

# レビュー時系列データからの分割統治による変化点検出法

The Change-Point Detection by Dividing and Conquering in Time Series Data of Reviews

山岸 祐己\*<sup>1</sup> 齊藤 和巳\*<sup>1</sup> 大久保 誠也\*<sup>1</sup>  
Yuki Yamagishi Kazumi Saito Seiya Okubo

\*<sup>1</sup>静岡県立大学大学院経営情報イノベーション研究科  
Graduate School of Management and Information of Innovation, University of Shizuoka

In recent years, those who write a review at online review sites are increasing. And we are considering a modeling of people's such evaluation action. However, there are a variety of scores, documents, and images on those sites, from the pure rating to the intentional or indifferent. Thus, in this paper, we propose a method for change-point detection in time series data of online review sites. The proposing method assumes a multinomial distribution model as a user's basic evaluation action, and is characterized by detecting an unusual period by a likelihood ratio test. Furthermore, in order to detect two or more unusual periods, this technique was performed in dividing and conquering. From an experimental result, it is shown that the periods when unreliable users posted reviews intensively have been detected.

## 1. はじめに

オンラインレビューサイトとは、商品やサービスについてのレビューを投稿することができるウェブサイトの総称である。オンラインレビューサイトについては、多様な分析や研究が展開されている [1]。現在、インターネットの爆発的な普及により、ネットショッピングの一般化と共にレビューサイトのユーザーが急増し、日々大量のレビューがあらゆるサイトに投稿されている。結果、1つの商品に対して多種多様なレビューが付くこととなり、有益なレビューを判別することが難しくなったため、レビューに付随する評点の平均点が一般的な評価指標として扱われるようになった。

しかし、殆どのレビューサイトが投稿回数制限やレビュー内容の吟味を行っていないため、主観的思考が強いユーザーによる極端な評価が書かれたレビューも飛び交い、この評点平均ですら信頼性を失いつつある。あまりにも肯定的なレビューは、商品の製造会社や関係会社が意図的に書いたのではないかと疑われ、あまりにも否定的なレビューは、競合他社や個人の嫌がらせとして見做される場合もある。さらには、金銭を受け取って好意的なレビューを書いたり書かせたりする「さくら」や「やらせ業者」の特定も相次いでいるため、オンラインレビューサイトに対する不信感は益々強まるばかりである。従って、オンラインレビューサイトにおけるレビューの変化点検出は重要な研究課題と言える。

本論文では、Swan と Allan [2] や Kleinberg [3] と同様に、回顧的 (Retrospective) な立場で異常を検出する新たな手法を提案する。我々は既に、ユーザーの基本評点行動として多項分布モデルを仮定し、尤度比検定により異常期間を検出することを特徴とする単一区間抽出法を提案している [4]。本稿では、この手法を拡張し、複数区間の抽出を可能とする手法を新たに提案する。

本論文の構成は以下となる。まず、評点時系列データから異常期間を検出する提案法について説明する。次に、実験で用いたデータセットの詳細を述べると共に、実験結果を報告する。最後に、本研究のまとめについて述べる。

## 2. 提案法

本章では、読者の便宜を計るため、既に提案している単一区間抽出法 [4] について説明した後に、本稿で新たに提案する複数区間抽出法について述べる。

### 2.1 単一区間抽出法

評点の時系列データを以下とする。

$$\mathcal{D} = \{(a_1, t_1), \dots, (a_N, t_N)\}. \quad (1)$$

ここで各評点は、1 から  $J$  の整数値で与えられるとする。即ち、 $a_n \in \{1, \dots, J\}$  となる。 $t_n$  はそれぞれの評点を与えられた時刻を指す。モデル記述の都合上、各評点  $a_n$  を以下のように  $J$ -次元ベクトルとしてダミー変数を導入する。

$$a_{n,j} = \begin{cases} 1 & \text{if } a_n = j; \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

いま、多項分布モデルを仮定し、評点  $j$  が与えられる確率を  $p_j$  とすれば、評点の時系列データの対数尤度関数は次式となる。

$$\mathcal{L}(\mathcal{D}; \mathbf{p}) = \sum_{n=1}^N \sum_{j=1}^J a_{n,j} \log p_j. \quad (3)$$

この尤度関数に対して最尤推定量を求めれば、以下となる。

$$\hat{p}_j = \frac{\sum_{n=1}^N a_{n,j}}{N} \quad (4)$$

一方、区間  $S = [t_u, t_v]$  が異常期間であるとして、通常とは違う多項分布に従うと考える。そこでは評点  $j$  が与えられる確率を  $q_j$  とし、区間  $S$  以外では評点  $j$  が与えられる確率を  $r_j$  とする。ここで、区間  $S$  に入る評点データを記述するため、以下の集合を導入する。

$$X(S) = \{u, u+1, \dots, v\}. \quad (5)$$

このとき、評点時系列データの対数尤度関数は次式となる。

$$\mathcal{L}(\mathcal{D}; \mathbf{p}, \mathbf{q}, S) = \sum_{n \in X(S)} \sum_{j=1}^J a_{n,j} \log q_j + \sum_{n \notin X(S)} \sum_{j=1}^J a_{n,j} \log r_j. \quad (6)$$

この尤度関数に対して最尤推定量を求めれば、以下となる。

$$\hat{q}_j = \frac{\sum_{n \in X(S)} a_{n,j}}{|X(S)|}, \quad \hat{r}_j = \frac{\sum_{n \notin X(S)} a_{n,j}}{N - |X(S)|}. \quad (7)$$

本提案法では、最も顕著な異常区間として、尤度比の対数を最大にする区間  $S$  を求める。即ち、次式を最大にする  $\hat{S}$  である。

$$\hat{S} = \arg \max_S \{\mathcal{L}(\mathcal{D}; \hat{\mathbf{p}}, \hat{\mathbf{q}}, S) - \mathcal{L}(\mathcal{D}; \hat{\mathbf{p}})\}. \quad (8)$$

今回の実験では、区間  $S$  の候補として、評点時系列データの任意の観測時間のペア  $[t_u, t_v]$  を考える。ここで、

$$A = \max\{\mathcal{L}(\mathcal{D}; \hat{\mathbf{p}}, \hat{\mathbf{q}}, S) - \mathcal{L}(\mathcal{D}; \hat{\mathbf{p}})\}, \quad (9)$$

とし、この  $A$  を区間  $\hat{S}$  における異常の度合いとして扱う。

一般に、データ数  $N$  が十分に大きいとき、 $A$  の 2 倍は漸近的に  $\chi^2$  分布となることが知られているため、予め設定した有意水準における自由度  $J - 1$  の  $\chi^2$  の棄却点を  $2A$  が超える場合は、「 $\hat{S}$  は異常区間ではない」という帰無仮説が棄却されることが示唆される。

## 2.2 複数区間抽出法

複数区間抽出法では、単一区間抽出法において帰無仮説を棄却するような  $\hat{S}$  が検出されたとき、 $\mathcal{D}$  を区間  $\hat{S}$  と  $\hat{S}_1$  により区切られた 2 区間の計 3 区間 ( $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$ ) に分割し、各区間に対して式 (5) から式 (9) までの手順を同様に行い、新たに  $\hat{S}_2$  などの検出を繰り返す。すなわち、新しい  $\hat{S}_i$  が設定された棄却点を超え続ける限り、この操作は繰り返される。このような手法を分割統治法 (Divide and conquer algorithm) と呼ばれ、そのままでは解決できない問題を小問題に分割することで最終的に問題を解決しようとする考え方として知られている。

## 3. データセット

今回使用するデータセットは、食べログ (<http://tabelog.com/>) のレストランレビューデータである。

食べログ (たべログ) とは、カカクコムグループ (<http://corporate.kakaku.com/>) が運営するグルメサイトであり、2005 年 3 月にサービスが開始された。このサイトには、ユーザー及びレビュー代行業者による全国のレストランの 5 段階評価が掲載されている。ユーザーは、アカウントを作成すると、「口コミ」と呼ばれるレストランのレビュー情報や画像の投稿ができ、後から口コミを修正することもできる。口コミの採点は「総合」の他に、「料理・味」「サービス」「雰囲気」「CP(コストパフォーマンス)」「酒・ドリンク」の 5 項目でも行われる。さらに、ユーザーは、「参考になった」と思った他のユーザーのレビューに対し「参考票」なるものを 1 票投じることができ、この「参考票」の数がレビューの信頼度として扱われている。店舗の 5 段階の評価指標は、ユーザーの

口コミ採点により上下変動するが、この評価指標に関する厳密な計算式は食べログより公開されていない。また、ユーザー登録してから一定回数以上の投稿を行った後でないと、店舗の評価に反映されず、極端に低い点数をつけることもできない。カカクコムはレビューのやらせ問題への対策として、点数算出アルゴリズムの大幅な変更を行ったことを 2012 年 3 月 1 日に発表した。

このデータセットは、2012 年の 1 月から 2 月にかけて食べログをクロールして取得したものであり、449447 レストラン、301086 ユーザー、3114507 レビューを有する。レビュー点数は、1~5 まで 0.5 点刻みで付けることができるため、実験時には点数を 2 倍にして 2~10 の整数値として扱っている。なお、1 人のユーザーが 1 つのレストランに対して口コミを多重投稿することはできない。

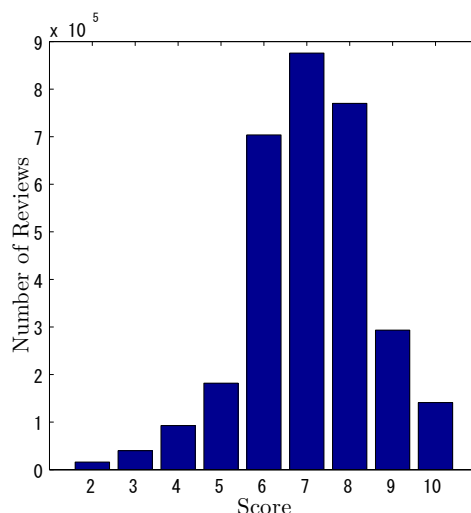


図 1: 食べログのレビュー点数分布

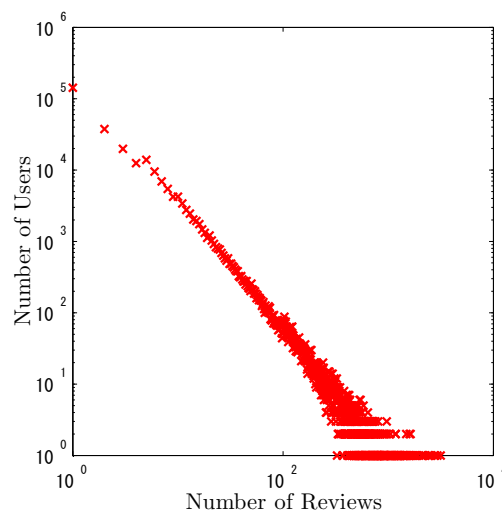


図 2: 食べログのレビュー投稿回数の分布

図 1 にレビュー点数の分布を、図 2 にレビュー投稿回数の分布を、図 3 にレビュー平均点数の分布を示す。図 1 より、レ

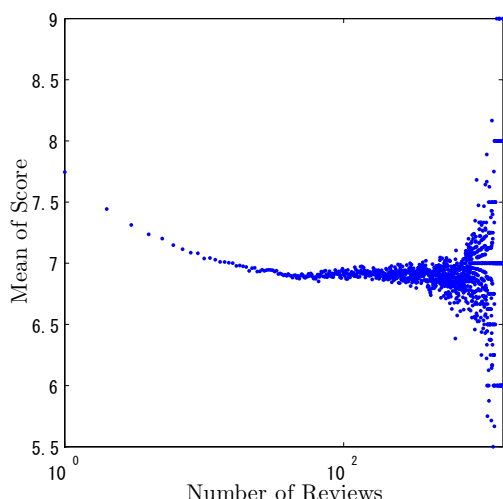


図 3: 食べログのレビュー平均点数の分布

ビュー点数の分布は、点数7のあたりが平均となるように、やや高得点側に偏った正規分布に近い形状となることが分かる。高得点側に偏る要因としては、そもそもユーザは自分が興味を持てるレストランなどに行くことより、自然な傾向と考えられる。一方、図2より、レビュー投稿回数の分布はスケール・フリー性を持つことが見て取れる。すなわち、多様なソーシャル・ネットワークの次数分布と同様な性質を有する。図3からは、投稿回数が少ないユーザー程、高評価のレビューを投稿していることがわかる。これについては、「投稿回数が少ないユーザーは極端に低い評価をつけることができない」という食べログのロジックが少なからず影響しているはずである。

#### 4. 実験結果と考察

今回、 $\chi^2$  検定の危険率は0.05とし、自由度はデータセットのレビュー点数(2~10)より8となるため、設定された棄却点は21.95である。 $\hat{S}$ が異常区間として検出されたレストランは1307件、検出された $\hat{S}$ は1643区間である。検出されたレストランの内、上位10件を表1に示す。図4, 6はレストランに付けられたレビュー点数の推移と $\hat{S}$ を示したもので、縦軸は点数、横軸は最初に投稿されたレビューからの日数である。 $\hat{S}$ は、検出された順に $\hat{S}_i$ としている。図5, 7は $\hat{S}$ おける $\hat{p}$ ,  $\hat{r}$ ,  $\hat{q}$ の確率分布の比較を示したもので、縦軸は確率、横軸は点数である。分割統治法を採用したことにより、図6のように $\hat{S}$ 内で新たな $\hat{S}$ が検出された場合は、 $\hat{S}$ 内の特にどの区間に注目すべきかが分かりやすくなっている。

検出されたレストランの $\hat{S}$ 内のレビューを直接参照すると、雑誌の紹介記事のように整いすぎているレビュー、投稿数が極端に少ないユーザーによる高評価レビュー、レストラン内部の者でしか知らないはずの情報が載せられたレビュー、極端に短文で高評価のレビュー、といった「不自然なレビュー」が数多く見受けられた。

しかし、1つずつレビューを参照していくのでは埒が明かないため、全区間 $\mathcal{D}$ で投稿したユーザー属性値の平均と、 $\hat{S}$ で投稿したユーザー属性値の平均との差分を見ることにした。ユーザー属性値として「平均得点数」「平均参考票」「投稿数」の3項目を選定し、 $A$ の上位96件のレストランを対象として、

各レストランの $\mathcal{D}$ でのユーザー属性値の平均と $\hat{S}_i$ でのユーザー属性値の平均との差( $\hat{S}_i - \mathcal{D}$ )をとった。対象とした96件の $\hat{S}_1$ のみで計算した場合と、 $\hat{S}_i$ 全てで計算した場合の2パターンの結果を表2に示す。結果より、投稿数の平均は、 $\hat{S}$ で投稿を行ったユーザーの方が明らかに少ないことが分かる。食べログの評点ロジック(一定回数以上の口コミ投稿を行わないと店舗の評価に反映されない等)から分かるように、投稿数は運営側からもユーザーからも重要視されており、投稿数そのものが信頼度として扱われることが多いため、 $\hat{S}$ には信頼度の低いユーザーが集中していることが伺える。さらに、僅かではあるが、レビューの信頼度とされる参考票の獲得平均も $\hat{S}$ 内で投稿したユーザーの方が少ない傾向にあることが分かる。

この実験の本来の目的は、不自然なレビューやユーザーを検出することである。しかし、この実験結果には、評価に時期的なブレが無いかを確認するため、対象とするレストランから異常値 $A$ を逆引き的に調べられるという有用性も期待できる。言い換えれば、「 $A$ の低さ」を「評価指標の信頼度」として考えることができるということである。

表 1:  $A$ の上位10レストラン

Rank	Restaurant	$A$	Reviews
1	Belgian Beer Cafe Antwerp Central	86.29	150
2	Rojiura Curry SAMURAI 平岸店	31.08	116
3	たま木亭	30.65	466
4	ブランジェリー コム シノワ	27.74	228
5	一蘭 天神店	25.96	261
6	あつた蓬莱軒 本店	25.94	587
7	ばんや	25.32	334
8	焼肉 芝浦	24.66	163
9	あつた蓬莱軒 松坂屋店	24.64	376
10	福砂屋 長崎本店	24.44	120

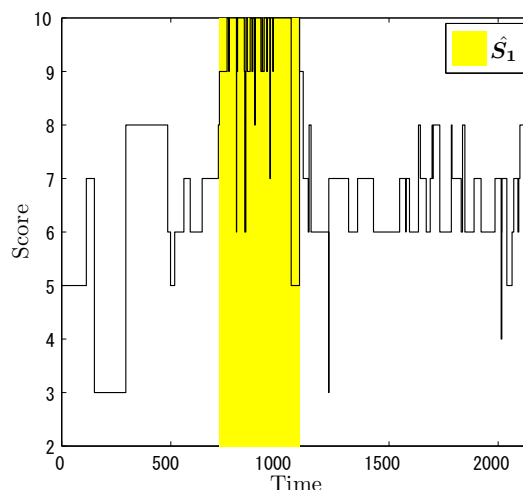


図 4: 「Belgian Beer Cafe Antwerp Central」の点数推移と $\hat{S}_1$

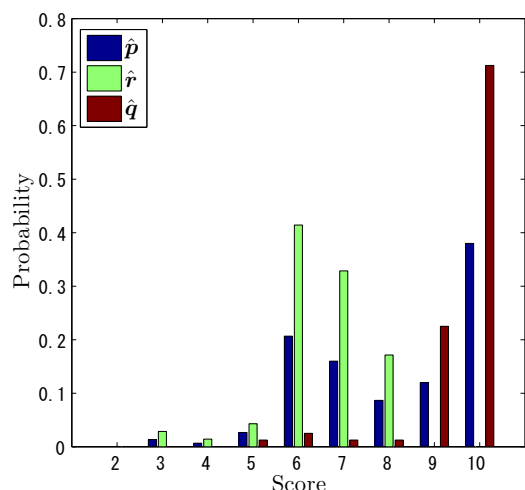


図 5: 「Belgian Beer Cafe Antwerp Central」の確率分布比較 ( $\hat{S}_1$ )

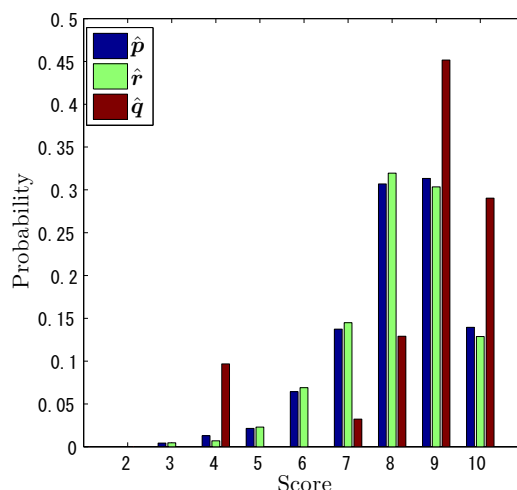


図 7: 「たま木亭」の確率分布比較 ( $\hat{S}_3$ )

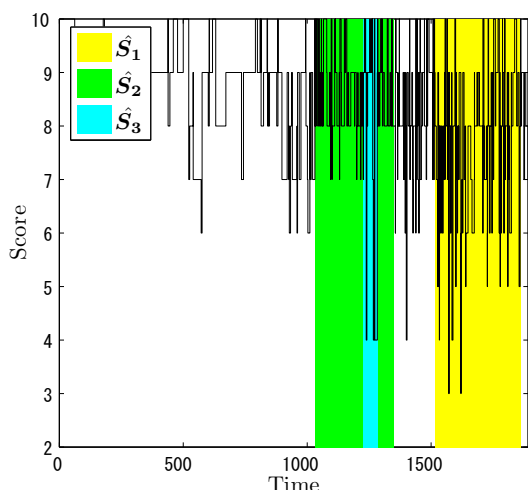


図 6: 「たま木亭」の点数推移と  $\hat{S}$

表 2: ユーザー属性値の平均の差分 ( $\hat{S} - D$ )

	平均得点数 の平均	平均参考票 の平均	投稿数 の平均
A の上位 96 件 の $\hat{S}_1$ (96 区間)	0.017	-0.284	-14.87
A の上位 96 件 の $\hat{S}_i$ (167 区間)	-0.034	-0.298	-12.08

## 5. おわりに

日本の代表的な大規模レストランレビューサイトの時系列データを用い、提案法によって、レストランに対する評価の変化点の検出を試みた。実験により検出された区間内のレビューを直接参照することにより、不自然なレビューが多数含まれていたことが確認できた。また、詳細な検証により、検出された

区間内で信頼度の低いユーザーが集中して投稿を行っていたことを示した。今後は、今回の結果を評点行動モデルの構築に活かすと共に、提案法を応用した評価指標の調整法を検討するつもりである。

## 謝辞

本研究は、科学研究費補助基金基盤研究 (C) (No. 23500312) の支援を受けて行ったものである。

## 参考文献

- [1] M.J.Salganik, P.S.Dodds, and D.J.Watts, "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market", Science 10, pp.854-856, February 2006.
- [2] R.Swan and J.Allan, "Automatic Generation of Overview Timelines", SIGIR 2000, pp.49-56, 2000.
- [3] J.Kleinberg, "Bursty and Hierarchical Structure in Streams", KDD 2002, pp.91-101, 2002.
- [4] 山岸 祐己, 斉藤 和巳, 大久保 誠也, "オンラインレビューサイトの評点時系列データからの異常検出", 第 4 回データ工学と情報マネジメントに関するフォーラム (DEIM2012), 2012.