情報処理学会論文誌　データベース　Vol.10　No.3　58–67 (Oct. 2017)

レビューサイトにおける多項分布に基づくレジームスイッチング検出手法と可視化への応用

山岸 祐己[1,a)]　斉藤 和巳[1,b)]

受付日 2017年3月10日, 採録日 2017年7月5日

概要：大規模レビューサイトにおけるレビューの評点情報は，ユーザの購買行動に大きな影響を与えている可能性がある．本論文では，それらのレビュー評点の付けられ方は，なんらかの理由によって時間とともに変化していると考え，その傾向変化をレジームスイッチングに基づくタイムラインとして可視化する手法を提案する．提案手法は，まず，各レジームにおけるユーザの基本評点行動は多項分布に従っていると仮定し，観測された評点時系列データにおける尤度を最大化することによって，それらのモデルパラメータとスイッチング時刻を推定する．そして，推定されたスイッチング時刻から確率関数を計算し，レジームスイッチングに基づくタイムライン，それに関連した樹形図，および，類似した変化傾向の評点確率関数の可視化結果を生成する．真のモデルパラメータから生成した人工データと，現実の大規模レビューサイトのデータを用いた実験において，提案手法が正確かつ解釈可能な可視化結果を生成できることを示す．

キーワード：時系列データ，レジームスイッチング，多項分布

Detecting Switching Regimes Based on Multinomial Distribution in Buzz Marketing Sites and Its Application for Visualizing

Yuki Yamagishi[1,a)]　Kazumi Saito[1,b)]

Received: March 10, 2017, Accepted: July 5, 2017

Abstract: The review scoring results in large-scale buzz marketing sites can greatly affect actual purchase activities of many users. In this paper, since the scoring tendency for an item usually changes over time due to several reasons, we propose a method for visualizing its scoring time series data as a timeline based on switching regimes. Namely, by assuming that fundamental scoring behavior of users in each regime obeys a multinomial distribution model, we first estimate the switching time steps and the model parameters by maximizing the likelihood of generating the observed scoring time series data, and then produce their timelines, their associated dendrogram, and the groups of similar timelines for scores as our final visualization results by calculating the probability function from the estimated switching regimes. In our experiments using synthetic time series data generated from a known ground truth model and real scoring time series data collected from a Japanese buzz marketing site, we show that our proposed method can produce accurate and interpretable visualization results for such time series data.

Keywords: time series data, regime switching, multinomial distribution

1. はじめに

近年，レビューサイトに投稿されているレビューは日々増加しており，それらのレビューはユーザの購買行動に大きな影響を与えている可能性がある．よって，特に大規模なレビューサイトは，商品やサービスの販売促進をコントロールする重要なメディアであると考えられる．大規模レビューデータを使った既存研究は多様に存在するが，たとえば，レビューサイトのコンテンツをいくつかのトピックに分類しようとする感情分析の試み [1] は，企業が自社の商

[1] 静岡県立大学
University of Shizuoka, Suruga, Shizuoka 422–8526, Japan
[a)] yamagissy@gmail.com
[b)] k-saito@u-shizuoka-ken.ac.jp

品やサービスが消費者によってどのように評価されているかを知るために有用であり，Twitter [2] やビジネス情報 [3] などの分類にも適用されている．

レビューの評価を徹底的に分析するためには自然言語処理が不可欠となるが，幸い多くのレビューサイトではレビューに数値得点を採用しているため，この数値を単純な評価結果として利用することができる．このレビュー評点は，各ユーザから各アイテムへ付けられているため，アイテムごとに評点時系列データが存在している．大規模なレビューサイトにおいて，これらレビュー評点の傾向は，何らかの理由によって時間とともに変化している可能性があるため，評点傾向の変化の原因を探るためには，過去のレビューを参照する必要がある．しかし，企業もしくはユーザが各々で過去の大量のレビューを調査することは困難であるため，レビューサイトにおけるレビュー評点傾向を，解釈しやすい視覚化結果として要約する技術は重要であるといえる．

時系列データの研究では，現時点の状況解析や将来予測に焦点を当てているものもあるが，今回の研究内容は，過去に何が起き，どのような変化をしていたかということに焦点を当てた研究 [4], [5] と類似する．本研究では，レジームスイッチングの検出問題を定式化し，推定されたレジームスイッチングに基づいた評点時系列データのタイムラインと，それに関する樹形図を生成する手法を提案する．ここで，各レジームにおけるユーザの基本評点行動は，多項分布に従っていると仮定し，スイッチングが起こるタイムステップとモデルパラメータは，観測された評点時系列データの尤度を最大化することによって推定する．

実験では，時系列データのバースト検出の最先端技術として Kleinberg の手法 [4] を比較対象とし，提案法の性能と特性を評価する．まず，レジームスイッチングの真のモデルパラメータから生成した人工データを用いた実験において，提案法が Kleinberg の手法よりも様々なケースのスイッチング間隔を検出できていることを示す．次に，現実の大規模レビューサイトのデータを使った実験において，提案法による可視化結果が，Kleinberg の手法による可視化結果よりも解釈が容易であることを示す．

2. 関連研究

本研究は，Kleinberg [4] や Swan ら [5] と同様に，回顧的（retrospective）な枠組みによる時系列データからの構造抽出を目的としている．たとえば，Kleinberg の研究は，文書ストリーム内のトピックの出現をバーストとして表現し，その入れ子構造を推定することによって，ある期間におけるトピックのアクティビティを要約し，それらの分析を容易にしている．この Kleinberg の手法は，バーストが自然に状態遷移として現れる隠れマルコフモデルを使用しており，電子メールメッセージの階層構造を識別することができている．レビューサイトでの適応を考えると，ある期間におけるレビューの投稿間隔やレビュー頻度が変化しているものについては，既存のバースト検出技術 [4] とともに，ウィンドウに基づく手法 [6] や複数ストリームを対象とした手法 [7] なども適応可能であるが，レビュー投稿間隔や頻度がほぼ一定のものについては，これらの既存手法の有効性は低いことが予想される．さらに，既存のバースト検出技術は，単一情報のバーストを検出するものであり，複数情報とその分布の変化に着目していないため，レビュー評点のような複数情報の傾向変化を検出することには向いていない．一方，Swan らの研究は，仮説検定に基づいた時間経過による特徴出現モデルを使用し，コーパス内の主要トピックに対応する情報をクラスタとして生成することに成功している．本研究も同様に，過去に起こった現象を理解するという目的を持っているが，あくまでレジームスイッチングに基づく変化を仮定しているため，このような研究のモチベーションとも離れている．

ここで，今回扱うようなレジームスイッチング検出は，ノベルティ検出や外れ値検出 [8] で使用される技術のような，機械学習の分野で広く研究されている異常検出や変化点検出の典型的技術とは大きく異なることを強調しておく．たとえば，異常検出に使用される統計的手法は，与えられたデータに対して統計モデル（インスタンスの大多数は正常であるという仮定）を適合させ，統計的検定によって未知のインスタンスがこのモデルに属するか否かを決定するものである．このような手法では，適用された統計的検定に基づき，学習モデルから生成される確率が低いインスタンスは異常とされる．本研究は，時間で変化するモデルパラメータをレジームスイッチングとして扱っているため，これらの典型的異常検出技術とは方向性が異なる．同様の方向性を持つ従来アプローチとしては，経済分野におけるレジームスイッチングモデルの研究 [9] があげられるが，これらの研究はガウシアンモデルに大きく依存している．意思決定支援の分野でも，オンラインレビューシステムにおける不正な評価を検出するための技術 [10] がいくつか開発されているが，これらの方法は明確に異常検出技術の領域に分類される．

一般に，レビューサイトでは，レビュー評点の平均点はアイテムの代表的情報として使用されている．確かに，平均点は収束した結果として理解しやすく，レビューの数が多くなるほどその信頼性が増すことも明らかである．しかし，あくまで過去から現在にかけて同じような傾向で評点がつけられていることを仮定したうえでの信頼性であるため，もしどこかの時期で評点分布が急激に変化していたとしたら，その信頼性は大幅に減少してしまう．評点分布の変化の理由としては，ユーザによる評価の偽装（サクラなど），商品やサービスの改善（改悪），流行の変化などが考えられる．このような評点分布の変化理由はアイテムに

とって重要な情報であるが，代表的情報である平均点や，直近のレビューからこれらの情報を推測することは困難である．よって，膨大な数のレビューからレジームスイッチングとして評点分布の変化を推測する研究は重要であるといえる．本研究では，代表的既存研究 [4], [5] と同様に，回顧的な観点で，レビュー評点のレジームスイッチングを可視化する手法を提案する．なお，本研究では評点変化を扱うので，レビューテキストの内容の時系列変化を視覚的に提供する手法 [11] とは対象データが異なる．

3. 提案手法

3.1 問題設定

レビューサイトにおける，あるアイテムの評点時系列データを $\mathcal{D} = \{(s_1, t_1), \cdots, (s_N, t_N)\}$ とする．ここで，s_n と t_n は，J カテゴリの評点と n 番目のレビューの投稿時刻をそれぞれ表す．$|\mathcal{D}| = N$ をこのアイテムに投稿されたレビュー数とすると，$t_1 \leq \cdots \leq t_n \leq \cdots \leq t_N$ となる．n はタイムステップとし，$\mathcal{N} = \{1, 2, \cdots, N\}$ をタイムステップ集合とする．また，k 番目のレジームの開始時刻を $T_k \in \mathcal{N}$，$\mathcal{T}_K = \{T_0, \cdots, T_k, \cdots, T_{K+1}\}$ をスイッチングタイムステップ集合とし，便宜上 $T_0 = 1$，$T_{K+1} = N + 1$ とする．すなわち，$T_1, \cdots, T_K$ は推定される個々のスイッチングタイムステップであり，$T_k < T_{k+1}$ を満たすとする．そして，$\mathcal{N}_k$ を k 番目のレジーム内のタイムステップ集合とし，各 $k \in \{0, \cdots, K\}$ に対して $\mathcal{N}_k = \{n \in \mathcal{N}; T_k \leq n < T_{K+1}\}$ のように定義する．なお，$\mathcal{N} = \mathcal{N}_0 \cup \cdots \cup \mathcal{N}_K$ である．

いま，各レジームの評点分布が J カテゴリの多項分布に従うと仮定する，$\boldsymbol{p}_k$ を k 番目のレジームにおける多項分布の確率ベクトルとし，$\mathcal{P}_K$ はそれら確率ベクトルの集合，つまり $\mathcal{P}_K = \{\boldsymbol{p}_0, \cdots, \boldsymbol{p}_K\}$ とすると，$\mathcal{T}_K$ が与えられたときの対数尤度関数は以下のように定義できる．

$$L(\mathcal{D}; \mathcal{P}_K, \mathcal{T}_K) = \sum_{k=0}^{K} \sum_{n \in \mathcal{N}_k} \sum_{j=1}^{J} s_{n,j} \log p_{k,j}. \tag{1}$$

ここで，$s_{n,j}$ は $s_n \in \{1, \cdots, J\}$ を

$$s_{n,j} = \begin{cases} 1 & \text{if } s_n = j; \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

のように変換したダミー変数である．各レジーム $k = 0, \cdots, K$ と各評点 $j = 1, \cdots, J$ に対する式 (1) の最尤推定量は $\hat{p}_{k,j} = \sum_{n \in \mathcal{N}_k} s_{n,j} / |\mathcal{N}_k|$ のように与えられる．これらの推定量を式 (1) に代入すると以下の式が導ける．

$$L(\mathcal{D}; \hat{\mathcal{P}}_K, \mathcal{T}_K) = \sum_{k=0}^{K} \sum_{n \in \mathcal{N}_k} \sum_{j=1}^{J} s_{n,j} \log \hat{p}_{k,j}. \tag{3}$$

したがって，スイッチングタイムステップの検出問題は，式 (3) を最大化する $\mathcal{T}_K$ の探索問題に帰着できる．

しかし，式 (3) だけでは $\mathcal{T}_K$ の導入によってどれだけ尤度が改善されたかという直接的な評価をすることができない．この問題において，レジームスイッチングを考慮しないときの尤度からの改善度合いを評価することは重要であるため，尤度比最大化問題として目的関数を構築し直す．もし，レジームスイッチングのような変化が存在しない，すなわち $\mathcal{T}_0 = \emptyset$ と仮定すると，式 (3) は

$$L(\mathcal{D}; \hat{\mathcal{P}}_0, \mathcal{T}_0) = \sum_{n \in \mathcal{N}} \sum_{j=1}^{J} s_{n,j} \log \hat{p}_{0,j}, \tag{4}$$

となる．ここで，$\hat{p}_{0,j} = \sum_{n \in \mathcal{N}} s_{n,j} / N$ である．よって，K 個のスイッチングを持つ場合と，スイッチングを持たない場合の対数尤度比は

$$LR(\mathcal{T}_K) = \mathcal{L}(\mathcal{D}; \hat{\mathcal{P}}_K, \mathcal{T}_K) - \mathcal{L}(\mathcal{D}; \hat{\mathcal{P}}_0, \mathcal{T}_0). \tag{5}$$

のように与えられる．最終的に，この問題は上記の $LR(\mathcal{T}_K)$ を最大化する $\mathcal{T}_K$ の探索問題に帰着できる．

3.2 解法

式 (5) を網羅的に解くと最適解が保証されるが，計算量が $O(N^K)$ となってしまうため，ある程度大きい N に対して $K \geq 3$ となってしまうと，実用的な計算時間で解くことができない．したがって，我々は任意の K について解くための高速な解法を提案する．以下では，まず貪欲法（A1）と局所探索法（A2）を説明し，さらにそれらを組み合わせた提案解法について説明する．

3.2.1 貪欲法

まず，貪欲法（A1）の手順について説明する．このアルゴリズムは，バックトラッキングをしないデータの 2 分割の繰返しである．つまり，すでに選択された $(k-1)$ 個のスイッチングタイムステップ $\mathcal{T}_{k-1}$ を固定したまま k 番目のスイッチングタイムステップ T_k を $\mathcal{T}_{k-1}$ に新たに追加することを繰り返す．一般的に，$2(\mathcal{LR}(\mathcal{T}_k) - \mathcal{LR}(\mathcal{T}_{k-1}))$ は，N が十分に大きいとき χ^2 分布に従うことが知られているため，我々はこのアルゴリズムの終了条件として χ^2 検定を採用する．この χ^2 検定の危険率は事前に設定する必要がある．貪欲法アルゴリズムの手順は以下となる．

A1-1. $k = 1$，$\mathcal{T}_0 = \emptyset$ のように初期化する．
A1-2. $T_k = \arg\max_{t_n \in \mathcal{T}} \{\mathcal{LR}(\mathcal{T}_{k-1} \cup \{t_n\})\}$ を探索する．
A1-3. $\mathcal{T}_k = \mathcal{T}_{k-1} \cup \{T_k\}$ のように更新する．
A1-4. もし $2(\mathcal{LR}(\mathcal{T}_k) - \mathcal{LR}(\mathcal{T}_{k-1}))$ が，設定された危険率と自由度 $J-1$ における χ^2 の棄却限界値よりも小さければ，$\mathcal{T}_K$ を出力して終了する．
A1-5. $k = k + 1$ とし，A1-2 に戻る．

ここで，A1-3 での $\mathcal{T}_k$ の各スイッチングタイムステップは，$T_{k-1} < T_k$ を満たすように再インデックスする．明らかに，このアルゴリズムの計算量は $O(NK)$ と高速であるため，大規模な N に対しても実用的な計算時間で結果を得る

ことが可能である．しかし，先ほども説明したように，このアルゴリズムはバックトラッキングを行わないため，プアーな局所解に陥ってしまうことが危惧される．

3.2.2 局所探索法

次に，局所探索法（A2）について説明する．このアルゴリズムは，A1 で得られた解 $\mathcal{T}_K$ から始まり，スイッチングタイムステップの改善を 1 つずつ試みるものである．つまり，k 番目のスイッチングタイムステップ T_k を一度取り去り，残った $\mathcal{T}_K \setminus \{T_k\}$ を固定して，より良い尤度を得られる T'_k を探索することを $k=1$ から K まで繰り返す．ここで，$\cdot\setminus\cdot$ は集合差を表す．もし，すべての k（$k=1,\cdots,K$）に対してスイッチングタイムステップの置換が行われない，すなわち，すべての k に対して $T'_k = T_k$ ならば，これ以上の改善は望めないとして処理を終了する．局所探索法のアルゴリズムは以下となる．

A2-1. $k=1$，$h=0$ のように初期化する．
A2-2. $T'_k = \arg\max_{t_n \in \mathcal{T}}\{\mathcal{LR}(\mathcal{T}_K \setminus \{T_k\} \cup \{t_n\})\}$ を探索する．
A2-3. もし $T'_k = T_k$ ならば $h = h+1$ とし，さもなければ $h=0$ として $\mathcal{T}_K = \mathcal{T}_K \setminus \{T_k\} \cup \{T'_k\}$ のように更新する．
A2-4. もし $h=K$ ならば $\mathcal{T}_K$ を出力して終了する．
A2-5. もし $k=K$ ならば $k=1$，さもなければ $k=k+1$ とし，A2-2 に戻る．

明らかに，このアルゴリズムの計算量は改善が終わらない限り増え続けてしまうが，ある程度大規模な問題に対しても，せいぜい貪欲法アルゴリズムの計算量 $O(NK)$ の数倍程度で終了することを我々はすでに実験によって示している [12]．

3.2.3 提案解法

もし，計算量を最低限に抑えることを目的として，単純に貪欲法アルゴリズムと局所探索法アルゴリズムを組み合わせると，

C1. A1 で $\mathcal{T}_K$ を得る．
C2. A2 で $\mathcal{T}_K$ を改善する．

となる．確かに，これだけでも十分な近似解が期待できるが，スイッチングタイムステップ数 K が貪欲法アルゴリズムによって決定されてしまうため，問題に対して不適切なスイッチングタイムステップ数のまま局所改善を行ってしまう恐れがおおいにある．したがって我々は，不必要なスイッチングタイムステップは極力追加せず，かつ必要なスイッチングタイムステップは極力追加することを目的とした，アルゴリズムの反復的な組合せを提案する．提案解法の手順は以下となる．

P1. A1-1 から処理を開始する．
P2. A1-4 の処理後に $k \geq 2$ ならば，$\mathcal{T}_k$ を $\mathcal{T}_K$ として出力する．
P3. $\mathcal{T}_K$ を A2 で改善し，改善した $\mathcal{T}_K$ を $\mathcal{T}_k$ として出力する．
P4. A1-5 から処理を再開させ，ステップ I2 へ戻る．

この手順では，スイッチングタイムステップが追加されるたびに局所探索法アルゴリズムを行うため，さらなる計算量の増加が予想されるが，ある程度大規模な問題に対しても，せいぜい貪欲法アルゴリズムの計算量 $O(NK)$ の数倍から十数倍程度で終了することを我々はすでに実験によって示している [12]．結局のところ，提案解法において事前に設定が必要となるのは，貪欲法アルゴリズムにおける χ^2 検定の危険率のみであり，これがスイッチングタイムステップ数を大きく左右するため，本手法におけるスイッチング感度のパラメータということになる．

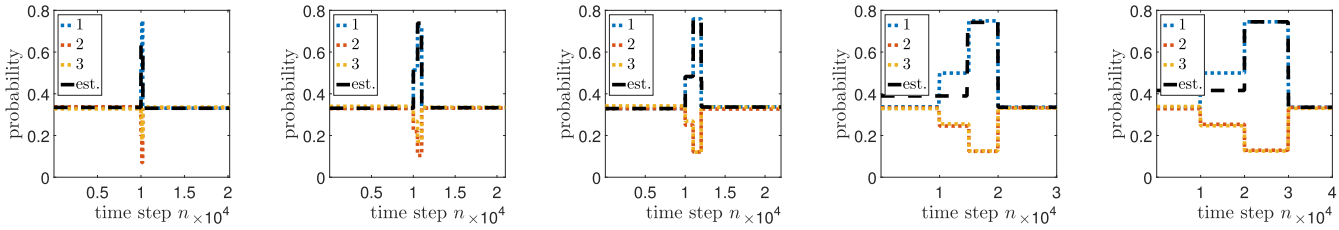
3.3 可視化

上記の解法によって得られた推定タイムステップ集合を $\hat{\mathcal{T}}_K$ とする．各カテゴリ j について，タイムステップ $n \in \mathcal{N}_k$（$0 \leq k \leq K$）における確率関数を $\hat{p}_j(n) = \hat{p}_{k,j}$ のように考える．そして，レジームスイッチングを視覚的に分析するために，J カテゴリの確率関数を同時にプロットしたタイムラインを生成することを考える．ただし，各レビュー評点を同等に扱うために，実際のレビュー投稿時刻 t_n ではなく，タイムステップ n に関する確率をプロットすることに注意されたい．実際のレビュー投稿時刻でプロットされたタイムラインは，レビューのバーストが起こると狭い時間範囲にプロットが埋め込まれて視認性が悪くなったり，長期間レビューが投稿されていない時間範囲もプロットでつなぐため無駄な情報が増えたりしてしまう．さらに，カテゴリ i と j の確率関数間のコサイン類似度に基づいて，以下の非類似度 $d(i,j)$ を考える．

$$d(i,j) = \sqrt{1 - \frac{\sum_{n\in\mathcal{N}} \hat{p}_i(n)\hat{p}_j(n)}{\sqrt{\sum_{n\in\mathcal{N}} \hat{p}_i(n)^2}\sqrt{\sum_{n\in\mathcal{N}} \hat{p}_j(n)^2}}}. \quad (6)$$

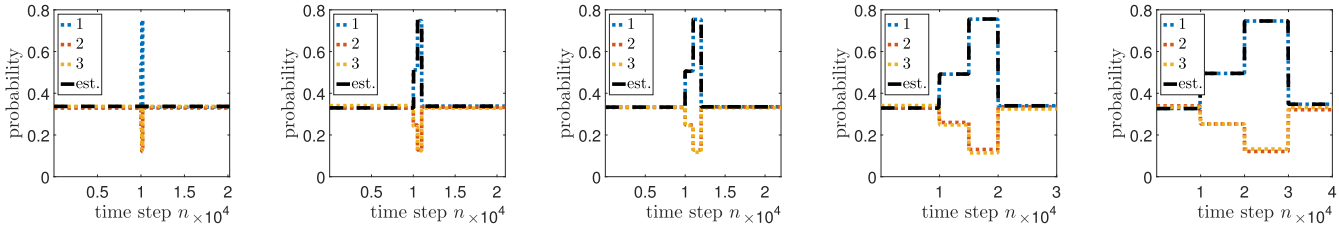
上記の非類似度を用いれば，Ward の最小分散法 [13] によって，これらの確率関数の樹形図を構築することができる．さらに，構築した樹形図に対し，適当なカットオフ点などを導入することで，評点カテゴリを $\{1,\cdots,J\} = G_1 \cup \cdots, G_H$ とクラスタリングし，同一クラスタ G_h に属す確率関数に限定しタイムラインを可視化する手段も提供する．ただし，H は得られたクラスタ数で，$h \neq h'$ に対し $G_h \cup G_{h'} = \emptyset$ である．

以下，本可視化結果のユーザへの提供シナリオについて説明する．一般に評点カテゴリには評価レベルが付随するので，各クラスタ G_h に対し，高評価カテゴリ群というような評価レベルによる解釈付与が可能な場合もある．よっ



(a) $\delta = 100$　(b) $\delta = 500$　(c) $\delta = 1{,}000$　(d) $\delta = 5{,}000$　(e) $\delta = 10{,}000$

図 1　Kleinberg の手法から得られたタイムライン
Fig. 1　Timelines obtained by Kleinberg's method.



(a) $\delta = 100$　(b) $\delta = 500$　(c) $\delta = 1{,}000$　(d) $\delta = 5{,}000$　(e) $\delta = 10{,}000$

図 2　提案法から得られたタイムライン
Fig. 2　Timelines obtained by proposed method.

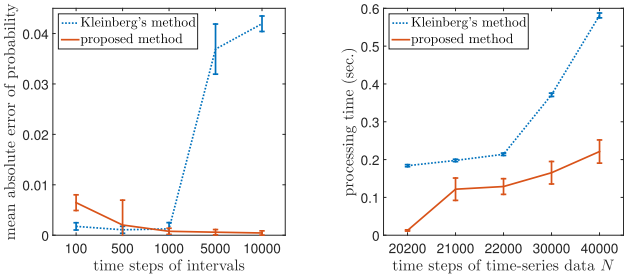
て，同一クラスタに限定した確率関数群のタイムライン可視化により，時間変化にともない，高評価カテゴリ群の段階的な減少，低評価カテゴリ群はほとんど変化しないなどの解釈付与も期待できる．すなわち，このようなシナリオなどを想定しない従来法とは一線を画し，本可視化法により，多様なユーザを対象に，レビュー評点傾向の解釈を容易にし，意思決定の支援に効果的に活用可能にすることを目的としている．なお本研究目的は，企業での商品評価分析やレビューサイトの公正運営などを目的とした，ある程度の知識を持った専門ユーザだけでなく，レビューサイトを利用する一般ユーザなどにも有用なタイムライン可視化法の基本技術の確立である．よって，一般ユーザを対象とした被験者実験により，本提案技術の有効性を実証評価する．

4.　実験結果

4.1　人工データによる実験

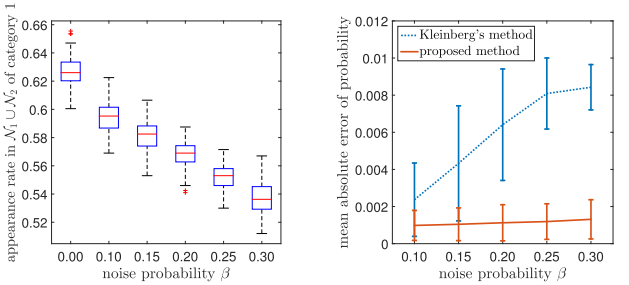
異なるアプローチによる時系列データの構造抽出手法の代表的かつ最先端の技術として，隠れマルコフモデルを用いた Kleinberg の手法 [4] を比較対象とし，提案法がレジームスイッチングの真のモデルを推定できるか実験する．ここでは，3 カテゴリの多項分布による単純なレジームスイッチングモデルを構築し，そのモデルによって生成された時系列データを用いて，両手法が真のモデルを正確に復元できるかどうかを調べる．より具体的には，真のスイッチングタイムステップ数を $K = 3$ とし，$\mathcal{T}_K^* = \{T_0^* = 1,\ T_1^* = 10{,}000,\ T_2^* = T_1^* + \delta,\ T_3^* = T_2^* + \delta,\ T_4^* = T_3^* + 10{,}000\}$ によって定義される各レジームの真のタイムステップ集合を考える．ここで，レジーム間隔 δ は今回 100，500，1,000，5,000，10,000 の 5 パターンに設定しているため，各 δ の最終タイムステップ N はそれぞれ 20,200，21,000，22,000，30,000，40,000 である．多項分布の確率設定は，最初と最後のレジーム（$\mathcal{N}_0$，$\mathcal{N}_3$）はすべてのカテゴリで同じ，すなわち $p_{0,1}^* = p_{0,2}^* = p_{0,3}^* = p_{3,1}^* = p_{3,2}^* = p_{3,3}^* = 1/3$ とし，2 番目と 3 番目のレジーム（$\mathcal{N}_1$，$\mathcal{N}_2$）は第 1 カテゴリだけ $p_{1,1}^* = 1/2$，$p_{2,1}^* = 3/4$，それ以外のカテゴリは $p_{1,2}^* = p_{1,3}^* = (1 - p_{1,1}^*)/2$，$p_{2,2}^* = p_{2,3}^* = (1 - p_{2,1}^*)/2$ とした．特に，第 1 カテゴリの確率設定は，Kleinberg のバーストスケール $\alpha = 3/2$ に沿ったもの（最初と最後のレジームでの確率と比較して，2 番目のレジームでは確率が α 倍，3 番目のレジームでは確率が α^2 倍）となっているため，Kleinberg の手法を第 1 カテゴリのみからなる時系列データ，すなわち $\mathcal{D}_1 = \{(s_n, t_n) \in \mathcal{D}; s_n = 1\}$ にのみ適用した．なお，Kleinberg のパラメータ設定は $\alpha = 1.5$，$\gamma = 1.0$，提案法の χ^2 検定の危険率は $p = 0.0001$ とした．

図 1 に Kleinberg の手法から得られたタイムラインを，図 2 に提案法から得られたタイムラインをそれぞれ示す．各カテゴリの破線は真のモデルで定義されたレジームスイッチングに基づいて計算された確率関数である．黒の破線は，Kleinberg の手法においては，$\mathcal{D}_1$ から得られたバーストレベル集合を用いて計算した第 1 カテゴリの確率関数であり，提案法においては，得られたスイッチングタイムステップを用いて計算した第 1 カテゴリの確率関数である．これらの図より，中程度の間隔のスイッチング（$\delta = 500$ または 1,000）では両手法とも高い精度でタイムラインを推定することができている（図 1 (b)，1 (c)，2 (b)，2 (c)）



(a) 推定誤差の比較結果　(b) 計算時間の比較結果

図 3　Kleinberg の手法と提案法の性能比較
Fig. 3　Performance comparison of Kleinberg's method and proposed method.



(a) ノイズ確率 β 毎の $\mathcal{N}_1 \cup \mathcal{N}_2$ における第 1 カテゴリーの発生確率　(b) ノイズ確率 β 毎の推定誤差の比較結果

図 4　ノイズを用いたときの Kleinberg の手法と提案法の性能比較
Fig. 4　Performance comparison of Kleinberg's method and proposed method when using noise.

が，長めの間隔のスイッチング（$\delta = 5{,}000$ または $10{,}000$）では Kleinberg の手法がスイッチングタイムステップ数を間違えており（図 1 (d)，1 (e)），最も短い間隔のスイッチング（$\delta = 100$）では提案法がスイッチングタイムステップを発見できなかった（図 2 (a)）．これらの結果より，今回の問題設定で期待されるような短かすぎないスイッチング間隔の場合は，提案法の方が Kleinberg の手法よりも優れていることが示唆されるが，Kleinberg の手法は非常に短い間隔のスイッチング，すなわちバーストを検出するという点において提案法よりも優れていることが分かる．

推定されたタイムラインの精度を定量的に評価するため，得られた第 1 カテゴリの確率関数の，真のモデルとの平均絶対誤差を以下のように導入する．

$$E(\hat{p}_1(n)) = \frac{1}{N}\sum_{k=0}^{K}\sum_{T_k^* \le n < T_{k+1}^*} |p_{k,1}^* - \hat{p}_1(n)|. \qquad (7)$$

図 3 (a) は Kleinberg の手法と提案法による推定結果の平均絶対誤差を示したものであり，プロットは各モデルで独立に生成した 100 サンプルによる試行の平均と標準誤差を表している．この図から，上記の考察を定量的に確認することができる．すなわち，$\delta = 500$ または $1{,}000$ では両手法とも誤差が小さいが，$\delta = 5{,}000$ または $10{,}000$ では提案法のみ誤差が小さく，$\delta = 100$ では Kleinberg の方が誤差が小さい．図 3 (b) は両手法の 100 サンプルによる試行の計算時間の平均と標準誤差をプロットしたものである．図より，今回の実験の場合は，提案法のほうが高速に問題を解いていることが分かる．

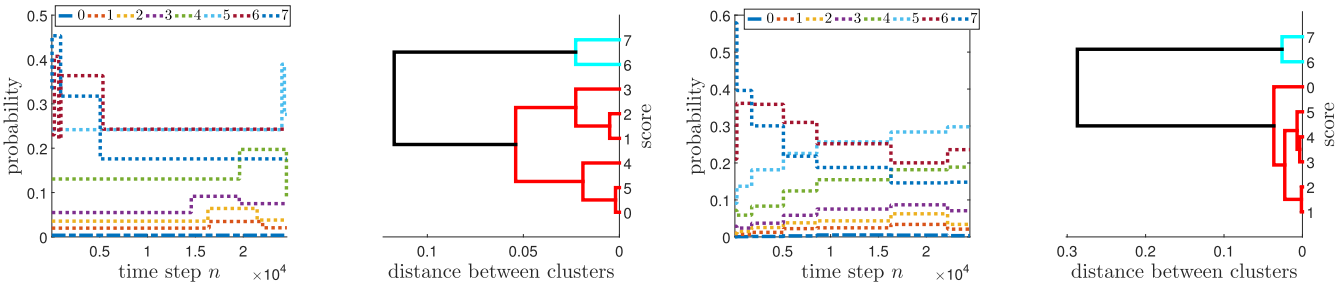
さらに，データにノイズが含まれる場合での精度について検証する．ノイズ発生確率 β を設定したとき，各タイムステップ $n \in \mathcal{N}_1 \cup \mathcal{N}_2$ において，確率 β で多項分布の確率設定を $\mathcal{N}_0$，$\mathcal{N}_3$ と同様に $p_{\cdot,1}^* = p_{\cdot,2}^* = p_{\cdot,3}^* = 1/3$ としてデータを生成する．すなわち，区間 $k \in \{1,2\}$ での各カテゴリ $j \in \{1,2,3\}$ へのノイズ付き確率は $(1-\beta)p_{k,j}^* + \beta/3$ である．本実験では，δ は両手法において精度が安定していた 1,000 とし，ノイズ確率 β は 0.10，0.15，0.20，0.25，0.30 の 5 パターンとした．なお，各ノイズパターンにおいて独立に 100 サンプルずつデータを生成している．**図 4** (a) はノイズ確率 β ごとの $\mathcal{N}_1 \cup \mathcal{N}_2$ における第 1 カテゴリの発生確率を示した箱ひげ図である．図より，ノイズ確率 β の増加にともなって，第 1 カテゴリの発生確率が低くなっていることが分かる．図 4 (b) はノイズ確率 β を導入したときの推定誤差を式 (7) に基づき比較評価した結果である．図より，提案法は Kleinberg の手法よりもノイズの影響を受けにくく，精度が安定していることが分かる．

4.2　現実データによる実験

ここでは，日本の大規模化粧品レビューサイトの “@cosme” [*1] から収集したレビューデータの中で，最もレビューが投稿されていた 2 アイテム（“Oshima Tsubaki Camellia Hair Care Oil (*Oshima Tsubaki*)” と “Conditioner Essential (*Albion*)”）のレビューデータを用いて実験を行う．実験におけるパラメータ設定は先ほどと同様だが，データのレビュー評点（0 から 7）に従い $J = 8$ カテゴリとして扱う．

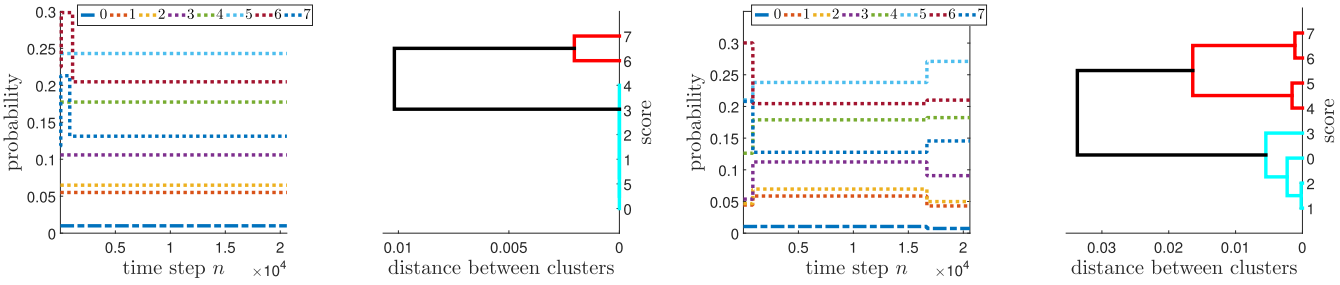
図 5 に Oshima Tsubaki のデータから得られた両手法のタイムラインとそれに関する樹形図を示す．ただし，Kleinberg の手法の場合は各カテゴリデータ $\mathcal{D}_j$ を個別に適応し，対応する確率関数を推定しているため，各タイムステップで確率の合計は 1 になっていないことに注意されたい．図 5 (a) から (d) より，樹形図については，高評価カテゴリ $j \in \{6,7\}$ がグループとして他のカテゴリから大幅に分離されているという点において，図 5 (b) と (d) は，似たような結果を出力しているが，タイムラインに至っては提案法の方が明らかに全体のカテゴリ傾向を把握しやすいことが見て取れる．一方，**図 6** (a) と (c) より，Albion のデータにおいてはバーストやスイッチングがあまり検出されていないため，Albion は過去から現在にかけてレビュー評点の傾向にほとんど変化がないことが示唆される．しかし，図 6 (b) と (d) に示されるように，後者ではカテゴリ

[*1] http://www.cosme.net/



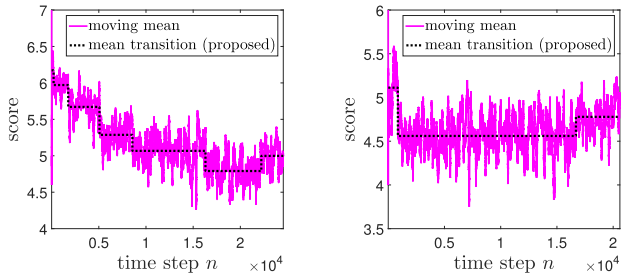
(a) Kleinberg の手法によるタイムライン　(b) Kleinberg の手法による樹形図　(c) 提案法によるタイムライン　(d) 提案法による樹形図

図 5　Oshima Tsubaki の推定タイムラインと樹形図
Fig. 5　Timelines and dendrograms for Oshima Tsubaki.



(a) Kleinberg の手法によるタイムライン　(b) Kleinberg の手法による樹形図　(c) 提案法によるタイムライン　(d) 提案法による樹形図

図 6　Albion の推定タイムラインと樹形図
Fig. 6　Timelines and dendrograms for Albion.



(a) Oshima Tsubaki のレビュー評点の移動平均　(b) Albion のレビュー評点の移動平均

図 7　レビュー評点の移動平均
Fig. 7　Moving means of scores.

群 $j \in \{4,5\}$ が陽に分離されているなど，カテゴリ分布の変化が乏しい場合でも，提案法は解釈しやすい樹形図を生成できていることが分かる．

提案法から得られた結果を詳細に分析するため，**図 7** (a) と (b) にレビュー評点の移動平均と提案法に基づく評点平均を示す．より詳細には，図中の実線はウィンドウサイズ 100 で計算した実データの移動平均を，破線は提案法で得られたレジーム内の評点平均を表している．両図より，提案法は実データの移動平均と自然に一致するようなレジームスイッチングを推定できていることが分かる．

4.3　被験者実験

本研究の目的は，レビューサイトを利用する一般ユーザなどにも有用なタイムライン可視化法の基本技術の確立であるため，レビューサイトや数理統計学に関する専門的な知識を持たない大学生 13 人に対して被験者実験を行った．今回の被験者実験の流れは以下のとおりである．

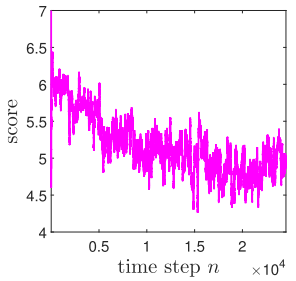
Exp 1.　対象アイテムの，ウィンドウサイズ 100 で計算したレビュー評点の移動平均の図を実験協力者に見せる．

Exp 2.　H クラスタに分割した両手法のタイムラインの図 $G_1, \cdots, G_H$（計 $2H$ 個）を，手法による違いが分からないように，ランダムな配置で実験協力者に見せる．

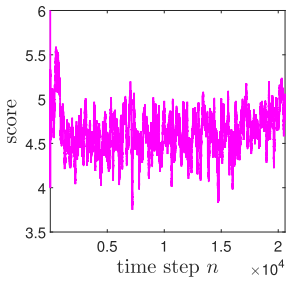
Exp 3.　移動平均の変動についての説明を，各タイムラインの図を参考にして，参考にした図がどれか分かるように，自由記述で実験協力者に書いてもらう．このとき，説明で用いる図の数の制限はせず，参考となる図がなかった場合は，図をいっさい用いない説明も許可している．

Exp 4.　全実験協力者の自由記述の内容から，タイムラインの図が用いられた数を手法ごとに集計する．

なお，実験で用いられた各図の軸や線の意味については詳細な解説が付与されている．今回の対象アイテムは，先ほど扱った Oshima Tsubaki と Albion であり，実験手順 Exp 1. において，それぞれのレビュー評点の移動平均を**図 8** のように表示した．また，タイムラインのクラスタ数は $H = 3$，すなわち $\{0, \cdots, 7\} = G_1 \cup G_2 \cup G_3$ となるよ



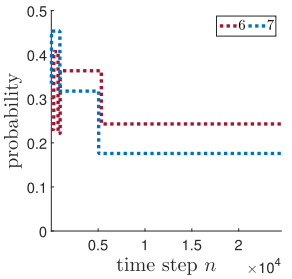
(a) Oshima Tsubaki



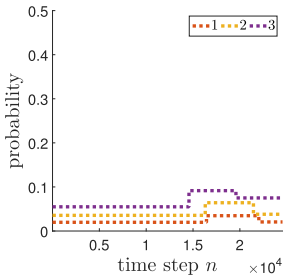
(b) Albion

図 8　ウィンドウサイズ 100 で計算したレビュー評点の移動平均

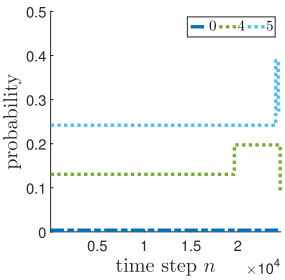
Fig. 8　Moving means calculated by window size 100.



(a) $G_{a,1} = \{6, 7\}$ のタイムライン



(b) $G_{a,2} = \{1, 2, 3\}$ のタイムライン



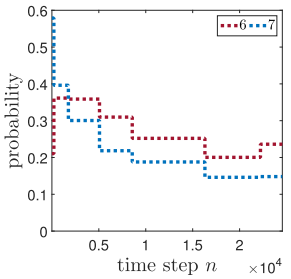
(c) $G_{a,3} = \{0, 4, 5\}$ のタイムライン

図 9　Kleinberg の手法によるクラスタリング後のタイムライン（Oshima Tsubaki）

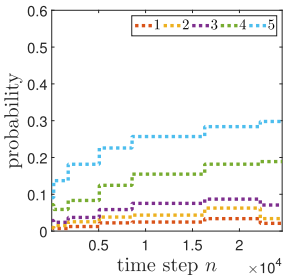
Fig. 9　Timelines of Kleinberg's method which divided by clustering results (Oshima Tsubaki).

うに各樹形図をカットオフし，Oshima Tsubaki を $G_{a,h}$，Albion を $G_{b,h}$ として，$G_{*,h}$ ごとのタイムラインを実験手順 Exp 2. において**図 9** (a)，**図 10**，**図 11**，**図 12** (c) のように表示した．

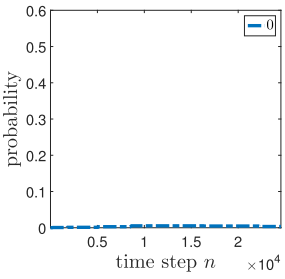
被験者実験の結果として，各タイムラインが自由記述で用いられた回数とその合計を，手法ごとに分けて**図 13** に示す．図 13 (a) より，Kleinberg の手法による $G_{a,3}$（図 9 (c)）は一度も採用されておらず，提案手法の $G_{a,1}$（図 10 (a)）と $G_{a,2}$（図 10 (b)）はともに採用数が多いことが分かる．また，図 13 (b) より，評点分布の変化が乏しい Albion においては，タイムラインが採用された合計回数に大きな違いは見られないが，評点分布の変化が激しい Oshima Tsubaki においては，提案手法の方が明らかに合計回数が多いこ



(a) $G_{a,1} = \{6, 7\}$ のタイムライン



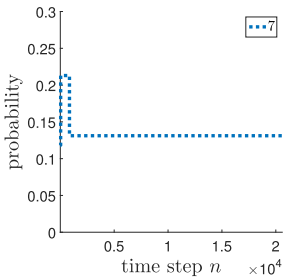
(b) $G_{a,2} = \{1, 2, 3, 4, 5\}$ のタイムライン



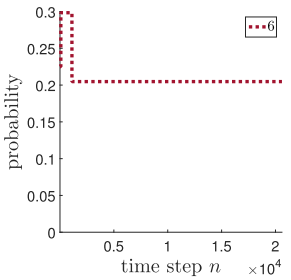
(c) $G_{a,3} = \{0\}$ のタイムライン

図 10　提案手法によるクラスタリング後のタイムライン（Oshima Tsubaki）

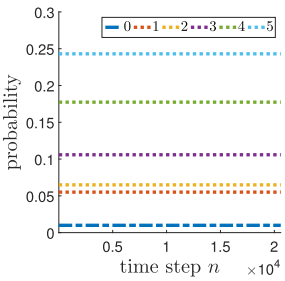
Fig. 10　Timelines of proposed method which divided by clustering results (Oshima Tsubaki).



(a) $G_{b,1} = \{7\}$ のタイムライン



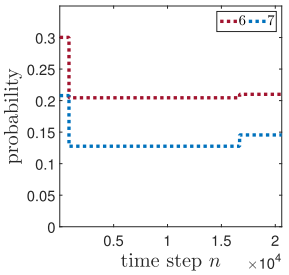
(b) $G_{b,2} = \{6\}$ のタイムライン



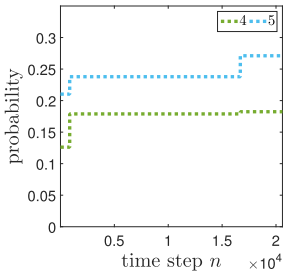
(c) $G_{b,3} = \{0, 1, 2, 3, 4, 5\}$ のタイムライン

図 11　Kleinberg の手法によるクラスタリング後のタイムライン（Albion）

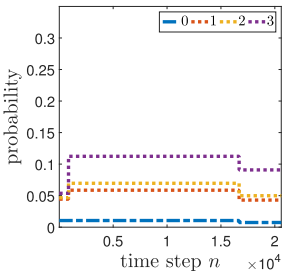
Fig. 11　Timelines of Kleinberg's method which divided by clustering results (Albion).



(a) $G_{b,1} = \{6, 7\}$ のタイムライン

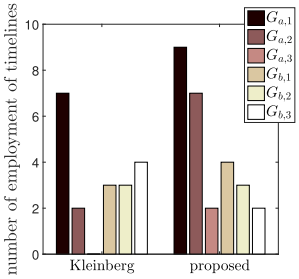


(b) $G_{b,2} = \{4, 5\}$ のタイムライン

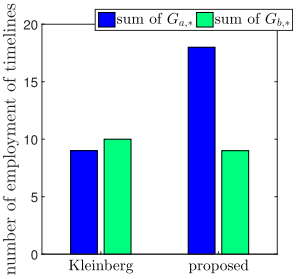


(c) $G_{b,3} = \{0, 1, 2, 3\}$ のタイムライン

図 12　提案手法によるクラスタリング後のタイムライン（Albion）
Fig. 12　Timelines of proposed method which divided by clustering results (Albion).



(a) 各タイムラインが自由記述で用いられた回数



(b) タイムラインが自由記述で用いられた回数の合計

図 13　被験者実験の実験結果
Fig. 13　Experiment results by participants.

とが見て取れる．つまり，専門知識がない一般ユーザがレビュー評点平均の変動原因を理解するうえで，提案手法の出力結果が有用であることが示唆される．

5. おわりに

本論文では，評点時系列データのレジームスイッチングに基づくタイムライン可視化法を提案した．提案手法は，各レジームのユーザの基本評点行動は多項分布に従っていると仮定し，観測された評点時系列データに対してレジームスイッチングモデルを適応することで解釈しやすい視覚結果を生成した．人工データを用いた Kleinberg の手法との比較実験では，非常に短いスイッチング間隔の場合を除いて，提案手法が正確に真のモデルを推定できることを示した．また，ノイズを含むデータに対する精度や，計算時間においても，今回の実験の場合は提案手法が優れていることを示した．現実データを用いた実験では，Kleinberg の手法と比較して，提案手法の方が解釈しやすいタイムラインと樹形図を出力できていることを示した．さらに，専門知識がない一般ユーザに対し，提案手法の出力結果が，評点平均の変動原因を理解するうえで有用であることを被験者実験により示した．

謝辞　本研究は，JSPS 特別研究員奨励費 16J11909 の支援を受けて行ったものである．

参考文献

[1] Melville, P., Gryc, W. and Lawrence, R.D.: Sentiment analysis of blogs by combining lexical knowledge with text classification, *Proc. 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD'09*), pp.1275–1284 (2009).

[2] Pak, A. and Paroubek, P.: Twitter as a Corpus for Sentiment Analysis and Opinion Mining, *Proc. 7th conference on International Language Resources and Evaluation* (*LREC'10*), pp.1320–1326 (2010).

[3] Glass, K. and Colbaugh, R.: Estimating Sentiment Orientation in Social Media for Business Informatics, *AAAI Spring Symposium: AI for Business Agility* (2011).

[4] Kleinberg, J.: Bursty and hierarchical structure in streams, *Proc. 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD-2002*), pp.91–101 (2002).

[5] Swan, R. and Allan, J.: Automatic generation of overview timelines, *Proc. 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR 2000*), pp.49–56 (2000).

[6] Zhu, Y. and Shasha, D.: Efficient Elastic Burst Detection in Data Streams, *Proc. 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD-2003*), pp.336–345 (2003).

[7] Sun, A., Zeng, D. and Chen, H.: Burst Detection from Multiple Data Streams: A Network-based Approach, *IEEE Trans. Systems, Man, and Cybernetics Society, Part C*, Vol.40, pp.258–267 (2010).

[8] Chandola, V., Banerjee, A. and Kumar, V.: Anomaly Detection: A Survey, *ACM Comput. Surv.*, Vol.41, No.3, pp.15:1–15:58 (2009).

[9] Kim, C.J., Piger, J. and Startz, R.: Estimation of Markov regime-switching regression models with endogenous switching, *Journal of Econometrics*, Vol.143, pp.263–273 (2008).

[10] Josang, A., Ismail, R. and Boyd, C.: A survey of trust and reputation systems for online service provision, *Decision support systems*, Vol.43, pp.618–644 (2007).

[11] 打田裕樹，吉川大弘，古橋　武，平尾英司，井口浩人：Web ユーザレビューにおける評価情報の時系列変化の可視化，知能と情報，Vol.22, No.3, pp.377–389 (2010).

[12] Yamagishi, Y., Okubo, S., Saito, K., Ohara, K., Kimura, M. and Motoda, H.: A Method to Divide Stream Data of Scores over Review Sites, *Proc. 13th Pacific Rim International Conference on Artificial Intelligence* (*PRICAI2014*), pp.791–800 (2011).

[13] Ward, J.: Hierarchical Grouping to Optimize an Objective Function, *Journal of the American Statistical As-*

sociation, Vol.58, pp.236–244 (1963).

山岸 祐己　（正会員）

静岡県立大学経営情報学部客員共同研究員．日本学術振興会特別研究員（PD）．2017 年静岡県立大学大学院経営情報イノベーション研究科博士後期課程修了．データマイニングの研究に従事．日本データベース学会会員．

斉藤 和巳　（正会員）

静岡県立大学経営情報学部教授．1985 年慶応義塾大学理工学部数理科学科卒業，1998 年東京大学博士（工学）．複雑ネットワークの研究に従事．電子情報通信学会，人工知能学会，日本神経回路学会，日本応用数理学会，日本行動計量学会，日本データベース学会各会員．

（担当編集委員　北山 大輔）