生しない工事が原因である苦情が観察された．また施設 F・G のレビュー文には，宿泊プランを原因とした苦情が観察された．

観察されたイベントはどちらも定常的に発生するイベントではない．またイベントの観察された月はレビュー投稿件数の最も多い月ではないケースが多い．

5.2.2 単語出現頻度の時間変化に着目した分析

イベントが発生した時期において，イベントを示す形態素は他の時期よりも出現頻度が高くなることが予想される．本研究では，前の月に比べて 100 位以上順位が上がった形態素はイベントを示す形態素として検出する．算出した結果の一部を表 2 に示す．

施設 A に対して 2009 年 11 月に投稿されたレビューでは「シングル」,「風呂」等の順位が上がった．単語「シングル」が記述されているレビュー文を調査すると,「シングルからツインに変更したおかげで，広く快適に過ごせた」という内容が 2 件観察された．また，単語「風呂」からは「お風呂が狭い」という内容が 2 件,「お風呂が広い」という意見が 1 件観察された．同宿泊施設に対して 2010 年 3 月に投稿されたレビュー文では「フロント」,「アメニティ」,「サービス」等の順位が上がった．単語「フロント」が記述されているレビュー文を調査すると「フロントの対応が良かった」という内容が 3 件確認された．単語「アメニティ」からは「アメニティが充実している」という内容が 1 件と,「女性用のアメニティが不足している」という内容が 1 件観察された．単語「サービス」からは「宿泊施設のサービスが行き届いて居て満足できた」という内容が 5 件観察された．

施設 B に対して 2010 年 1 月に投稿されたレビューからは「プラン」,「料金」等が上位に上がった．単語「プラン」が記述されているレビュー文からは「特典付きのプランがあったので宿泊する事にした」という内容が 2 件観察された．単語「料金」が記述されているレビュー文からは,「宿泊料金が手頃である」という内容が 2 件と,「駐車料金が安いので良い」という内容が 1 件観察された．同宿泊施設の 2010 年 3 月には単語「送迎」,「飛行機」等の順位が上がった．単語「送迎」,が記述されているレビュー文から「空港までの送迎バスのサービスが良い」という内容が 3 件観察された．単語「飛行機」からは「部屋から飛行機が見える」という内容が 2 件観察された．

6．考察

エントロピーを用いる事で，評価が偏っているレビュアによる評価への悪影響を除外することができた．選別されたレビュアによる評価を用いて各宿泊施設に対して月毎に分散を取って分布を調査すると，分散の大きい月が比較的増加した事が確認された．レビュアの選別によって，評価値が変動した月を発見しやすくなったと考えられる．一方，近年ではレビューサイト「食べログ」等で，ステルスマーケティングによる評価の不正操作が問題視されている．ステログは食べログにおいて不正評価を行うレビュアを判別する web サイトとして注目されており，投稿回数・投稿字数を用いてレビュアの信頼性を評価している．本研究におけるエントロピーの利用目的は偏ったレビュアを除外することである．エントロピーを用いてレビュアの信頼性を評価する場合，投稿されたレビューの字数を考慮した評価も今後検討したい．

評価値に着目した分析では，イベントを原因とする苦情が複数観察された．記述されている内容はどれも定常的に発生するイベントではない．ただし，観察されたイベントである工事やプランは必ずしも一ヶ月以上の間継続するイベントであると言えず，実際にイベントが発生していた期間は一ヶ月未満の可能性がある．評価期間を検討する事で，イベント検出精度が上がる可能性がある．

単語の出現頻度に着目した分析において，検出された形態素が記述されているレビュー文の内容は複数観察された．内容は類似している事が多い．「お得なプラ

ン」等のイベントを示すレビューは少なく,「お風呂が広い」等のイベント以外の内容が多く観察された．またレビューの内容は宿泊施設に対して満足感を示す事が多く，苦情に関するレビュー内容は少なかった．評価値に着目した分析結果とは異なる結果が得られた．出現頻度に着目した分析では，イベントに繋がる単語が検出されていないため，分析手法の改良を行う予定である．検出された単語の中には「部屋」や「残念」,「お願い」等，単体では何を示すか推測できない名詞が観察された．検出すべき単語を精査する事で，イベントを示す単語を正確に抽出できる可能性がある．

7. 結論

本研究は，レビュアが宿泊施設に対して与える評価値のエントロピーを用いて有用なレビューを抽出し，レビューを時系列順に並べた時の評価値の分散の変化に着目し，イベントの発生した時期を推定した．低い評価値に着目する事で宿泊施設に対する評価値が時系列的に変化した原因を抽出できる事を示した.しかし，単語の出現頻度に着目した分析では原因は抽出されなかった．

今後の課題を以下に述べる．

本研究では月単位でイベントの発生時期の推測を行ったが，発生するイベントは長期的に影響を与えるものばかりではなく，短い期間のみ発生する可能性も考えられる．短い期間発生するイベントを発見するため，時間の粒度を検討する．

また単語の出現頻度に着目した分析ではイベントの発見は困難であるため，分析手法の検討を行う．

謝辞

貴重なデータを提供していただいた楽天株式会社に深い感謝の意を表する．

本研究の一部は科学研究費補助金基盤研究(A)(#21240005)による．

参考文献

[1] 小倉達也，宍戸開，今藤紀子，山口実靖，淺谷耕一　「レビューサイトにおけるレビューの特性とそれを考慮した評判情報の抽出に関する一考察」　DEWS2008　2008

[2] 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 2006

[3] 小林のぞみ，乾健太郎，松本裕治，立石健二，福島俊一　「意見抽出のための評価表現の収集」自然言語処理，Vol.12，No.2　2005

[4] 藤井　絵美子，上野　剛，中本　政一，東　高宏，加藤　直樹，羽室　行信　「ホテル業界におけるロコミ情報に基づいた顧客満足度予測モデルの構築とポジショニング分析」　WebDB Forum 2011

[5] 楽天データ公開 http://rit.rakuten.co.jp/rdr/index.html

[6] 楽天トラベル http://travel.rakuten.co.jp/

[7] Marilena Oita, Pierre Senellart “Deriving Dynamics of Web pages: A Survay”　TWAW2011 2011

[8] S E Robertson, S Walker, S Jones, M M Hancock-Beaulieu “Okapi at TREC-3”　TREC 1994

[9] MeCab（和布蕪） http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html